Information Research - Vol. 30 No. iConf (2025)

A cultural heritage data curation framework for knowledge discovery

Ruijie He and Xiaoguang Wang

DOI: https://doi.org/10.47989/ir30iConf47323

Abstract

Introduction. Cultural heritage data resources (CHDR) serve as an important foundation for the revitalization of cultural heritage. By systematically investigating the construction process of CHDR, it helps to further clarify the resource construction process and promote the construction of CHDR oriented towards openness, collaboration, and intelligence.

Method. By means of literature review, case analysis, open coding, and other methods, 105 documents and 75 projects were analysed.

Analysis. On the basis of d-KISTI Model, a series of technical methods of CHDC are summarized from the literature and used as coding reference for further project analysis. Through the analysis of project cases, the processing and construction process of cultural heritage data resources are summarized and sorted out.

Results. A cultural heritage data curation (CHDC) framework for knowledge discovery has been proposed. There is still a broad scope for improvement in aspects such as structured expression, multi-sensory interactivity, sharing mechanisms, and copyright protection in the construction process of CHDR.

Conclusion(s). This paper focuses on the development needs and reality of resource construction in the field of cultural heritage from the perspective of knowledge discovery, as well as emphasizes the importance of collaborative opening and embracing technology in this process.

Introduction

With the proliferation of data volume and the advancement of data processing technology, data has become a core resource for research in various fields, driving the exploration and discovery of new knowledge and new laws. Against the background of the initial success of cultural heritage resources digitization, further datafication construction has become a new impetus for data-intensive academic research and an important foundation for cultural heritage to move towards intelligence. By understanding the current status of cultural heritage datafication and sorting out the process and methods of CHDR construction, we can further clarify the needs and deficiencies of data construction, promote data reuse and discovery through data resources construction, and help the field of humanities adapt to the development of the data era.

Considering that cultural heritage resources will be preserved for a long time after digitization and datafication as important basic resources for subsequent use, this paper will combine the data life cycle and data curation related theories and methods, summarize the cultural heritage data curation methodology process through literature research, analyse the related projects from data collection to data preservation and utilization, and examine the current situation of cultural heritage resources through the perspective of knowledge discovery. We will examine the current status of CHDR construction from the perspective of knowledge discovery, construct a framework for CHDC oriented to knowledge discovery, summarize the current situation and shortcomings, and provide a reference direction for optimization and improvement of CHDR construction and cultural knowledge discovery.

Literature review

Lifecycle, data lifecycle and cultural heritage data lifecycle

Life cycle refers to the entire process of an object from its creation to its demise. '*Once something has the three important attributes of continuity, irreversibility, and iteration, it can be considered to be examined from a life cycle perspective*' (Ma & Wang, 2010). The concept of the data life cycle, which focuses on the entire process of data from its creation to its destruction, emphasizing the characteristics of data and its applications at different stages and aiming to ensure its sustainability, has played an important role in bridging the gap between data management and research.

Resources building in the field of cultural heritage includes resource collection, processing, transmission, use and storage. Typically, the links are continuously iterative and cannot be skipped, which is in line with the scope of application of the life cycle. It is worth noting that most cultural heritage resources do not exist in extinction or destruction. This stage is instead replaced by the permanent preservation of cultural heritage resources. Nowadays, the construction of data resources for cultural heritage has become an important foundation for the long-term development of digital humanities. Data curation based on data lifecycle is more in line with the concerns of preservation, maintenance, and care in this field. This paper defines the lifecycle of CHDR as,

> *the entire process from the creation to the permanent preservation of CHDR, encompassing stages such as collection, processing, storage, and utilization, with the ultimate goal of achieving value-added and reusable data, characterized by its permanence and iterability.*

Data curation and cultural heritage data curation

Data curation initially emerged as a response to challenges related to long-term data preservation, scientific utilization, and effective value addition. It focuses on comprehensive management of data with a clear lifecycle, involving planning and management at the point of data creation. Throughout the entire lifecycle, active management and evaluation are employed to maintain and enhance the resource for both current and future use, thereby achieving value addition (Feng & Richards, 2018; Lee et al., 2020). In recent years, as the domains of data resources have diversified, governmental data (Gao et al., 2020) and humanities data (Laksmi et al., 2024) have gradually become noteworthy digital curation objects.

In the field of digital humanities, '*effective management and preservation of cultural heritage data contribute to the full realization of the value of cultural heritage resources*' (Laksmi et al., 2024). '*In general, data curation and digital humanities have similarities in terms of future goals and practical paths*' (Poole, 2017). '*The foundational construction and sustainable utilization of data resources are important issues of concern for both*' (Gao & He, 2023).

Knowledge discovery

'*Knowledge discovery is defined as the entire process of discovering useful knowledge from data*' (Fayyad et al., 1996). '*Earlier studies focused on data identification and data utilization*' (Guo, 2022). In recent years, as a goal-oriented theoretical perspective, knowledge discovery has been used in the construction and optimization of technical system platforms and digital service systems (Hu et al., 2022; Zhao et al., 2022), and combined with technical means such as knowledge mapping, artificial intelligence, machine learning (Han & Sun, 2021), which is regarded as an important value-added process of data (Hao & Gu, 2023). Knowledge discovery can uncover the value and application potential embedded within CHD, leading to the effective enhancement of CHD resources, and promoting in-depth applications of CHD across various fields.

Therefore, this paper integrates data lifecycle theory and data curation concepts to explore the full lifecycle management and value realization of cultural heritage data. By analysing and researching the stages of data generation, collection, processing, storage, and utilization, and combining the significant value-adding role of knowledge discovery for cultural data, a CHDC framework oriented towards knowledge discovery is constructed. Furthermore, by delving into the current practical situation of CHDR construction, this paper thoroughly discusses the management, sustainable value addition, and reusability of CHDR. It provides theoretical support and practical guidance for building a sustainable ecosystem for digital humanities and cultural heritage in the digital space.

Methodology

Foundation model

As a classification method for data curation-related operations, the d-KISTI Data curation Life cycle Model (d-KISTI Model) (Rhee，2024) draws on the DCC model. A series of operations in the process of data curation are divided into sequential operations (such as creation and collection, evaluation and selection, extraction, and preservation, etc.), accidental operations (such as re-evaluation, disposal, migration, etc.), and life-cycle operations (such as referring to the core and key parts of data curation work, respectively). And the various stakeholders and their actions, in addition to the data itself, that may influence the data curation process. This model has been widely consulted in the process of design and development and has been verified in practice in the environment of digital library. It encourages resource users to examine the preservation and utilization of data resources from the perspective of methodology, and conforms to the goal of opening, protection, and utilization of cultural heritage resources construction. Therefore, this paper chooses this model as a basic reference, examines the methods and processes of cultural heritage data resources construction, and explores the current situation of data resources construction in combination with project practice.

Document analysis

> *As a type of literature review coexisting with narrative review, systematic review is able to collect relevant research information in a more comprehensive and specific way,*

> *analyse and evaluate the literature that meets the criteria through a series of principles and methods* (Qiu, 2010).

‘*Systematic review has the characteristics of problem orientation, screening criteria, comprehensive search, and effective quality*’ (Aromataris & Pearson, 2014). This method helps researchers to find the most comprehensive ‘evidence’ possible in relation to the research question and has been widely used in research on topics such as information behaviour (Wang, 2018) and digital services (Tu & Liu, 2023).

> *The systematic review method can be roughly divided into the basic steps of problem identification, literature search, critical assessment, extraction and integration, interpretation, and analysis* (Aromataris & Pearson, 2014).

It can also be combined with looping, backtracking and other operations to enhance the comprehensiveness of the data sources, and to continuously revise and improve the research analysis and conclusions. This paper adopts the basic idea of systematic review and collects and refines the methods and techniques in the process of building CHDR based on the d-KISTI model.

Retrieval strategy

Digital resources are foundational to data resources in the CDH resource construction process, making digitization essential. This paper includes ‘*digital resource*’ and ‘*data resource*’ in its search scope, employing terms like ‘*cultural heritage*,’ ‘*digital resources*,’ ‘*data resources*,’ and their translations, along with Boolean operators and wildcards for an effective search strategy, focusing on the ‘*subject*’ field to enhance search relevance and coverage. Finally constructing the basic search formula as follows: Chinese literature search formula: SU%= ‘文化遗产’ AND SU%=(‘数字资源’ + ‘数据资源’ + ‘资源建设’), English literature search formula: TS = ‘*cultural heritage*’ AND TS = (‘*digital resource**’ OR ‘*data resource**’ OR ‘*resource* construction*’). We selected China national knowledge infrastructure (CNKI) database and web of science (WoS) core collection as the search source. We also tracked the citations and references that met the requirements, supplemented with Google scholar to obtain the citations and references that were not in the selected databases without limiting the time of publication, and used the snowball tracking strategy to minimize the omissions due to the incompleteness of the search terms.

Assessment and screening

According to the research objectives and research questions of this paper, the reference standards for literature evaluation and screening are determined as follows: the construction process of CHDR is taken as the research content, including the construction of methodology or theoretical system, the research related to the practice of project platform construction, and the type of literature is limited to the journal papers and conference papers in both Chinese and English. First of all, using the Chinese literature search formula to search CNKI database, allowing synonym expansion, we got 1059 pieces of literature, from which 759 journal papers and 40 conference papers were screened out, and by browsing the titles, keywords and abstracts, we initially got 75 journal papers and 2 conference papers, and supplemented 5 pieces of literature after tracking the cited literature and references to the above 77 papers, which amounted to a total of 82 pieces of Chinese The total number of Chinese papers was 82. Then we used the English literature search to search the WoS core collection, and obtained 118 papers, including 76 journal papers and 42 conference papers, screened 10 journal papers and 7 conference papers, and supplemented 6 cited papers and references, totalling 23 English papers.

Extract and integrate

Read the full text of the 105 selected documents, draw on the methods and ideas of grounded theory, use the d-KISTI model as the first-level coding of the reference frame for document information extraction, and extract and collect the second-level coding corresponding to each operation stage through literature reading. Due to space limitations, some open coding results are shown in Table 1.

One-level coding	Secondary coding	Partial source data
Conceptualise	Metadata	It is necessary to make a metadata description of the stele and rubbings to reveal the content features of the stele such as title, responsible person, summary, writing style, and material characteristics such as size and material.
		The main goal of this research is to develop a model to aggregate diverse CHI resources on the Web based on One-to-One Principle of Metadata. The One-to-One Principle helps to distinguish digital copies and their source related to cultural heritage objects.
Create & collect	Crowdsource collaboration	Drawing on the concept of Internet co-construction and sharing, museums should build an open platform for the public and use the wisdom of the masses to crowdsource data to build digital humanities projects.
		The construction of intelligent data cannot be separated from group collaboration and social crowdsourcing. Build a crowdsourcing platform or online community for intelligent data collation, processing and organisation in the field, and the public participate in data crowdsourcing through social networks.

Table 1. Document coding results (part)

Analysis

Combining the results of literature reading and information extraction, the CHDC method process is summarised, as shown in Figure 1.

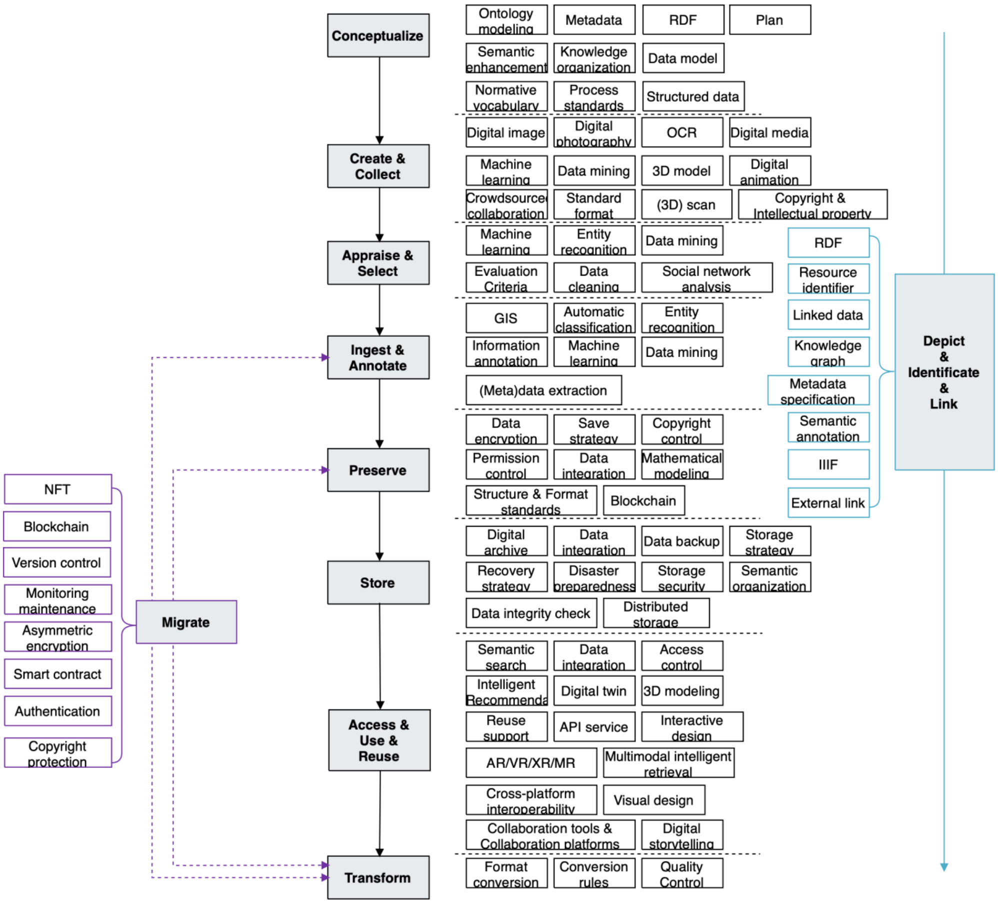

Figure 1. CHDC method and process

Sequential operations are represented by black boxes and connected in sequence by solid black arrows. The conceptualisation stage aims to apply knowledge organisation methods to construct framework standards, develop metadata specifications in combination with international standards create an ontology model, utilise the list of subject headings to carry out semantic extensions, and ultimately establish a systematic framework for describing the resources. In the specific process, the three stages of creation and collection, appraisal and selection, and ingestion and annotation are usually executed simultaneously. First, digitize physical objects through photography and scanning to form a cultural heritage database in a specific field. Then combined with machine learning to accomplish entity recognition, automatic categorisation, data extraction and other data mining and cleaning, as well as the semanticisation based on the extraction and annotation of resources. Only three papers mentioned the appraisal and selection, mostly with the degree of topic relevance as the default measure. It is noteworthy that crowdsourcing collaboration has become one of the effective modes of CHDR collection and processing, but the issue of copyright and intellectual property rights in the process of data collection has also come to the forefront. In order to realize the long-term preservation of CHDR, it is also necessary to consider appropriate preservation methods and storage strategies, and to realise data integration and reuse with as low a capacity as possible, low risk, high efficiency, multimodality, and strong correlation. Data backup, integration and security are essential requirements for cultural heritage resources management, combined with integrity checks to ensure data validity and availability, resulting in standardised digital archives. Semantic organisation and correlation of data are key to storage design, while blockchain encryption methods facilitate control and tracking of copyright and ownership. In addition, meeting the needs of different types of users to participate in the exploration and discovery process of CHDR through multi-dimensional and multi-sensory experience interactions is a key link in realizing the value of datafication and an important stage in realizing the intelligentization of CHDR and the provision of intelligent services. In addition, multi-dimensional experiential interactions to satisfy users exploring CHDR are key to datafication value realization and intelligent service delivery. Interactive design, visualization design, virtual reality and its extension technology are widely used in the construction, display and utilization of various data resources, intelligent semantic search and recommendation based on multimodal data integration, and cross-platform interoperability design have been paid attention to, and the development and establishment of collaborative tools and cooperation platforms have become the key infrastructure for data openness and reuse. It is also possible to enhance the possibility of data access and reuse by transforming changes in the external form of data such as format, which is common in the transformation of metadata format and resource modality, but not much has been mentioned so far.

Occasional operations are located on the left side of the framework diagram and are connected to the possible stages of the operation by dashed arrows. Resource selection is mostly done at the '*assessment and selection*' stage and emphasizes attention to the flow of copyright and ownership during the migration process. The use of NFT, blockchain, smart contract, asymmetric encryption, and other methods to realize copyright traceability and identity authentication has been preliminarily explored and is considered for monitoring of CHDR.

The full life cycle operations are located on the right side of the framework diagram, with long straight lines running vertically through the sequential and occasional operations. The stage of description, identification and linkage aims to realize accurate description, identification, and linkage of cultural heritage resources by actively linking internal and external resources through RDF, Linked Data, Knowledge Graph and other methods and technologies, and by using them as a reference to coordinate and plan each attribute and characteristic of the data. This is not only an effective way to expand the data utilization scenario, but also an important basis for enhancing the possibility of data discovery.

Project analysis

Sample selection and coding results

Considering the relevance of the project, the impact of the award and the accessibility of the data, we have chosen the 2022 China digital annual conference excellent project cases of humanities (CDH) and award-winning projects of digital humanities awards (DHA) are selected as the research objects. Due to the small number of projects in CDH2022, excellent projects of CDH2021 are also included in the research scope. After removing one duplicate project, a total of 75 digital humanities projects became our research objects.

Based on the above series of summaries of the operation process of the literature related to CHDC, we will further combine the bottom-up open coding method to collect, visit, read and analyse the literature materials (such as papers, news, etc.) and platform websites of 75 projects. First of all, we take the names of each stage in Figure 1 as the reference of first-level coding, and the specific technical methods in it as the second-level coding to form the initial coding scheme. Then based on this, the project data is investigated and analysed, and the coding scheme is constantly adjusted, supplemented, and improved in the process. The final coding scheme was obtained as shown in Table 2.

Stage Name	Concrete Action	Number of Items
Standards, guidelines & specifications Design	/	2
Initial conceptualization	Metadata & Vocabulary Reuse	5
	Knowledge organization	6
	Ontology modeling	0

Stage Name	Concrete Action	Number of Items
Data collection & Creation	Physical data collection	31
	Digital data collection	23
	Crowdsourcing collaboration	7
	Evaluation & Selection	7
	Storage	2
Conceptual improvement	Vocabulary building	6
	Metadata specification	13
	Data & Model structuring	12
Extraction & Annotation	Extraction	10
	Annotation	4
Identification & Linking	External linking	12
	Internal linking	7
Presentation & Display	Accessibility	29
	Visualization	28
	Interactivity	23
	Experientiality	3
	Reusability	30
	Linkability	1
	Collaborability	11
Others	Copyrights & IPR	7
	Data transformation	1

Table 2. Coding scheme

Findings

The design of standards, guidelines and specifications is the first step. The operational content mainly focuses on digitization operation, process specification and operation guidelines, reflecting the transitional characteristics of the integration and construction stage of resources in cultural heritage resources domain towards the integration and utilization stage. The initial conceptualization stage does not seek to be exhaustive, but rather to design the basic framework of the resource and improve the interoperability of the data through the reference and reuse. The operations are mainly focused on the reuse of metadata sets and vocabularies and the initial establishment of the research project's knowledge organization methodology. The data collection and creation stage in this field has not only seen the traditional researcher-led model, but also the enhancement of the breadth and diversity of cultural heritage data through the power of the public. Most projects choose to digitize physical materials directly and use existing resources through correlation or extraction, avoiding the waste of costs associated with repeated collection of similar information.

Data appraisal and selection are the key to ensuring the quality of massive datasets, with key metrics including copyright, accuracy, and uniqueness. Copyright is concerned with ensuring legitimacy, accuracy assessment ensures data quality, and uniqueness requires enhancement of the character and value of CHDR. In terms of data storage, cultural heritage data projects are concerned about the association between data and databases, for example, by matching metadata tables with encoding index tables to help facilitate join operations.

The refinement of the conceptualization stage involves extending the initial word list framework and constructing specialized metadata specifications in conjunction with standardized word lists. Given the diversity and complexity of cultural heritage data, researchers need to customize the ontology to accommodate domain specificity and integrate data from different sources to achieve heterogeneous data mapping and integration. The extraction and annotation stage enriches the data content and improves the information density through key information refinement and data annotation, including text entity relationship extraction, image recognition classification, and enhances the data semantics by combining with the open annotation model, while actively adopting the annotation method combining manual and machine learning. The identification and linking stage build a complex semantic network through internal and external data association for effective resource identification and linking. External linking utilizes databases such as entity names or Wikidata or interoperability frameworks, while internal associations use knowledge graphs and keywords to facilitate multidimensional analysis and data exploration.

The task of the external presentation and display stage is to present processed cultural heritage data to the user in a rich and understandable way and providing interactive channels to meet diversified needs. This stage consists of seven aspects: accessibility emphasizes open access to the resource, visualization enhances the visual experience through static and dynamic means, interactivity allows users to explore the resource and have a personalized experience, experiential creates immersive environments through digital technologies, reusability considers the re-creation and sharing of the resource, linkability emphasizes the overall openness of resources, and collaborability supports interaction and cooperation among users. Although some projects achieve these aspects, there is still room for improvement in the areas of open licensing, API usage, and user community building. In addition, some projects have focused on copyright and intellectual property protection and data transformation.

In general, the operations involved in CHDR building projects are adapted to the specific needs of the project, but usually include the development of digitization specifications, technical collection, and conversion of resources, ensuring data copyright and quality, refining conceptualization and structured representations, as well as enhancing linkages and multi-format presentations of the resources, and ultimately, the sharing and use of the resources.

A CHDC framework for knowledge discovery

Based on the literature review of life cycle and data curation, we carried out literature reading and collating on the theme of CHDC, obtained a series of specific technical methods, and summarized the main methodological framework of CHDC based on the investigation and analysis of 75 project cases. Since the focus of CHDC is to satisfy the knowledge mining and utilization of cultural heritage resources, we believe that the concept of knowledge discovery can be used to emphasize the goal of this process, and a framework for CHDC oriented to knowledge discovery is proposed, as shown in Figure 2.

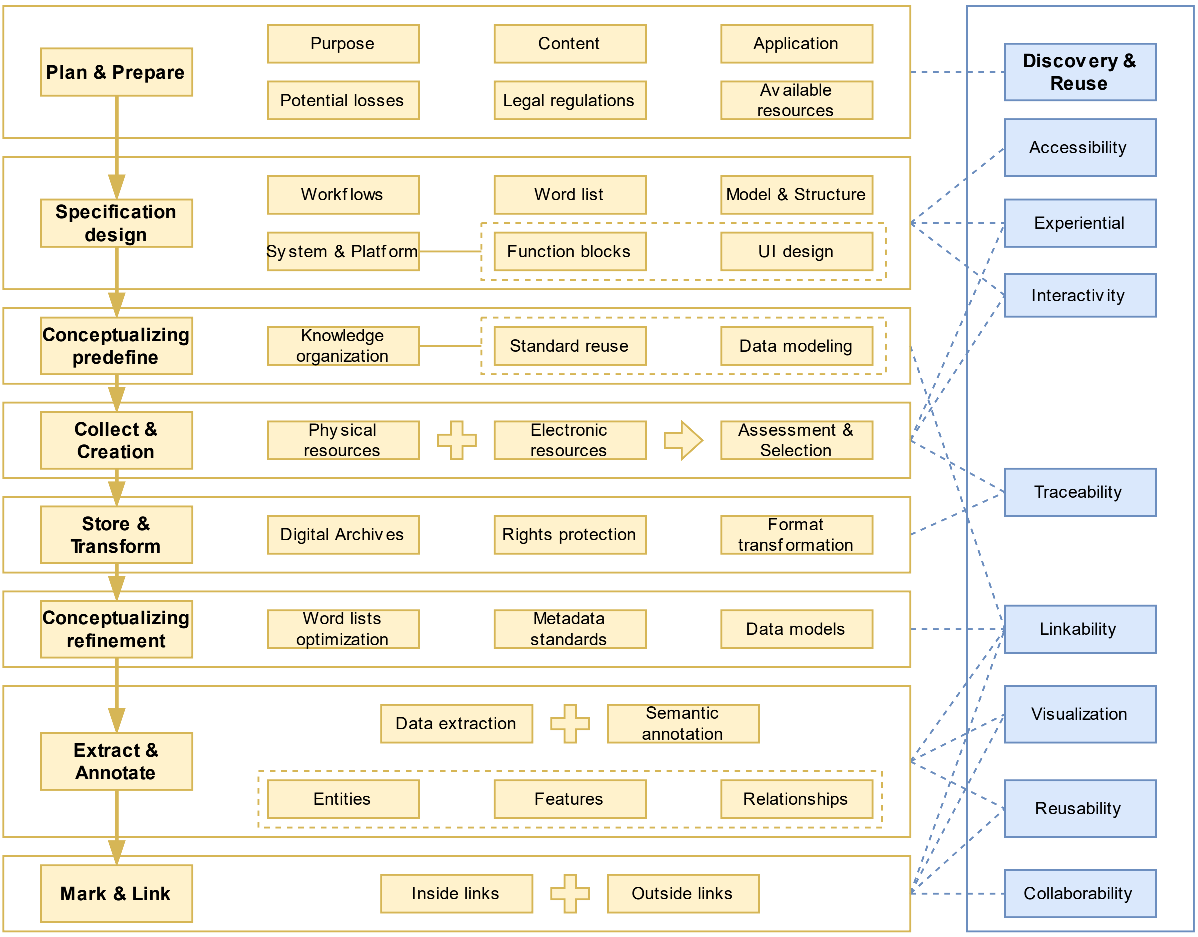

Figure 2. A CHDC framework for knowledge discovery

Explanation of the content of the framework

In the stage of planning and preparation, the project team needs to predefine the objectives, content, and use of cultural heritage resources, assess potential damages, clarify legal norms, and research existing resources. The aim is to provide a framework proposal of overall objectives and construction models for knowledge discovery and reuse. The specification design stage needs to develop specifications for data collection and processing, research and organize metadata standards and ontology models. Start early in the project and iterate with implementation to ensure that the dataset or system platform helps users to accurately understand and utilize the data. The conceptualizing predefinition stage requires the formation of metadata sets, word lists, and ontology models suitable for the project based on research and standards, to enhance the discoverability and availability of data, and to prepare for the open sharing of data. The stage of data collection and creation aims to collect and screen physical and electronic resources through technological means to create a digital archive and lay the foundation for a data resource for knowledge discovery potential. Attention also needs to be paid to copyright and data traceability. The stage of storage and transformation focuses on the traceability of copyrights and intellectual property rights, providing tools, rules and mechanisms for format transformation and safeguarding the project's own copyrights and intellectual property rights. The stage of conceptualizing refinement focuses on refining and perfecting word lists, metadata standards, and data models to form a complete and consistent data model and working to support data linking and retrieval. The stage of extraction and annotation involves extracting and annotating data entities, features, and relationships, integrating different data sources, and enhancing the semantic richness of the data. At the same time, it is necessary to consider the possibility of displaying the discovery results through visualization methods to extend the data application scenarios. The stage of marking and association focuses on realizing the links between inside and outside resources, facilitating the integration of different modal resources, and enhancing comprehensibility and discoverability. It is also necessary to realize semantic enrichment through methods such as knowledge mapping to enhance the explorability of resources as much as possible. The stage of discovery and reuse is the core part of the later stages of building CHDR, focusing on user experience and engagement. And the discovery and reuse of data value is realized through resources, tools, and methods.

Characterization of the framework

Examining CHDR building through the data curation lens

Based on the analysis of the specificity of the life cycle of CHDR, this paper investigates the methodological process of CHDR construction and establishes a CHDC framework accordingly with reference to the d-KISTI model. On the one hand, '*traditional data lifecycle theory and data curation methods focus on data in a broad sense or big data, mainly focusing on research data and scientific data*' (Yoon et al., 2022), and less on the exploration and construction of cultural heritage data. On the other hand, the traditional construction of CHDR lacks a unified methodology and process specification, and the introduction of the data curation method can provide a more complete process of data collection, organization, analysis, display and utilization. On the other hand, the traditional CHDR construction lacks unified methodological process specification, the introduction of data curation method can provide a more complete process guidance from data collection, collation, analysis to display and utilization, which can help to systematize and standardize the process of data resource construction in this field.

Explaining the process of building CHDR around knowledge discovery

With the help of knowledge discovery, the potential value of CHDR can be more fully realized. Combined with the demand and realization of knowledge discovery, this paper proposes to integrate knowledge discovery into each stage of CHDR construction, so as to realize the continuity and sustainability of the whole process from data collection to knowledge discovery. At the early stage of the construction process, it emphasizes the predetermination and target determination of knowledge discovery, and provides support for knowledge discovery through specification design, data collection and processing, and conceptualization construction, so as to give users more ability to discover and explore the resources.

Emphasizing copyright and intellectual property rights for CHDR

'*Data curation explicitly requires the specification of rights and accountability*' (Brink, 2017) and the adoption of appropriate technical means to protect the security and legitimacy of data. In this paper, the active attention and protection of copyright and intellectual property rights are explicitly proposed in the framework of CHDC, and how to incentivize the creation and reuse of cultural heritage data and safeguard the legitimate rights and interests of resource owners or providers still need to be actively explored in conjunction with data encryption and transmission technologies.

Combining literature research and project platform practice

In the process of framework construction, this paper fully combines two means of literature research and project research. Through extensive literature research, in-depth understanding of CHDR construction related theories and methods and technologies, to provide a basis for reference for the framework construction; at the same time, combined with the specific practice of the project platform to verify and optimize the framework, to improve the relevance and applicability of the framework.

Future works

Actively promote universal structured expression design

Standardizing structured representations of CHDR by means of feature classification, metadata description, and ontology modelling facilitates knowledge storage, representation, and retrieval, and is essential for resource protection and intelligent application. Despite the current diversity of structured representations in the cultural heritage domain and the lack of unified standards, many projects choose to reuse and extend existing standards in similar domains. While this approach solves the problem of generalizability, there are still challenges of semantic inconsistency and cross-domain compatibility. In addition, further research is needed on how to make the organization of resources comprehensible to users from different backgrounds in order to attract them to explore the deeper laws and knowledge of cultural heritage.

Wide range of technological approaches to hyper-realistic multi-sensory interactivity

With the development of virtual simulation technology, the field of cultural heritage is shifting to scenario-based and experiential communication based on digital media, emphasizing multi-sensory and multi-angle immersive experiences. Panoramic image technology provides users with a virtual environment for close observation and hands-on experience, promoting the multidimensional presentation of cultural heritage information (Jiang & Zhang, 2023). However, current resource building projects mostly focus on data opening and display, lack interactive design from the user's perspective, and are mainly in the form of passive browsing. Cultural heritage projects can explore the intelligent generation and interaction of characters or scenes and utilize new digital media technologies to enhance the discoverability and exploration of knowledge, creating a more interesting and unique cultural experience.

Building resource synergy and open sharing mechanisms

There are deficiencies in the collaborative openness and sharing of CHDR, which are mainly manifested in the lack of platforms and standards for interoperability between projects, leading to the phenomenon of resource silos; an imperfect data ecology for user participation, limited channels for data access and reuse, and a lack of incentive mechanisms.

> *Realizing collaborative openness is a change in the way of knowledge production in the digital environment, involving the providers, deliverers, and users of resources, which requires the establishment of a support system*' (Wu et al., 2022).

The initial awareness of collaborative openness should be strengthened, translated into technical specifications and standards, drawing on excellent cases and designing an open platform, so that the concept of collaborative openness can be carried through the whole process of resource construction.

Optimize the means of copyright protection

'*The construction of culture heritage data resources involves original humanistic data, digitized materials of traditional literature and processed data*' (Ou, 2023), and the copyright issues are complex and risky. Uncertainty exists in the attribution of copyright of original materials, the legality of self-constructed materials, and the citation of third-party databases. The open standard of new-born resources is not uniform, the diversity of resource types leads to the complexity of the degree of authorization and openness, and the rules of open license agreements of different platforms are different. The definition of copyright in data reuse requires continuous exploration of legal policies and technical measures.

Conclusions

Digitization enhances the preservation, efficiency and value of cultural heritage resources and promotes the development of related construction projects. This paper systematically studies and analyses the construction of digital CHDR, including standardized design, conceptualization, data collection, refinement, extraction and annotation, identification, and association, display and copyright, and proposes a framework for knowledge discovery. The research shows that although resource conceptualization and opening are more mature and interactive display has become mainstream, the diversity and complexity of cultural heritage resources requires customized solutions, while insufficient datafication platforms, limited application of virtual technologies, and copyright issues still constrain the development of large-scale collaborative knowledge discovery.

The digitization and datafication of CHDR is the key to their adaptive use, and in the era of artificial intelligence, sorting out process frameworks and standard specifications can help to improve the standardization of resource construction and promote data opening and sharing. The shortcomings of this paper lie in the possible bias in the understanding of the selected projects and the lack of expert support and empirical testing of the proposed framework. Future research will conduct empirical tests and practical applications, focus on the application of AI in cultural heritage resource construction, and collect expert feedback to conduct experiments from the user's perspective.

Acknowledgments

The authors would like to acknowledge the support of reviewers for their helpful comments. This research work has been supported by the Major Program of National Social Science Foundation of China (21&ZD334).

About the author(s)

Ruijie He, corresponding author, is a postgraduate student at the School of Information Management at Wuhan University, Wuhan, Hubei, China. Her research interests are digital humanities and data governance. Her contact address is ruijiehe_ch@163.com

Xiaoguang Wang is a professor at the School of Information Management at Wuhan University, Wuhan, Hubei, China. His contact address is wxguang@whu.edu.cn

References

Aromataris, E., & Pearson, A. (2014). The systematic review: an overview. AJN The American Journal of Nursing, 114(3), 53-58. https://doi.org/10.1097/01.NAJ.0000444496.24228.2c

Brink, J. A. (2017). Big data management, access, and protection. Journal of the American College of Radiology, 14(5), 579-580. https://doi.org/10.1016/j.jacr.2017.03.024

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37. https://doi.org/10.1609/aimag.v17i3.1230

Feng, Y., & Richards, L. (2018). A review of digital curation professional competencies: theory and current practices. Records Management Journal, 28(1), 62-78. https://doi.org/10.1108/RMJ-09-2016-0022

Gao, D., & He, L. (2023). Digital humanities researches from the perspective of data intelligence empowerment: data, technology, and applications. Library Tribune, 43(09), 107-119. https://doi.org/10.3969/j.issn.1002-1167.2023.09.014

Gao, G. Zhu, M., & Duan J. (2020). Research on government big data service standardization system based on data curation. E-Government, 216(12), 110-120. https://doi.org/10.16582/j.cnki.dzzw.2020.12.012

Guo, S. (2022). Research on library's knowledge service model of data-driven knowledge discovery. Journal of Library and Information Science, 7(05), 23-27. https://doi.org/10.3969/j.issn.1005-6033.2022.05.005

Han, T., & Sun, M. (2021). Intellectualization of scientific research and knowledge service: conceptualization, realization and opportunity. Information Studies: Theory & Application, 44(10), 41-49. https://doi.org/10.16353/j.cnki.1000-7490.2021.10.006

Hao, X., & Gu, X. (2023). Al reshapes the view of knowledge: knowledge creation and education development under the influence of data science. Chinese Journal of Distance Education, 43(05), 13-23. https://doi.org/10.13541/j.cnki.chinade.2023.05.002

Hu, Y., Li, X., & Zhu, X. (2022). The knowledge discovery of integrating subject words and citation: data optimization and content visualization. Journal of Intelligence, 41(10), 130-137+155. https://doi.org/10.3969/j.issn.1002-1965.2022.10.019

Jiang, N., & Zhang, J. (2023). The evolution, presentation, and management of cultural heritage information from the perspective of digital humanities. Journal of Tongji University (Social Science Edition), 34(06), 82-93. https://doi.org/10.3969/j.issn.1009-3060.2023.06.009

Laksmi, L., Suhendra, M. F., Shuhidan, S. M., & Umanto, U. (2024). The readiness to implement digital humanities data curation of four institutional repositories in Indonesia. Digital Library Perspectives, 40(1), 80-95. https://doi.org/10.1108/DLP-04-2023-0031

Lee, H., Yoon, S., & Park, Z. (2020). “SEMANTIC” in a digital curation model. Journal of Data and Information Science, 5(1), 81-92. https://doi.org/10.1108/10.2478/jdis-2020-0007

Ma, F., & Wang, J. (2010). A Literature Review of Studies on Information Lifecycle(I)——the Perspective of Value. Journal of the China Society for Scientific and Technical Information, (5), 939-947. https://doi.org/10.3772/j.issn.1000-0135.2010.05.024

Ou, Y.(2023). Risks and prevention strategies of data copyright in digital humanities application service. Journal of Library Science in China, 49(01), 118-128. https://doi.org/10.13530/j.cnki.jlis.2023008

Poole, A. H. (2017). “A greatly unexplored area”: Digital curation and innovation in digital humanities. Journal of the Association for Information Science and Technology, 68(7), 1772-1781. https://doi.org/10.1002/asi.23743

Qiu, X. (2010). Systematic review: a more scientific and objective overview method. Documentation, Information & Knowledge, (01),15-19. https://doi.org/10.13366/j.dik.2010.01.016

Rhee, H. L. (2024). A new lifecycle model enabling optimal digital curation. Journal of librarianship and information science, 56(1), 241-266. https://doi.org/10.1177/09610006221125956

Tu, Z., & Liu, X. (2023). The contents and forms of digital scholarship services: a systematic review and comparative study. Library and Information Service, 67(08), 104-114. https://doi.org/10.13266/j.issn.0252-3116.2023.08.010

Wang, Y. (2018). User value of science and technology information: an integrative conceptual framework based on systematic literature review. Publishing Journal, 26(01), 82-89. https://doi.org/10.13363/j.publishingjournal.2018.01.018

Wijesundara, C., & Sugimoto, S. (2018). Metadata model for organizing digital archives of tangible and intangible cultural heritage and linking cultural heritage information in digital space. Library and Information Science Research E-Journal. https://doi.org/10.32655/LIBRES.2018.2.2

Wu, J., Shi, J., & Xu, J. (2022). Open data ecosystem for digital humanities: constituent elements and model frameworks. Library and Information Service, 66(22), 44-54. https://doi.org/10.13266/j.issn.0252-3116.2022.22.004

Yoon, A., Kim, J., & Donaldson, D. R. (2022). Big data curation framework: Curation actions and challenges. Journal of Information Science, 0(0). https://doi.org/10.13266/10.1177/01655515221133528

Zhao, Z., Liu, Y., Zhu, L., & Wu, X. (2022). Research on patterns and methods for knowledge construction and reuse in a complex information environment. Journal of the China Society for Scientific and Technical Information, 41(12), 1266-1279. https://doi.org/10.3772/j.issn.1000-0135.2022.12.005