5 A semantic enrichment approach to linking and enhancing Dunhuang cultural heritage data

*Xiaoguang Wang, Xu Tanⁱ, Heng Gui,
and Ningyuan Song*

Introduction

The rapid development of information technologies and the data-driven research paradigm of Digital Humanities (DH) have led to a wave of document digitisation projects on books, newspapers, periodicals, photos, ancient texts etc. to meet the growing demands of data resources (Hockey 2004; Berry and Fagerjord 2017). In the last three decades, many institutions from the gallery, library, archive and museum (GLAM) community have made contributions to the digitisation and organisation of cultural heritage (CH). Some renowned GLAMs in Europe, such as the *Louvre* (Louvre 2021) and the *British Museum* (The British Museum 2021), first offered digital platforms for their collections in the mid-1990s. In April 2009, the *United Nations Educational, Scientific and Cultural Organization* (UNESCO) launched the *World Digital Library Project*. It provided free access to a vast wealth of artistic works from human history to a worldwide audience. It now contains 1,057,175 item files from 158 GLAMs in 60 countries (U.S. Library of Congress and UNESCO 2009).

The need to organise digital resources has created new challenges for CH informatics (Hyvönen 2016). Specifically, since most digital collections, archives, indexes and databases are still stored in separate repositories, they have been characterised as isolated islands of data, often called data silos. Because of the distinct formats, descriptive frameworks, vocabularies and recording methods used by different institutions, a collection of data held by one group is hardly and partly accessible to other groups (Zeng and Tan 2021). From the perspective of data processing, inconsistencies and idiosyncrasies in methods cause errors and the duplication. From the perspective of usability, the isolated data is only machine-readable instead of machine-understandable or machine-processable (Zeng 2017). From the perspective

i The research team affirms that Xiaoguang Wang and Xu Tan (txishuo@163.com) have equal contributions to this chapter.

of user-friendliness, it is difficult to comprehensively display rich cultural knowledge to audiences without contextual information.

Data silos also diminish the value of Dunhuang CH data and impede the development of Dunhuangology. Dunhuangology is an academic study of the city of Dunhuang, China, in particular its cultural and historical heritage (MediaWiki 2019). To manage and maintain Dunhuang CH data and support Dunhuangology research, many GLAMs have constructed databases for Dunhuang CH, for instance, the *International Dunhuang Project* (International Dunhuang Project 2016), the *Dunhuangology Database* (Lanzhou University Library 2013) and the *Digital Library of Dunhuang Literature* (Shaanxi Normal University Press 2016). However, these existing databases have a limited scope in advancing the field of Dunhuangology and cannot support DH research (Hu 2018; Wang et al. 2019a). Since artefacts from Dunhuang are scattered worldwide, the execution of projects and the establishment of databases are usually carried out by different institutions. In the absence of domain-specific standards and specifications, the databases are generally incompatible with each other in their methods of digital acquisition, processing and storage, which results in data silos. Researchers with shared interests continue to invest enormous resources in building up their own databases. However, despite a significant trend in DH that could support the further development of Dunhuangology, existing DH resources and infrastructure cannot currently meet the requirements of academic research.

We therefore explore the ideas and techniques of semantic enrichment for enhancing CH data and supporting Dunhuangology research. The concept, processing methods and related Knowledge Organization System (KOS) of semantic enrichment are reviewed in section “Background”. We then present a broad description of the *Digital Dunhuang Project* (DDP) in section “The Digital Dunhuang Project”. In section “The semantic enrichment approach to Dunhuang CH data”, we focus on the construction of the *Dunhuang Mural Thesaurus* (DMT) and the analysis and enhancement of Dunhuang CH data to accelerate the development of DH research infrastructure for Dunhuangology. Also, this chapter describes how to transform the DMT into Linked Data in section “The semantic enrichment approach to Dunhuang CH data”. We contend that linked open data (LOD) is the best means to facilitate the sharing, association, enrichment and reuse of Dunhuang CH data. Finally, we discuss the future development of DH research infrastructure in the field of CH.

Background

The rapid development of DH, alongside the huge demand for large-scale and high-quality data, has made data processing an essential and necessary part of DH research. As one of the foundations of this research, the digitisation of CH resources has led to the creation of massive, heterogeneous

and multi-modal CH data, to which the semantic enrichment approach is well placed to process. Semantic enrichment is a strategy that applies semantic techniques to enhance the value of data. For example, by constructing links and associations between CH data and related digital resources, providing rich contextual information improves users' understanding of CH content. It is an effective means to enhance the quality of data in the CH field (Zeng 2019).

Virtually any digital file format (text, audio, video or office formats) is suitable for semantic enrichment, but the semantic enrichment of text is the most mature among them (Retresco 2021). In addition to different file formats, the semantic enrichment approach can process the data in different structures such as structured, semi-structured and unstructured data. The algorithm employed in semantic enrichment uses explicit semantic relationships such as hierarchical relationships and associative relationships to identify people, places, organisations, items, events and other entities (Damjanovic et al. 2011). Subsequently, the algorithm determines the relative importance of the entities by considering the relevance of the information and how the entities contribute to the meaning of the text. Lastly, by identifying or recognising entities in the text, structured and machine-readable data is created from unstructured text, and this interoperable standardised content is published as LOD and made freely available. This approach can also be flexibly implemented with either a combination of human and machine processing or be fully automated (Zeng 2017). By applying semantic enrichment strategies or approaches, the retrievability and reusability of data can be improved within the CH field and the construction of research infrastructure construction can be significantly accelerated.

One of the most productive international semantic enrichment projects for CH data was launched by Europeana, the leading service platform of the European cultural community. It contains the metadata of more than 50 million digitised CH items (such as books, music, artworks) across more than 3,000 European institutions. However, the diverse types of unstructured and semi-structured data in Europeana constitute the main challenge to data processing (Europeana Collections 2021). Based on the semantic enrichment strategy, Europeana enriches its data providers' metadata by automatically linking text strings in the metadata to controlled terms from LODs or KOS vocabularies (Europeana Pro 2020). The *Europeana Semantic Enrichment Framework* was created to standardise the entire process (Isaac et al. 2017). Europeana asserts that the better way to understand the semantic enrichment process is to distinguish conceptually between the three main stages of datafication. These major stages include:

- 1 Analysis: focusses on analysing the metadata fields in the original resource descriptions, selecting potential resources to be linked to and deriving rules to match and link the original fields to the contextual resource.

- 2 Linking: the process of automatically matching the values of the metadata fields to the values of the contextual resources and adding contextual links (whose values are most often based on equivalent relationships) to the dataset.
- 3 Augmentation: the process of selecting values from the contextual resource to be added to the original object description. This might include not only (multilingual) synonyms of terms to be enriched but also further information, e.g., broader or narrower concepts (Isaac et al. 2017; Europeana Pro 2020).

Similarly, Zeng (2019) extensively reviewed the semantic enrichment practices from the perspective of structured, semi-structured and unstructured data. She summarised four semantic enrichment practices, methods and approaches that are applicable to the GLAM and CH fields:

- 1 enhancing existing metadata quality and discoverability with more contextualised meanings by using KOS vocabularies and other resources that are available as LOD. For example, *Europeana* enriches names of concepts with values from the *Art & Architecture and Thesaurus* (AAT) and the *Library of Congress Subject Headings* (LCSH). Their vocabularies are published as LOD to align concepts and values automatically (Europeana Pro 2020);
- 2 transforming semi-structured data into structured data through entity-based semantic analysis and annotation to extend access points. For example, *Kent State University* uses the semantic analysis tool *OpenCalais* to transform the descriptive information of finding aids, such as creator histories, scope, content notes and abstracts, to extract access point candidates (Davidson and Gracy 2014; Gracy and Zeng 2015);
- 3 mining unstructured data and generating knowledge bases to support knowledge discovery and enabling one-to-many usages of GLAM data across data silos while delivering intuitive online user interfaces. For example, drawing on oral history transcripts, the *Pratt Institute's Linked Jazz* project uses Linked Data to uncover meaningful connections between documents and data related to the personal and professional lives of jazz artists and to expose relationships between musicians that reveal their community network (Pattueli 2012);
- 4 making the heterogeneous contents from diverse providers semantically interoperable via shared ontology infrastructure. For example, the *Semantic Computing Research Group* (SeCo) has designed a circular data publication system to improve interoperability. Here, a shared semantic ontology infrastructure is at the heart of the system. It includes mutually aligned metadata and shared domain ontologies, modelled using semantic web standards such as the Resource Description Framework Schema (RDF Schema) and the Web Ontology Language (OWL).

If content providers outside of the circle provide the system with meta-data about a CH object, the data is automatically linked and reciprocally enriched, thus forming a *Giant Global Graph* (Hyvönen 2016).

This process of semantic enrichment puts entities, such as the label on the mural and location, at the center and reflects the transformation from document-centric to entity-centric knowledge modelling. Regardless of which step of the process that the data has reached, KOS and external resources are always critical. For instance, in the case of projects related to Europeana, each place-name in the metadata is enriched by creating links with the global geography database and knowledge-sharing platform *GeoNames*, the *Integrated Authority File* (widely known as Gemeinsame Normdatei or GND), and the community-based encyclopedia *DBpedia* (which extracts structured contents from Wikipedia and serves them as Linked Data). The links, together with the use of URIs, provide each place in the metadata with geographic coordinates and multi-language names, thus achieving the semantic enrichment of those place names (Europeana Pro 2020). There is a growing number of high-quality KOS that are commonly used. Table 5.1 below lists the recognised KOS that are relevant to this chapter.

Table 5.1 Principal KOS per domain and metadata standard

<i>KOS</i>	<i>Domain</i>	<i>Introduction</i>
DCMI (Dublin Core Metadata Initiative) Metadata Terms	General	This document is an up-to-date specification of all metadata terms, including properties, vocabulary encoding schemes, syntax encoding schemes, and classes (DCMI Usage Board 2020).
AAT (Art & Architecture Thesaurus Online)	Art and architecture	AAT is a thesaurus provided by Getty Vocabularies. It is a powerful conduit for knowledge creation, research, and discovery for digital art history and related disciplines (Getty Research Institute 2017).
CDWA (Categories for the Description of Works of Art)	Art	CDWA is a set of guidelines for the description of art, architecture, and other cultural works. CDWA represents common practice in this field and advises best practices for cataloguing based on surveys and consensus-building within the user community (Getty Research Institute 2016).
PM-DCH (Preservation Metadata for Digital Cultural Heritage)	Cultural Heritage	Sets out the general principles of cultural heritage metadata. Specific metadata standards or principles for categories of CH such as sculptures, porcelain, bronzes, furniture, potteries, caves, murals, and others (Peking University Library 2017).

The Digital Dunhuang Project

The DDP is a world-famous, pioneering project which has yielded greater access for art critics to the priceless Dunhuang artefacts and their rich CH resources and metadata. As a representative project in the CH field, the DDP provides an instructive example of digitisation and datafication.

Dunhuang CH and DDP

Dunhuang, famous as a significant stop on the ancient Silk Road, is a historical city that witnessed the blending of western and eastern cultures over thousands of years. Located in the Dunhuang oasis, the construction of the Dunhuang Grottoes began in AD 366. In 1987 the grottoes were added to UNESCO's list of *World Heritage Sites*. The Dunhuang Grottoes constitute the largest shrine and collection of Buddhist art, of immeasurable academic and cultural value to studies in history, culture and archaeology. Among these, the Mogao Caves contain more than 45,000 square meters of murals and more than 2,000 painted sculptures preserved in 492 caves (UNESCO World Heritage Center 2021).

The cultural and artistic value of the Dunhuang CH collection draws great attention from researchers. As a result, a specific area of study called *Dunhuangology* became internationally famous (Li and Li 2009). To permanently preserve Dunhuang culture and to further the development of Dunhuangology, Jinshi Fan proposed the construction of *Digital Dunhuang* in the 1980s (Fan 2014). Subsequently, the *Dunhuang Academy of China* (DAC), the *Mellon Foundation*, the *University of California Berkeley*, the *Chinese Academic of Science*, *Zhejiang University* and *Wuhan University* participated in the project.

In keeping with the national strategy for the preservation of cultural artefacts, the DDP is pursuing digitisation of the entire corpus, including collecting, processing and storing Dunhuang CH using advanced technology. The project integrates diverse data types, including photos, videos, 3D data and other literature data, into a digital repository of cultural artefacts found in the caves that can be shared globally over the Internet. Among the research outputs of the DDP are the *Mellon International Dunhuang Archive* (ARTSTOR 2021), the *Dunhuang Academic Resources Databases* (Dunhuang Academic 2019), the *E-Dunhuang* website (Digital Dunhuang 2020) and the research upon which this chapter is based.

The capacity and distribution of Dunhuang CH data

As a result of digitisation initiatives, the diverse digital data of Dunhuang have grown substantially. By 2018, DDP had finished digitising the collections of 221 caves, processing images of 141 caves, producing panoramic images of 144 caves and digitising the negatives of 45,000 photographs

(Wu et al. 2019). The total size of the digitised data reached 1PB (Digital Dunhuang 2020). Being an indispensable component of the Dunhuang corpus, the CH data plays an essential role in the preservation of Dunhuang. However, the massive, heterogeneous and multi-model properties of the data present challenges for its efficient management, organisation and practicability (Wang et al. 2019a).

The Dunhuang CH dataset is complex, reflecting both the multi-disciplinary nature of its data organisation and that of CH more widely. The disciplines represented within its corpus include religion, art, history, archaeology, linguistics, literature, ethnography, geography, philosophy, technology, architecture and various other themes or domains of knowledge. In terms of data classification, it contains both digitally native and non-native information. For instance, the native data includes digital photographs and digital archaeological reports. In contrast, the non-native data consists of the digital 3D model data of the caves, the digital images of murals, digitally scanned copies of transcripts and manuscripts and the digital text of Dunhuangology research articles, books, newspapers and bibliographies. In addition, Dunhuang CH data are also multi-model, including text, images, audio, video, multispectral data and 3D modelling data. The majority of Dunhuang CH data are unstructured images of murals, ancient manuscripts and 3D models of caves (Wu et al. 2019).

The semantic enrichment approach to Dunhuang CH data

The semantic enrichment approach proposed in this chapter describes a complete process of applying semantic enrichment to Dunhuang CH data. As shown in Figure 5.1, the process consists of data acquisition and processing, the construction of the DMT, analysis and enrichment of Dunhuang

Figure 5.1 The workflow of the semantic enrichment approach to Dunhuang CH Data.

CH data and the creation and publishing of LOD-DMT. Firstly, academic books and papers, images and the *Dunhuang Academic Resources Database* metadata were collected, cleaned and preprocessed before being stored in the local database. Then, with the help of Dunhuangology experts, DMT was designed as a domain KOS with reference to the structure of AAT and constructed using a combination of Natural Language Processing (NLP) technology and manual inspections. Semantic analysis, linking and augmentation were implemented by considering the metadata specification and the knowledge units within the corpus. Meanwhile, Dunhuang CH data were connected with the DMT and with selected external resources for data enhancement. The DMT is also published as LOD-DMT for better access, storage and sharing of open data.

Data acquisition and processing

This chapter describes various types of data, including electronic textual data from books and papers, high-resolution digital images of murals, digitally scanned copies of the Dunhuang atlas and the metadata from research papers, newspapers and bibliographies related to Dunhuangology (as shown in [Table 5.2](#)).

Table 5.2 Primary dataset for enhancing Dunhuang CH data

<i>Data</i>	<i>Content and quantity</i>	<i>Format</i>	<i>Notes</i>
Books	<i>General Catalogue of Dunhuang Caves</i> (敦煌石窟内容总录): Contains an introduction to all of the 492 caves in the Dunhuang Mogao Grottoes. <i>Dunhuang Dictionary</i> (敦煌学大辞典): Contains more than 60 classes of 6,925 entries of posthumous writings and related research literature.	PDF and Text	Optical Character Recognition (OCR) and manual inspection
Papers	Dunhuang Research (敦煌研究), Journal of Dunhuang Studies (敦煌学辑刊): Contains the bibliography and full texts of 784 journals.	PDF, CAJ, and Text	Downloaded from the China National Knowledge Infrastructure
Images and metadata	High-resolution digital images of 5 classic Buddhist murals with a resolution of 300 ppi. Digital scan copies of the Dunhuang atlas: contains 2,877 images of the Dunhuang caves, murals, sculptures, and manuscripts less than 10 MB with a resolution over 96 ppi. Metadata from the Dunhuang Academic Resources Database.	PDF, JPEG, and Excel	Provided by DAC.

The *General Catalogue of Dunhuang Caves* and the *Dunhuang Dictionary* are the two most comprehensive and influential encyclopedias in Dunhuangology (Dunhuang Academic 1996; Ji et al. 1998). We converted the digitally scanned copies of these works from PDF into a textual format using OCR. In the process, manual reviewing was conducted to verify accuracy.

The journals *Dunhuang Research* and the *Journal of Dunhuang Studies*, contained in the Chinese Social Sciences Citation Index (CSSCI, Chinese Social Science Research Assessment Center 2017), were selected as the textual data source. In total, 985 candidate articles were collected by searching with the keyword “mural” (壁画) in the principal Chinese literature database *China National Knowledge Infrastructure* (CNKI 2021). After manual de-duplication, data cleaning and format conversion of the data, 784 items of literature related to the murals were selected.

The *Dunhuang Academic Resources Database* is developed by DAC, aiming to share academic resources for the study of Dunhuangology. It provides the metadata for research articles, bibliographies, conferences, degree theses and newspapers in the field. With the help of DAC, the metadata of the *Dunhuang Academic Resources Database* was collected and stored in our local database. In addition, the image data included five high-resolution digital images of the Dunhuang murals and digitally scanned copies of the Dunhuang atlas, as well as associated basic metadata provided by DAC.

The design and construction of the Dunhuang mural thesaurus

Due to the lack of any Dunhuang CH domain-specific thesauri, the construction of the DMT is described in this chapter to demonstrate the semantic enrichment approach in the analysis, linking and augmentation of the data. The thesaurus is a KOS that normatively organises words in a hierarchical structure and other associated structures. It usually consists of authority terms and standardised domain-specific terms, concepts and semantic associations. Compared with authority files and directories, thesauri have significant advantages in interoperability and data connectivity. As a result, they are an important and commonly used KOS.

The DMT was designed as a controlled and structured vocabulary in the Dunhuang CH field (Wang et al. 2019b). Using the standard terms from DMT, fast processing and extraction of the detailed metadata from constructed texts and semantic enrichment of complex image contents became feasible. As outlined in this chapter, the DMT was constructed by a combination of a top-down process and a bottom-up process, which corresponded to the structural design step and content acquisition step, respectively.

Top-down process: Structural design

The DMT was designed by adopting a top-down structure, which contained four categories and five facets. We researched the structure of AAT,

carrying out a deep analysis of its domain generality, organisation, annotations and metadata standards. The AAT is a hierarchical thesaurus that follows the widely accepted international ISO and NISO standards for thesauri construction, of which the hierarchical structure consists of facets, hierarchy, flags and concepts. Taking into account the design of AAT, we designed the DMT with four categories: *Facet*, *Hierarchy*, *Concept* and *Instance*. The *Facet* is located at the top level of the structure, representing a set of *Concepts* with similar characteristics. The *Concepts* and *Hierarchy* are found on the next level. They were used to describe the relationships within the thesaurus. The *Instances* are entity examples beneath the level of *Concepts* and *Hierarchy*.

The rich contents of the Dunhuang murals are the subject of a range of academic disciplines, including religion, history, art etc. We researched articles and authority books to obtain a better understanding of the murals, their subjects and the hierarchical relationships between them. Finally, with the guidance of domain experts from DAC, five facets were selected for the structure of the thesaurus: *Agents*, *Objects*, *Activities*, *Time* and *Physical Attributes*.

Bottom-up process: Content acquisition and extending thesaurus

To fill and extend the thesaurus and optimise its structure, we employed a bottom-up process. The content acquisition step was carried out by a combination of automated processes and manual inspections. Using the prepared textual database (including the *General Catalogue of Dunhuang Caves*, the *Dunhuang Dictionary*, the *Dunhuang Research* and the *Journal of Dunhuang Studies*), the NLP technology was used to tokenise the sentences and filter candidate words. Specifically, we used the *Jieba* Python package (Sun 2013). This tool is based on the *Dynamic Programming and Hidden Markov Model* (HMM) and is designed to recognise and type words. Contextual information such as the *Named Entities* corpus (including location names, Buddhist terms, Dynasty calendars, names of positions held) were added with manual inspections of the data to improve auto-selection accuracy and efficiency. During this process, 18,050 candidate words were selected. Subsequently, the filtering, proofreading and classification processes were carried out manually by Dunhuangology experts and by our team.

The step of extending the thesaurus proceeded simultaneously, in which the thesaurus structure was extended and modified to improve comprehensive and accuracy wherever it could not already accommodate the new entities. Sub-categories were added based on the current categorisation and the characteristics of the murals. For instance, extra categories were added to distinguish *Buddhist Deity* and *Secular Human* in the *Agent* facet, and *Buddhist Time* (based on the Buddhist calendar) was added to the *Time* facet. By iteratively modifying the computer algorithm and thesaurus

structure, we eventually created 25 sub-categories nested under five facets and selected 4,276 words as entities in the thesaurus.

Analysis and enrichment of Dunhuang CH Data

Semantic analysis, linking and augmentation and LOD-DMT publishing were the primary processes of the semantic enrichment approach that we used to process the Dunhuang CH data. Semantic analysis was used in the pre-enrichment phase, which focusses on analysing the metadata and fine-grained units of knowledge and selecting target resources for metadata enhancement. Linking and augmentation constituted the main enrichment phase, focusing on property values, identities of entities and data description and reliability. The publication of LOD-DMT was an initial exploratory phase in working towards the creation of a sustainable ecosystem of Dunhuang CH data.

Semantic analysis

ANALYSIS OF THE METADATA SPECIFICATION

Because of the absence of any domain-specific metadata about Dunhuang artefacts, we made reference to the CDWA, the PM-DCH (Peking University Library 2017), and metadata from the Dunhuang Academic Resources Database to design a metadata specification for the Dunhuang caves, murals, sculptures and manuscripts. The PM-DCH adopts principles for categories of CH and provides specific metadata specifications for objects such as sculptures, porcelain, bronzes, furniture, potteries, caves and murals, which are the main types of cultural artefacts found in Dunhuang. We reused the general core elements for artworks (*Name, ID, Subject, Material, Techniques, Construction, Specification and Description*) from PM-DCH and CDWA to construct a domain-specific metadata specification for the Dunhuang caves, murals, sculptures and manuscripts. In addition, the property *Topic* was added to describe the artwork type of the murals. *Location* was also added to describe the location of the murals, following the advice of Dunhuangology experts. Specifically, we added the sub-property *Cave belonging to, Caves* and *Bearing to Location* to express which cave the murals belong to and where the caves are located within the grottoes.

ANALYSIS OF THE FINE-GRAINED UNITS OF KNOWLEDGE

In the field of Knowledge Organization, knowledge units are usually used for the description of the people, events, times, places and things in images and texts (Xia 2017). By following this idea and reusing the information in documents and artefacts created by ancient ancestral artists, related

to these categories listed above, was used to create a set of specifications to describe the fine-grained knowledge units. For example, the metadata entity *Label* (榜题/题识) on the mural, which may be a title of a story or name of an entity, was selected as the knowledge unit. Based on the content of the mural, the label has six core elements: *Name*, *Type*, *Subject*, *Created Data*, *Current Location* and *Description*. Furthermore, by comparing this to the DMT, the sub-properties *Construction*, *People*, *Animals*, *Objects* and *Story* were added to *Description* to increase the detail of the metadata in the label.

ANALYSIS OF THE TARGET RESOURCES SELECTION

The selection of target datasets is the most crucial step in improving the quality of CH data. Besides the DMT, we make reference in this chapter to the criteria set out by the *Task Force on Enrichment and Evaluation* of Europeana to determine the external target resources for selection. These criteria were organised around seven dimensions: *availability*, *access*, *granularity*, *coverage*, *quality*, *connectivity* and *size* (Isaac et al. 2015). We made minor adjustments to these criteria based on the particular requirements of the Dunhuang CH field and the Chinese context. The *availability* and *access* criteria refer to the targets that should be available on the Web and should be reusable under an open license. The *granularity* and *coverage* criteria determine whether the targets should have the same coverage as the source data or can complement the source data. The *quality* criterion refers to the quality of the target in terms of semantic and data modelling, especially the structure and representation of the target. *Connectivity* describes how to select the target with incoming and outgoing links to other targets. The *size* of the target dataset, i.e., the number of concepts, is also a criterion of selection. A high number of classes, properties and triples is preferable for the *size* criterion. Finally, we chose *DMT*, *DBpedia*, *Wikidata*, *AAT*, *Geonames*, the *CNKI* and the *China Biographical Database* (Harvard University et al. 2017) as the external target resources for selection.

Linking and augmentation

DESCRIBE AND ENRICH THE PROPERTY VALUE

In this chapter, we used the *OpenRefine* tool to automatically map the strings in the specification to the values in the selected KOS vocabularies and other resources. It replaced the local values with the URI of the contextual resources. For instance, when enriching the concept *Big Fahua Temple*, the tool automatically found its property *dc:type* has the corresponding concept *Temple* in the external resource *AAT* and replaced the property with the new one.

CREATE IDENTITY LINKS TO REPRESENT EQUIVALENCE

When different external resources point to the same original resource (e.g., using both DMT and Wikidata to describe the “Bodhisattva”), the property *OWL:sameAs* is imported to create the equivalent association between the resources.

ENHANCE DATA DESCRIPTION AND RELIABILITY
VIA THE TRACEABLE DATA MODEL

To distinguish the various type of descriptions of mural images, we organised the descriptions into a *reality layer* and a *fiction layer*. The *reality layer* contained the properties from reality, including descriptions about the mural itself (*identified with the edm:providedCHO*). The web resources information (*identified with edm:webResource*) were likewise added in this layer for data traceability. The fictional layer contains the properties of the content of the mural (*identified with dh:ContentObject*), including general descriptions of the image and the semantic entities. The two layers are associated via the property *crm:P62_depicts*. Based on this, each knowledge unit was linked with all relevant related information. In this way, users can trace the links between any knowledge unit and track back to the original image or resource provider to get further information.

LOD-DMT publishing

SEMANTIC DESCRIPTION

We set up an ontology model by reusing the *Getty Vocabulary Program Ontology* (GVP Ontology, Getty Research Institute 2015), the *Simple Knowledge Organization System* (SKOS, W3C 2012) and *DCMI Metadata Terms* for the standardised semantic description of LOD-DMT. In keeping with the hierarchical structure of thesauri, five core properties were defined: *skos:ConceptScheme*, *gvp:Facet*, *gvp:Concept*, *gvp:Hierarchy* and *dho:Instance*. The *Facets* and *Hierarchy* properties reuse the GVP ontology model. The corresponding class expresses the property *Concepts* in SKOS. The *Instance* class is defined as the subclass of *Concepts*. In addition, the *Object* properties of the data model, corresponding to the associations between the data, are represented by *skos:inScheme*, *skos:hasTopConcept*, *skos:narrower*, *skos:broader*, *skos:exactMatch* and *skos:related*.

SEMANTIC TRANSFORMATION

We described the semantic transformation process for LOD-DMT based on the semantic description provided by the SKOS model. In this process, each entity was given a unique and corresponding HTTP URI. Meanwhile, the associations between terms were established via the object properties of

SKOS. The tools *D2RQ* and *OpenRefine* were adapted during this process to provide automatic semantic transformations. Subsequently, verification of the consistency of the semantic transformations was carried out via the *qSKOS* tool in order to ensure the high quality of the Linked Data.

ASSOCIATION ESTABLISHING

We created associations between the LOD-DMT and AAT by using the *skos:exactMatch* property. Establishing associations between databases is a basic requirement of Linked Data. Although vocabularies in DMT are usually domain-specific, general terms or concepts do exist. We associated those vocabularies with corresponding vocabularies in AAT by using the *skos:exactMatch* property. The corresponding vocabularies are determined via *SPARQL* query searching and manual inspection. Finally, after semantic description, transformation and establishing associations, 3,853 terms and over 27,500 triples were created in LOD-DMT, and 816 terms were associated with AAT.

Knowledge service and visualisation

We established a platform to visualise the LOD-DMT (Center for Digital Humanities of Wuhan University 2019). The platform contains basic services like a Web UI, concept resolving, themes guidance, intelligent searching and a term service. In addition, it also provides advanced searching and fuzzy searching and an API with multiple data forms. While the serialisation of the Linked Data enables machine learning, visualisation modules are also provided for better human understanding. We designed several visualisation functions for the platform, including knowledge graphs, sunburst charts, tree charts and circle packing.

Discussion

To realise the efficient knowledge organisation and management of Dunhuang CH data and to construct a sustainable DH research infrastructure for Dunhuangology, this study explores the idea and technologies of semantic enrichment through the construction of DMT, and by organising, linking and enhancing the CH content with the KOS and other contextual resources. The following section summarises the original ideas and best practices presented in this study, and discusses how they could be further expanded by future work:

Data silos and DH research infrastructure

The data silos problem is the major challenge for CH informatics. It diminishes the value of Dunhuang CH data and holds back the development

of Dunhuangology. Many databases have been developed by GLAMs in the Dunhuang CH field. However, those databases still cannot support Dunhuangology research. In this chapter, we use the ideas and technologies of semantic enrichment to manage, maintain and enrich Dunhuang CH data, in which the enhanced data can serve as a sustainable basis for the further development of Dunhuangology. By creating links with KOS vocabularies and other resources, the quality of existing CH data is enhanced with contextualised meaning. Moreover, by transferring the isolated data into Linked Data, CH data's discoverability, usability, reusability and value can be maximised. Our research demonstrates that semantic enrichment is a crucial technique in DH research infrastructure that should be applied widely across CH and other disciplines in the Humanities.

Domain KOS, DMT and LOD-DMT

Domain KOS is the essential contextual resource for semantic enrichment, especially for structurally complex and semantically rich CH data. However, in the Dunhuang CH field, the domain KOS for vocabulary control and contextual enriching was lacking, which hinders the efficient organisation of Dunhuang CH data and limits the growth of the field. The DMT is the first large-scale domain vocabulary in the Dunhuang CH field. It is constructed by pooling the knowledge of domain experts and NLP technology and building a vocabulary by combining automated and manual methods. Unlike the DMT, LOD-DMT provides an example of our semantic enrichment approach. It supports Dunhuangology by providing convenient and connective access to researchers and creates a self-sustaining ecosystem of Dunhuang CH data by transforming it into Linked Data. Linked Data can be seen as a higher-level KOS. It provides the means to construct a network of associations between datasets with rich semantic contents. The construction of LOD-DMT follows the *Best Practice Recipes for Publishing RDF Vocabularies* (W3C Working Group 2008), and linking it with AAT ensures that is shareable, reusable and better quality data.

Knowledge service and user demands

The evolution of DH demonstrates the vast potential of the data-driven research paradigm in academics, but difficulties in data retrieval and data usability impede Humanities researchers. Semantic enrichment can help scholars find more relevant knowledge by adding associations between CH data and contextual resources. We designed our platform based on LOD-DMT, but we believe that it is not sufficient for the deep needs of DH researchers. Future work needs to focus in greater depth on the user experience. Questionnaires and eye-tracking experiments could be employed to explore scene-oriented and personalised knowledge retrieval.

Conclusion

This chapter presents a semantic enrichment approach to linking and enhancing Dunhuang CH data. It fills in a blank in knowledge organisation in the Dunhuang CH field, provides a methodological reference for the construction of DH research infrastructure and demonstrates how to build high-quality data and an effective digital platform. By cooperation with Dunhuangology experts, the DMT was designed and constructed as a domain KOS using NLP technology. Domain-specific metadata specifications, fine-grained units of knowledge and target resources were analysed during the pre-enrichment phase. Links and associations with related digital resources were built to augment the Dunhuang CH data. Finally, we created and published LOD-DMT to explore the semantic enrichment approach and build a self-sustaining ecosystem in the Dunhuang CH field. In the future, we will focus on evaluating the knowledge service provision platform based on the enhanced data and exploring various service methods to improve user experience.

Acknowledgements

The authors would like to express their heartfelt appreciation to Mr. Shengping Xia from Dunhuang Research Academy for providing access to the valuable Dunhuang CH data. Thanks also to Marcia Lei Zeng, Ying-Hsang Liu and Koraljka Golub for providing inspiration and helpful feedback.

Bibliography

- ARTSTOR. 2021. "The Mellon International Dunhuang Archive." Accessed March 25 2021. <https://www.artstor.org/collection/mellon-international-dunhuang-archive/>
- Berry, David M. and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge, CB: Polity.
- Center for Digital Humanities of Wuhan University. 2019. "The linked data publishing platform of Dunhuang Mural Thesaurus." Accessed May 06 2021. <http://dh.whu.edu.cn/dhvocab/home>
- Chinese Social Science Research Assessment Center. 2017. "Chinese Social Sciences Citation Index." Nanjing University. <http://cssci.nju.edu.cn/>
- CNKI. 2021. "China National Knowledge Infrastructure (cnki.net)." Accessed March 25 2021. <https://www.cnki.net/>
- Damjanovic, Violeta, Thomas Kurz, Rupert Westenthaler, Wernher Behrendt, Andreas Gruber, Sebastian Schaffert. 2011. "Semantic enhancement: The key to massive and heterogeneous data pools". In *Proceedings of the 20th intl IEEE ERK (Electrotechnical and Computer Science) conference*, pp. 413–416. https://www.researchgate.net/publication/266603290_Semantic_Enhancement_The_Key_to_Massive_and_Heterogeneous_Data_Pools
- Davidson, Sammy, and Karen Gracy. 2014. "Helping Users Finding the 'Good Stuff': Using the Semantic Analysis Method (SAM) Tool to Identify and Extract

- Potential Access Points from Archival Finding Aids.” In SAA Research Forum, Society of American Archivists. <http://files.archivists.org/pubs/proceedings/ResearchForum/2014/posters/DavidsonGracy-PosterAbstractBio14.pdf>
- DCMI Usage Board. 2020. “About DCMI.” DCMI: DCMI Metadata Terms, Accessed May 06 2021. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-1>
- Digital Dunhuang. 2020. “Enter Mogao Grottoes Here.” Accessed March 25 2021. <https://www.e-dunhuang.com/>
- Dunhuang Academic. 2019. “Dunhuang Academic Resources Database.” <http://dh.dha.ac.cn>
- Dunhuang Academic. 1996. *General Catalogue of Dunhuang Caves*. Beijing: Cultural Relic Pressing House.
- Europeana Collections. 2021. “Welcome to Europeana Collections.” Accessed March 25 2021. <https://classic.europeana.eu/portal/en/about.html>
- Europeana Pro. 2020. “Europeana Semantic Enrichment.” Accessed March 25 2021. <https://pro.europeana.eu/page/europeana-semantic-enrichment>
- Fan, Shijin. 2014. “The Meaning of Digital Dunhuang Is Permanent Utilisation.” *Chinese Academy of Social Sciences*. http://ex.cssn.cn/lxx/sxpy/201408/t20140801_1276184.shtml
- Fresa, Antonella. 2013. “A Data Infrastructure for Digital Cultural Heritage: Characteristics, Requirements and Priority Services.” *International Journal of Humanities and Arts Computing* 7 (supplement): 29–46. doi:10.3366/ijhac.2013.0058
- Getty Research Institute. 2016. “Categories for the Description of Works of Art (CDWA),” Last modified February 28, 2019. http://www.getty.edu/research/publications/electronic_publications/cdwa/
- Getty Research Institute. 2017. “Art & Architecture Thesaurus Online” The Getty. Last modified March 7, 2017. <http://www.getty.edu/research/tools/vocabularies/aat/>
- Getty Research Institute. 2015. “Getty Vocabulary Program ontology - Report generated by Parrot”, Accessed May 06, 2021. <http://vocab.getty.edu/ontology#anchor1342708431>
- Gracy, Karen, and Marcia Lei Zeng. 2015. “Creating Linked Data within Archival Description: Tools for Extracting, Validating, and Encoding Access Points for Finding Aids.” In Digital Humanities Conference of the Alliance of Digital Humanities Organizations (ADHO). <https://dh-abstracts.library.cmu.edu/works/2211>
- Harvard University, Academia Sinica, and Peking University. 2017. “China Biographical Database Project (CBDB), Accessed May 06 2021. <https://projects.iq.harvard.edu/cbdb>
- Hockey, Susan. 2004. “The History of Humanities Computing.” In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 1–19. Malden, MA: Blackwell Publishing Ltd. doi:10.1002/9780470999875.ch1
- Hu, Xiao. 2018. “Usability Evaluation of E-Dunhuang Cultural Heritage Digital Library.” *Data and Information Management* 2 (2): 57–69. doi:10.2478/dim-2018-0008
- Hyvönen, Eero. 2016. “Cultural Heritage Linked Data on the Semantic Web: Three Case Studies Using the Sampo Model.” Invited Talk to Be Published in the Proceedings of VIII Encounter of Documentation Centres of Contemporary

- Art: Open Linked Data and Integral Management of Information in Cultural Centres. <https://seco.cs.aalto.fi/publications/2017/hyvonon-vitoria-2017.pdf>
- International Dunhuang Project. 2016. "The International Dunhuang Project: The Silk Road Online." Accessed March 25 2021. <http://idp.bl.uk/>
- Isaac, Antoine, Hugo Manguinhas, Valentine Charles, and Juliane Stiller. 2015. "Selecting Target Datasets for Semantic Enrichment." https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/EvaluationEnrichment_SelectingDatasets_102015.pdf
- Isaac, Antoine, Valentine Charles, Yorgos Mamakis, and Juliane Stiller. 2017. "Europeana Semantic Enrichment Framework." Accessed March 25 2021. <https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBjlpEx0Y/edit#heading=h.bz78ahr2Isjd>
- Ji, Xianlin, Changru Tang, Wenjie Duan, and Ke Ning. 1998. *Dunhuang Dictionary*. Shanghai: Shanghai Lexicographical Publishing House.
- Lanzhou University Library. 2013. "Dunhuangology Digital Library." Accessed March 25 2021. <http://lib.lzu.edu.cn/Html/Database-cn/2008-12/26/20081226134573.html>
- Li, Ruiyuan, and Hongen Li. 2009. "The Exploring and Experience of the Dunhuangology Literature Dataset Construction, Using the Dunhuangology Literature Dataset of Dunhuang Academic China as Example." In *2009 Annual Conference of Professional Library Branch of Library Society of China*.
- Shaanxi Normal University Press. 2016. "Dunhuangology Literature Digital Library." Accessed March 25 2021. <http://dunhuang.hanjilibrary.com/>
- Louvre. 2021. "Louvre." Accessed March 25 2021. <https://www.louvre.fr/en/>
- MediaWiki. 2019. "Dunhuangology." Wiktionary. Accessed May 06 2021. <https://en.wiktionary.org/wiki/Dunhuangology>
- Pattueli, M. Cristina. 2012. "Personal Name Vocabularies as Linked Open Data: A Case Study of Jazz Artist Names." *Journal of Information Science* 38 (6): 558–565. doi:10.1177/0165551512455989
- Peking University Library. 2017. "The Metadata Standard Draft of Cultural Relics Digital Protection." Accessed March 25 2021. <https://www.lib.pku.edu.cn/portal/cn/news/0000001494>
- Retresco. 2021 "How Does Semantic Enrichment Work?" Accessed March 25 2021. <https://www.retresco.de/en/encyclopedia/semantic-enrichment/>
- Schreibman, Susan, Ray Siemens, and John Unsworth. 2004. *A Companion to Digital Humanities*. Edited by Susan Schreibman, Ray Siemens, and John Unsworth. Malden, MA, Oxford: Blackwell. doi:10.1002/9780470999875
- Sun, Junyi. 2013. "Jieba - GitHub." Accessed March 25 2021. <https://github.com/fxsjy/jieba>
- The British Museum. 2021. "The British Museum." Accessed March 25 2021. <https://www.britishmuseum.org/>
- U.S. Library of Congress, and UNESCO. 2009. "World Digital Library." Accessed March 25 2021. <https://www.wdl.org/en/>
- UNESCO World Heritage Center. 2021. "Mogao Caves." Accessed March 25 2021. <http://whc.unesco.org/en/list/440>
- W3C Working Group. 2008. "Best Practice Recipes for Publishing RDF Vocabularies." Accessed March 25 2021. <https://www.w3.org/TR/2008/NOTE-swbp-vocab-pub-20080828/>
- W3C. 2012. "SKOS Simple Knowledge Organization System - home page (w3.org)". Accessed May 06 2021. <https://www.w3.org/2004/02/skos/>

- Wang, Xiaoguang, Xu Tan, and Qingyu Duan. 2019a. "Enhancing Scholar Supportive Data: Surveying the Landscape of Information Resources for Digital Dunhuang." *Proceedings of the Association for Information Science and Technology* 56 (1): 798–800. doi:[10.1002/pra2.180](https://doi.org/10.1002/pra2.180)
- Wu, Jian, Tianxiu Yu, Chunxue Wang, Xiaohong Ding, Liang Zhao, and Liang Jin. 2019. "Review of Digital Dunhuang Project." *Communications of the CCF* 9: 10–16.
- Xia, Cuijuan. 2017. "The Research on the Digital Humanities Platform Construction Based on the Linked Open Resources Service." *Journal of Library and Information Science* 43 (1): 47–70. doi:[10.6245/JLIS.2017.431/717](https://doi.org/10.6245/JLIS.2017.431/717)
- Wang Xiaoguang, Hanghang Cheng, Huinan Li, Xu Tan, Qingyu Duan. 2019b. "Chinese Dunhuang Mural Vocabulary Construction Based on Human-Machine Cooperation." In *Digital Humanities conference (DH 2019)*. <https://dev.clariah.nl/files/dh2019/boa/1015.html>
- Zeng, Marcia Lei, and Xu Tan. 2021. "Semantic Enrichment of Data – Interpreting the New Trend of LAM Data in Supporting Digital Humanities." *Digital Humanities Research (China)* 2021 (1): 64–72
- Zeng, Marcia Lei. 2017. "Smart Data for Digital Humanities." *Journal of Data and Information Science* 2 (1): 1–12. doi:[10.1515/jdis-2017-0001](https://doi.org/10.1515/jdis-2017-0001)
- Zeng, Marcia Lei. 2019. "Semantic Enrichment for Enhancing LAM Data and Supporting Digital Humanities. Review Article." *Profesional de La Informacion* 28 (1): 1–35. doi:[10.3145/epi.2019.ene.03](https://doi.org/10.3145/epi.2019.ene.03)