Postprint: An Analysis of the Connotation, Characteristics, and Principles of Semantic Organization in the Archival Datafication Process

Authors: Qi Tianjiao, Feng Huiling

Date: 2023-04-01T16:02:50+00:00

Abstract

[Objective/Significance] In the archival datafication stage, archival utilization and services need to meet user demands at the data level, breaking through the limitations of page-level reading and file-level utilization. This requires constructing a new semantic-level archival organization model in the organization phase to enable fine-grained analysis and mining of archival content, context, and structural data, and to prepare resources, methods, and technologies for the value-added development and intelligent knowledge services of archival resources. [Method/Process] Using methods such as literature research and case analysis, and based on the characteristics of the archival datafication stage, this study analyzes the basic connotations of archival semantics, semantic associations, and semantic organization, compares and analyzes the differences and characteristics between archives and other information resources in the semantic organization process, and explores the basic principles of conducting inward and outward archival semantic organization under the principles of semantic integrity, chain association, and network multidimensionality. [Result/Conclusion] Archival semantic organization is a new model of archival organization based on the meaning and associations of data, aiming to discover semantics and semantic associations from the content, context, and structural data of archival resources. Archival semantic organization is the core link in achieving archival datafication and a critical step toward making archives machine-understandable and machine-actionable. Through archival semantic organization, originally discrete, distributed, and domain-dependent archival content, context, and structural data can be clarified in meaning, formalized in encoding, and linked in relationships, enabling archival data to be machine-understandable and machine-actionable, making automated associative organization, storage, and provision of archives possible, and ultimately supporting intelligent acquisition, utilization, and services of archival resources based on human-machine interaction and machine-machine interaction.

Full Text

The Connotation, Characteristics, and Principles of Semantic Organization in the Process of Archival Datafication

Qi Tianjiao, Feng Huiling

School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [Purpose/Significance] In the stage of archival datafication, archival utilization and services must meet users' needs at the data level, breaking through the limitations of page-level reading and file-level utilization. This requires constructing a new semantic-level archival organization model during the organization phase to enable fine-grained analysis and mining of archival content, context, and structural data, and to prepare resources, methods, and technologies for archival value-added development and intelligent knowledge services. [Method/Process] Using literature investigation and case analysis methods, and based on the characteristics of the archival datafication stage, this paper analyzes the basic connotation of archival semantics, semantic relations, and semantic organization, compares the differences and features between archival and other information resources in the semantic organization process, and explores the basic principles of inward and outward archival semantic organization under the principles of semantic integrity, chain association, and network multidimensionality. [Result/Conclusion] Archival semantic organization is a new archival organization model based on the meaning and association of data, aiming to discover semantics and semantic relations from the content, context, and structural data of archival resources. Archival semantic organization is the core link in achieving archival datafication and the key step toward making archives machine-understandable and machine-operable. Through archival semantic organization, originally discrete, distributed, and field-dependent archival content, context, and structural data can be clarified in meaning, formalized in encoding, and linked in relations. Archival data thus becomes machine-understandable and machine-operable, enabling automated associative organization, storage, and utilization of archives, which ultimately supports intelligent acquisition, utilization, and services of archival resources based on human-machine and machine-machine interaction.

Keywords: archives; archival datafication; semantics; semantic relation; semantic organization

Classification Number: G271
DOI: 10.13266/j.issn.0252-3116.2021.09.001

After two decades of digitizing existing archives and managing incremental electronic records, the storage of digital archival resources has become extremely rich. However, existing archival organization methods can only support file-level

archival utilization and services based on page reading. They cannot achieve machine-understandability of archival content or enable associative utilization of content, context, and structure. A critical question for advancing archival management is how business-driven archives can establish connections with the background of record creation to provide information "feedback" that supports business decisions. Former National Archives Administration Director Yang Dongquan once pointed out: "I want to utilize archives without having to search for them myself—they should be automatically pushed to me. This requires doing something more important, more labor-intensive, and more valuable: datafying archives." This points to the goal that archival datafication should achieve when facing new computer technologies such as the semantic web and artificial intelligence: transcending the limitations of file-level searching and page reading to provide users with more intelligent archival services.

Archival datafication represents a higher stage beyond archival digitization. While digitization converts archival information from analog or physical signals into digital "0" and "1" formats through scanning and word processing, transforming fixed objects in archival texts into discrete bits stored in computer systems or databases rather than paper media, datafication reorganizes these discrete bits into structured, standardized, open, and universal data objects. It then conducts corresponding machine operations based on the different forms and categories of these data objects. The key to archival datafication lies in organizing scattered bits into meaningful and related data collections—that is, data organization. The core objective of this data organization is to enable machines to understand data meanings and thereby achieve automated machine operations on data. Therefore, organizing data at the level of meaning becomes particularly important, which is precisely what this paper discusses as semantic organization.

The application of semantic technologies in the information resource organization process has been a research hotspot in the library, archives, and museum fields in recent years. Existing research has primarily reflected on the shortcomings of information organization, advocating for a shift toward finer-grained information organization and more intelligent information dissemination and application. It emphasizes that classification and description of information resources should develop toward deeper semantic revelation and relationship discovery. Consequently, "semantic organization" has become a frequently discussed concept in the information resource management field. Unfortunately, there is currently no unified formulation of the "semantic organization" concept. Some scholars approach it from the object of semantic organization, arguing that it includes three basic levels: "semantic description, ontology transformation, and publication as linked data." Others have explored specific semantic organization approaches for different types of information resources in various domains, including: ontology model construction and layered semantic description for intangible cultural heritage multimedia resources oriented toward cultural dissemination and inheritance, semantic organization of public opinion information for government decision-making, semantic organization research on scientific data

in e-Science environments, and semantic association research on scientific and technical reports. At the level of semantic technology application pathways, the basic 思路 of existing research outcomes involves knowledge extraction, ontology construction, knowledge graphs, and ontology-based retrieval for information resources. However, because supporting technologies such as the semantic web, knowledge engineering, and artificial intelligence are rapidly evolving, academic research on information resource semantic organization is in a stage of rapid conceptual expansion and technological integration, accompanied by issues such as confused terminology and concept usage. For instance, concepts like metadata, vocabularies, ontologies, and linked data repeatedly appear in numerous research outcomes, yet the underlying principles and their relationship to the transformation of information resource organization have not been clearly explained.

Driven by the broader information resource field, the archival field has also begun to explore semantic technology application solutions. For example, genealogy websites such as FamilySearch.org and Ancestry.com use ontology technology to reconstruct genealogical archival data, establishing multidimensional semantic relationships such as spatiotemporal relations to reveal hidden relationships between individuals and other knowledge within genealogical archives, and providing users with multidimensional retrieval. Kanagawa University's Center for Non-Written Cultural Heritage in Japan has datafied some folk tool materials and built a folk tool database based on ontology technology. In 2011, the French National Archives published its thesaurus in RDF format, providing users with linked data-based semantic query services. Chinese scholars have also proposed applying related technologies to improve archival organization. For instance, Qiu Li argued that semantic web technology should be used to achieve data transformation, description, and classification, and intelligent Agent technology should be used to integrate and optimize service information for ambiguous users. Ma Yinyuan, based on SWOT analysis, analyzed the influencing factors and countermeasures for applying linked data methods in archival knowledge services, concluding that linked data application is an important direction for future archival knowledge services. However, scholars have discussed but not yet reached a consensus on the precise meaning of "archival semantic organization" and its unique connotation in the archival field. For example, some scholars, starting from the question of "what to organize," believe that archival semantic organization mainly includes the organization of semantic relations in archival information resource content systems and knowledge systems, and the mapping relations between vocabulary systems and metadata systems. Other scholars, starting from the question of "how to organize," believe that the archival semantic organization process includes metadata semantic transformation, archival data semantic analysis and representation, semantic organization and storage, and semantic retrieval and services. Such connotations are essentially operational pathways. However, what exactly archival semantics, semantic relations, and semantic organization refer to remains unclear.

Against this background, this paper addresses the fundamental research ques-

tions of what archival semantic organization is and how to implement it. Using literature investigation and case analysis methods, it examines the semantic traditions in the development of archival organization, compares archival semantic organization with archival entity organization and archival information organization, defines the connotation of archival semantic organization in the context of datafication, and explores the principles and rules for applying semantic technologies to archival organization based on the characteristics of archival resources. This provides semantic-level solutions to the core issues of data organization in the archival datafication process.

2. The Semantic Organization Tradition in Archival Organization

Archival semantic organization is a new application of semantic technology in archival organization, representing new thinking and new methods emerging in the datafication process. However, semantic organization did not appear out of thin air. The long-term development of archival organization has also formed a latent semantic organization tradition that can lay a solid foundation for new-era archival semantic organization development.

2.1 The Semantic Organization Tradition in Archival Entity Organization

Archival entity organization targets the physical entities (carriers) of archives, aiming to achieve the ordering of collection entities. In China, archival entity organization borrowed from the former Soviet Union's "General Regulations for State Archives," with the principle of fonds as the core, and then classified and filed according to the physical characteristics of archives. Its basic components include dividing fonds, classifying archives within fonds, filing, and arranging within files, thereby classifying archives based on their source, time, content, and form characteristics. The connotation of archival arrangement proposed by He Baorong, Chen Zhaoyu, and Song Shiqin, which divides the work into "systematization and basic cataloging" and follows the principle of "arranging archives according to the historical connections between documents," continues to influence practice today. Here, "historical connections" refer to the relationships between documents within a file in terms of source, time, content, and form. Archival entities are mostly classified and shelved in file form based on archival source (organization), time (year), content (subject), and form (type), with content (subject) classification representing a semantic-level classification organization method at the content theme level.

2.2 The Semantic Organization Tradition in Archival Information Organization

Archival information organization targets archival retrieval information, aiming to achieve the ordering of retrieval information. Archival information organiza-

tion has a long history, with classification cards, aperture cards, punched cards, and edge-punched cards all being important tools in manual management environments. The results of archival information organization generally manifest as archival catalogs, indexes, and compilation products—secondary and tertiary literature. Classification and subject methods are the two most important methods of archival information organization, both operating at the semantic level and analyzing archival content for indexing and retrieval. Whether class numbers and class names in classification methods or subject terms in subject methods, they essentially represent expressions of concepts reflected in archives. The archival information classification system is a hierarchical number retrieval system that enumerates known categories and expands them progressively, based on the principles of concept division and generalization, reflecting the subordinate, derivative, and parallel relationships in archival content. To unify archival retrieval classification methods on the basis of systematic collections and break through the limitations of year, organization, and type classification in archival entity classification, the "Chinese Archives Classification" proposes an archival classification system that "takes unified classification principles and notation systems as prerequisites, uses functional division as classification standards and bases, combines systematic classification with faceted classification, and has semi-faceted properties." This principle of "functional division" essentially analyzes the semantics of archival content from the perspective of social functions in which archives participate. On the other hand, the subject method in archival information organization uses natural language words to describe various concepts in archives and arranges these concepts alphabetically. Unlike the hierarchical organization of classification methods, subject methods use faceted organization to reveal archival themes, with standardized natural language words as indexing and retrieval bases, essentially forming an archival subject dictionary. Such subject dictionaries have strong potential to transform into "data dictionaries"—ontologies in the current context.

2.3 The Semantic Organization Tradition in Network Archival Information Organization

Whether content classification in archival entity organization or classification and subject methods in archival information organization, although both represent semantic-level organization of archival content and themes, their categories or subject terms are expressed in artificial marks or natural language—understandable to humans but not machines. Such semantic organization methods primarily target humans rather than machines. Machine-understandable archival organization first emerged in the management of network archival resources. Network archival resource organization requires screening, sorting, describing, indexing, analyzing, storing, and utilizing large amounts of dispersed and disordered archival information on the internet to form systematic structures. Traditional manual and expert-dependent organization methods cannot handle massive network archival resources. Automated organization methods, including automatic classification, automatic indexing, automatic compilation

and management of classification tables and vocabularies, automatic compilation of catalogs, indexes, abstracts, and automatic searching of online information sources, can more effectively process complex multimedia information such as text, graphics, images, and audio-video. Hypertext links can connect these complex resources to form larger resource networks. However, this network organization and association of network archival resources remain at the file level, and hypertext links themselves do not carry semantics. Why network archival resources have such relationships and links still requires users to judge for themselves.

With the exploration of network archival resource organization, more scholars have begun paying attention to the development of the semantic web and using semantic technologies to improve metadata-centered archival resource organization methods. Scholars have most frequently discussed how to use semantic web technologies such as ontologies and linked data for archival information identification, description, and reasoning, solving the heterogeneity problem between archival information and archival information systems, or applying them to digital archives to address issues such as knowledge association, semantic interoperability, cross-media semantic retrieval, and semantic aggregation of digital archival resources.

3. The Connotation of Archival Semantic Organization

Terry Cook once said: "By showing the world how to avoid drowning in a 'sea' of meaningless data and how to explore interconnected meanings or knowledge, we reaffirm the applicability of our profession." This thinking about finding connections has profoundly influenced theoretical research and practical exploration in the archival field over the past two decades. Today, the new wave of semantic technology and data management thinking has brought "transactions" and "relations" back to the core of data organization. Discussing "interconnected meanings or knowledge" in the new computer environment means discussing interconnected data with clear meanings and the knowledge networks formed through associative organization. This is the essential pursuit of archival semantic organization discussed in this paper. Therefore, archival semantic organization must begin with discussing the connotation of data meaning (semantics), the association of data with clear meanings (semantic relations), and how to organize data meanings and associations (semantic organization).

3.1 The Connotation of Archival Semantics

Semantics refers to the meaning of data, which needs to be expressed through formalized languages that follow certain grammatical rules—meanings that machines can understand. Data meanings expressed in natural language must be transformed into formalized language expressions. Archival semantics refers to the meanings of all archival data itself and data describing archives, including the semantics of archival content data, context data, and structural data. This differs from the meanings of traditional archival text content or metadata, as

archival semantics uses formalized language expression, with clear meanings that machines can understand.

3.1.1 The Semantics of Archival Content Archival content refers to the information contained in archives that expresses the author's intent. Archival content is generally expressed in natural language, with meanings depending on the words and syntactic structures used by the author during record creation, relying on certain linguistic systems and contextual relationships that may vary with language and context. Therefore, identifying the semantics of archival content means identifying the correspondence between words in the text and the concepts they refer to. To obtain content semantics, semantic annotation (indexing) methods are generally used, aided by tools such as ontologies to identify concepts in the text, evolving information organization from file units to concept units. The semantics of archival content are concentrated in five aspects: time, person (organization), place, event (subject), and object. These can be obtained by annotating these five elements in the text and disambiguating them. For example, in the Taiwan History Digital Library, a document titled "Li Du Mai Jin Gen Pu Yuan Zi" has semantic entities such as place names like "Jian Dong Shang Bao Qi Fen Zhuang" and personal names like "Wu A Wang" annotated in its text. After formalized encoding and expression, these content semantics can be associated with related semantics in other archives to extract knowledge about the same time, person, place, and event across these archives. The finer the granularity of archival content semantic annotation, the higher the degree of semantic relationship revelation in the content, and the deeper and more effective the discovery, aggregation, and mining of knowledge contained in archival resource text content.

3.1.2 The Semantics of Archival Context Archival context refers to the environment in which archives exist. Any archive lacking context information is incomplete and lacks evidential value. All archives have institutional context, business context, procedural context, and documentary context. The semantics of archival context refer to the meanings of data describing these contexts. Institutional context refers to the organizational system to which the archive creator belongs; business context refers to the business functions, activities, and transactions that generated the archives; procedural context refers to the procedures of record creation, transfer, archiving, and management; and documentary context refers to the relationships between the record and other records within the same fonds or archival collection.

Context-drivenness is an important characteristic distinguishing archival resources from other information resources. Archival resources are not "static" resources. Throughout their lifecycle evolution, context information profoundly influences the content and structure of archival resources. The identification and acquisition of archival context semantics are crucial for establishing relevant relationships between archival texts and their sources, businesses, procedures, and collections, and for expanding external associations at the file level, collection

level, and even fonds level. For example, as shown in Figure 1 [Figure 1: see original paper], the Taiwan History Digital Library (THDL) identifies the role and position of historical official documents in "transmission up and down" based on documentary context semantics, providing an "alternative viewing window" for the relationships between each historical archive and other archives, offering users more browsing recommendations for historical archives.

3.1.3 The Semantics of Archival Structure Archival structure refers to the organization and expression methods of archival content information, where organization methods include main text and attachments, and expression methods include format, carrier, version, etc. The semantics of archival structure refer to the meanings of data describing this structure. In traditional paper archival resources, the semantics of archival content and structure were inseparable. However, with archival digitization and the development of electronic records, archival content and structure have gradually separated, with their mutual influence decreasing. For example, in many information management systems, certain file content data can be filled in while structural data is automatically generated from templates, ultimately presenting documents with fixed structures suitable for human reading. In this context, the semantics of archival structure are more easily identified and obtained independently, generally reflected in the meanings of data describing long-term preservation information of archival resources, including format, version, and carrier data. In the datafication state, structural semantics are essential for defining the lifecycle stage of records. Files of different versions and formats may exist in different institutional, business, procedural, and documentary contexts and may contain different text content information. Therefore, the semantics of archival structure often serve as an important bridge for establishing relationships between archival content semantics and context semantics.

Through the association between archival content, context, and structure semantics, new associations among archives will also be established. As shown in Figure 2 [Figure 2: see original paper], content, context, and structure data at different levels may have inheritance or other relationships in meaning and nature. These associations between semantics at different levels can establish connections between archives at different levels, thereby forming multi-level data and semantic networks, as shown in Figure 3 [Figure 3: see original paper].

3.2 The Connotation of Archival Semantic Relations

Archival semantics are embedded in archival content, context, and structural data, which exist not only in records but also at various levels of archival collections. Content, context, and structural data at different levels may have inheritance or other relationships in meaning and nature. These associations between semantics at different levels can establish connections between archives at different levels, thereby forming multi-level data and semantic networks.

3.2.1 Multi-level Description-triggered Multi-level Associations The concept of multi-level archival description originates from Oliver W. Holmes's five-level archival arrangement theory, which proposes conducting archival resource description and arrangement at five levels: repository, records group, series, file unit, and document. The 1992 Twelfth International Congress on Archives established the "Madrid Principles" for modern archival description: the principle of provenance, the principle of respect for fonds, and the principle of reflecting administrative levels. The principle of reflecting administrative levels requires archival description to fully reflect the hierarchical levels of "fonds-subfonds-category-file." ISAD(G) explicitly proposes conducting multi-level description at four levels: fonds (subfonds)-series (subseries)-document-entity, thereby supporting vertical cataloging and retrieval within fonds and horizontal correlation retrieval between levels. Multi-level description provides multi-layered, multidimensional, and complete archival description, laying the foundation for establishing associations between different archival levels, and also provides multi-level entry points for archival retrieval. Descriptive data at any level can be traced upward or extended downward to obtain higher- or lower-level archival descriptive data, and can also be expanded horizontally to retrieve related archives.

China's "Rules for Archival Description" (DA/T18-1999) only specifies content, context, and structural description items and formats at the document and file levels, with almost no standard data definitions, item references, or format specifications for fonds- and category-level description. Under such a thin hierarchical description system, archival semantics mostly derive from content, context, and structural data at the document level, unable to obtain more semantics from file, category, and fonds levels, and thus unable to establish rich semantic associations across all four levels and between levels.

Therefore, to obtain more complete archival resource semantic associations, multi-level archival resource semantic description work must be completed. This includes: (1) conducting multi-level description on fonds-category-file-document four-level archival resources, setting description items and formats for each level, clarifying inheritance or partial inheritance relationships between higher- and lower-level description items to avoid duplication in multi-level description, while emphasizing the particularities and specialized description of content, context, and structure at different levels; (2) conducting semantic description on multi-level descriptive data at the four levels, analyzing semantic clustering and association relationships of descriptive data in content, context, and structure across different levels.

3.2.2 Multi-level Networks Formed by Multi-level Associations After multi-level description, semantics at the four levels of fonds-category-file-document can be further analyzed, extracted, and associated, thereby establishing semantic association networks in both vertical and horizontal directions: (1) Vertical networks refer to establishing semantic associations between the four

levels of fonds-category-file-document. These associations are based on partial inheritance relationships of content, context, and structural semantics between upper and lower levels. Generally, vertical networks are the main thread of archival semantic networks, inheriting archival cataloging traditions but changing from strict hierarchical tree structures to providing multiple archival description and retrieval expansion/reduction methods including up-down, skip-level, and single-level options. (2) Horizontal networks refer to establishing same-level semantic associations at the four levels, such as fonds-fonds, category-category, file-file, and document-document. These associations are based on correlation relationships in content, context, and structural semantics at the same level. Generally, such correlation relationships require external open resources as association bridges. For example, two different archival collections can establish upstream-downstream functional relationships between their respective institutions through an "institution directory" external resource, thereby further establishing associations in institutional context semantics between institutional fonds and enabling semantic links from one institution's fonds to another's.

Based on these basic vertical and horizontal semantic association networks, more cross-level and cross-collection semantic networks can be established, providing more archival retrieval points and service entry points.

3.3 The Connotation of Archival Resource Semantic Organization

Archival semantic organization refers to the process of clarifying, formally encoding, and linking the meanings of archival content, context, and structural data. It includes two major components: identifying, understanding, analyzing, and expressing archival semantics, and establishing associations between multi-level archival semantics. No semantics can realize value independently. Classifying, clustering, and associating semantics according to their relationships can form concepts or knowledge describing the objective world. Therefore, the essence of semantic organization is analyzing relationships between semantics and establishing associations between them. Based on whether archival semantics originate from within or outside archival collections, archival semantic organization can be divided into inward semantic organization and outward semantic organization.

3.3.1 Inward Semantic Organization The "inward" in inward semantic organization refers to within a specific archival collection. The organized semantics come from data in the same data source, without involving cross-collection or cross-data source association or integration issues. Inward organization is also the process of establishing archival vertical semantic relationship networks. For most historical archives digitized from traditional paper sources, where historical research or public memory construction is the main goal, inward semantic organization primarily refers to semantic organization of archival text content within a collection oriented toward a specific historical research theme or public memory direction. That is, based on associations between person (organization),

time, place, event, or object semantics in text content, it establishes associations between different documents within a collection on a certain theme, forming a complete and detailed content network about that theme. For most born-digital archives, where providing business evidence and supporting business decisions are the main service goals, inward semantic organization primarily refers to multi-level associations of content, context, and structural semantics within "fonds-category-document-component" in a certain institutional archival collection (or institutional fonds). As shown in Figure 4 [Figure 4: see original paper], semantic organization at each level involves associations between content, context, and structural semantics, ultimately constructing a four-level network structure.

3.3.2 Outward Semantic Organization The "outward" in outward semantic organization refers to outside a specific archival collection—that is, transcending collection limitations to search for and associate related semantics from more external data sources, including semantic associations between different archival collections and between archival collections and other types of data collections. Outward semantic organization treats archival collections as components of a larger domain of archival collections, emphasizing integration and associative organization of different archival collections and even different data collections based on domain knowledge, thereby forming descriptions, development, and utilization of broader domain knowledge. Archival semantic outward organization is also the process of establishing archival horizontal semantic association networks. The key to outward semantic organization is establishing machine-understandable links to create associations between data from different data sources, essentially achieving semantic association through semantic integration. For most business-driven electronic archives, outward semantic organization means comprehensively establishing social relationship networks that transcend institutional functions and archival provenance limitations among related institutions, functions, and archival collections. As shown in Figure 5 [Figure 5: see original paper], archival collections from different institutions were originally independent. Due to upstream-downstream relationships between institutions, functions, and businesses, and the generation of corresponding business evidence when these relationships occur, series of archives spanning institutions and functions eventually form, making originally dispersed and independent archival collections related. This requires that when organizing archives, we should not only inwardly organize business activities, transactions, and various archives generated within institutional functions but also outwardly organize related institution directories, function lists, business activity records, transaction logs, behavioral data, etc. By associating with public external resources, we can establish associations between different archival collections in contextual semantics, especially institutional and business semantics, and establish relationships between archives within a collection and archives outside it, forming a complete 梳理 of functions and activities across the entire business domain, ultimately providing archival resource services for the entire business domain rather than

just for a specific function or activity.

4. The Principles of Archival Semantic Organization

Archival semantic organization is the core link in advancing archival datafication, addressing three core issues: Where do archival semantics come from? How are archival semantics associated? How can archival semantics and semantic associations be understood by machines?

4.1 Structured Semantic Sources

Semantics are the meanings of data, and machine-understandable and operable semantics mainly come from structured data strictly defined by data models. For archives dominated by unstructured data, ensuring complete semantic sources hinges on the structurization of unstructured data. Annotating archival texts and describing archival content, context, and structure are the two most important methods, with resulting archival annotation data and archival description data being the two most important types of archival semantics.

4.1.1 Transcription and Semantic Annotation of Archival Content

For many historical resources digitized from traditional paper archives, Optical Character Recognition (OCR) is currently the most commonly used text transcription method. However, OCR-recognized and transcribed data can only support machine recognition and matching of characters and remain non-semantic in terms of data meaning—that is, these transcribed data are still non-semantic. Further annotation of transcribed data is needed, analyzing, ordering, clustering, and associating them at the semantic level. Unlike general archival resource description (metadata) that mainly occurs at the document level, archival resource semantic annotation emphasizes sinking to the item level—that is, conducting detailed description of "things" in archival resource content rather than the archival resources themselves. Existing archival content annotation often uses manual tagging. For example, the U.S. National Archives and Records Administration (NARA) launched the "Citizen Archivist Initiative" in 2011, encouraging citizens to help structure and describe NARA's holdings by adding tags, annotations, and transcriptions. NARA also issued a specialized tagging policy for this purpose. Since 2012, citizens have contributed hundreds of thousands of tags, metadata, transcribed texts, video subtitles, and digital images through this crowdsourcing approach, making important contributions to understanding and describing collection content.

In addition to crowdsourced manual tagging, some automated semantic annotation tools are available for archivists to conduct item-level annotation. For example, the "Communication and Empire: Chinese Empires in Comparative Perspective" project funded by the European Research Council developed the MARKUS automated semantic annotation tool, which currently supports automatic annotation of semantic entities in Chinese and Korean texts, including

names of people, places, times, official titles, and institutional names. It also supports manual and batch annotation of custom keyword lists or tags in all languages. MARKUS has established automatic associations with a series of conceptual models or databases, such as language-specific dictionaries, domain-specific vocabularies, the China Biographical Database (CBDB), and the China Historical Geographic Information System (CHGIS), providing references for annotation concepts.

Whether manual or automated, annotation outputs become important components of archival resource description data. Some annotation data may duplicate or be synonymous with archival resource description items, such as subject terms and subject tags; some may be related to description items, such as historical institution names and record-generating institutions; some may supplement description items with important additions for understanding archival resource content, such as historical personal names and place names. Together, these constitute a database describing archival resources that, after further semantic association, forms an archival semantic unit network.

4.1.2 Archival Metadata Description and Semantic Enhancement
OCR recognition and semantic annotation are generally used for acquiring archival content semantics, while acquiring archival context and structural semantics mainly relies on archival metadata acquisition and semantic description. Metadata is the most commonly used archival description and management tool and the most important source of structured archival data in current information management systems. Metadata (data about data) has the basic function of defining and describing data. The part of metadata that captures the meaning of data corresponds to the term "semantic" in the semantic web discussed in this paper. In the paper archives period, metadata was scattered in file covers, internal catalogs, and other places, requiring repeated recording. In the electronic records or information management system era, archival metadata has basic attributes of structurization, centralization, and standardization, generally expressed in XML format. For example, the "Encoded Archival Description" (EAD) commonly followed in current archival description is an XML-based archival description rule. However, XML format is an implicit semantic expression method and can be considered non-semantic. Therefore, further semantic enhancement of archival metadata is needed to make the semantics contained in metadata explicit.

For example, the large-scale European digital humanities heritage resource project "Europeana," initiated by the EU, stores over 58 million digital cultural resources from more than 3,700 European libraries, archives, museums, and other collecting institutions, all of which have undergone basic metadata description. To optimize retrieval and utilization of massive metadata, Europeana adds new information to metadata at the semantic level—a process called "semantic enrichment." Europeana's metadata semantic enrichment process mainly consists of three stages: (1) analyzing existing metadata sets,

selecting reference datasets, and developing matching and association rules between metadata and reference datasets; (2) matching metadata items and their values with fields and values in reference datasets, and automatically adding inter-data relationships from reference datasets to metadata sets; (3) adding data items and their values that exist in reference datasets but not in current metadata sets to metadata sets, including semantically identical or similar concepts, superclass or subclass concepts, etc. After these three stages of semantic enhancement, archival metadata semantics become explicit, enabling clearer establishment of semantic associations.

4.2 Clarified Semantic Relations

Capturing archival annotation and description data and obtaining semantics from structured data is the first step in archival semantic organization. However, no semantics can exist independently. The connotation and boundaries of archival semantics require more semantic relations to define them, and relations between semantics also determine relations between data. The core of semantic organization is establishing semantic associations to define data associations and form archival data networks. Semantic relations are essentially relations between concepts, determined by the extension of concepts, which in turn further influences the definition of concept intension. Therefore, establishing archival semantic relations means finding relations between concepts contained in archival content, structure, and context data. Ontology is the most suitable tool for completely expressing conceptual systems in archival resources and strictly defining and formally expressing relations between concepts. Ontology, used in the semantic web to describe resource metadata (metadata vocabularies), is a recognized conceptual set in a certain domain, where concepts contain recognized semantics expressed through various associations between concepts.

Using Finland's national-level semantic collective memory platform CultureSampo as an example, the platform hopes to integrate numerous heterogeneous archives through semantic association and build a complete Finnish national memory based on digital archival resources. To this end, CultureSampo first established a national digital resource ontology FinnONTO, semi-automatically transforming commonly used national vocabularies into lightweight ontologies. Through collaboration between domain experts, mappings were established between these cross-domain ontologies, ultimately forming a large national ontology—KOKO. KOKO includes a top-level ontology YSO (defining 20,600 concepts), a museum domain ontology MAO (defining 6,800 concepts), an agriculture and forestry domain ontology AFO (defining 5,500 concepts), an applied arts domain ontology TAO (defining 2,600 concepts), and a photography ontology VALO (defining 1,900 concepts). These ontologies provide a basic framework and basis for CultureSampo to establish associations between annotation data and metadata for various digital resources.

4.3 Formalized Semantic Expression

Semantic expression methods are divided into implicit expression, informal expression, and formalized expression based on whether machines can directly understand them. General archival text content uses natural language expression, an informal semantic expression method understandable to humans but not machines—that is, without semantics. General archival metadata uses XML semantics, still at the grammatical level describing data structure rather than data meaning, thus an implicit semantic expression method that can also be considered non-semantic. Therefore, to make archival semantics machine-understandable, archival metadata must be re-expressed using formalized languages. Formalized semantics is a model-theoretic semantics that defines semantics through "statements" with certain structures and models. RDF triples (one triple is one statement) are a model-theoretic semantic expression method—that is, a formalized expression method. The transformation of archival semantics from implicit to formalized expression methods involves converting archival resource annotation and description data and their semantic description data from XML documents to RDF documents. RDF can be based on XML syntax, providing a pathway for transforming many existing XML-format archival metadata into RDF format.

The UK Archives Hub's linked data project achieves formalized expression of archival semantics through RDF transformation of archival metadata. Archives Hub is a non-profit organization providing cross-searching access to archival description data from 363 institutions across the UK, storing metadata for approximately 1.74 million collection-level archival resources without physically holding any archives. To enable machines to automatically understand archival data semantics and intelligently serve user retrieval needs, Archives Hub launched the Locah (Linked Open Copac Archives Hub) project, exploring basic steps for formalized expression of archival semantics: (1) constructing an archival linked data model (ontology); (2) searching for and reusing existing vocabularies (or ontologies) to fill missing concepts in the archival linked data model; (3) adding URIs to archival metadata; (4) transforming archival metadata from EAD data into RDF XSLT stylesheets; (5) publishing archival linked metadata and creating linked data views; (6) using SPARQL language for data semantic association, etc. Transforming archival metadata from EAD format into RDF XSLT stylesheets is the key step for explicitly describing metadata semantics. RDF XSLT stylesheets can encapsulate the data model constructed in step (1), providing a simple, standardized, reusable solution for formalized transformation of archival metadata into linked data.

In summary, through semantic annotation and metadata semantic enhancement, archival content, context, and structure are transformed into structured data with hidden semantics. Clarified conceptual models establish semantic relations between these data, which are then expressed through machine-understandable formalized languages to construct a machine-understandable data network rich in structure, semantics, and relations. This is the basic

principle of archival semantic organization. Moving from basic principles to implementation for different types of archival semantic organization requires further exploration using more semantic technologies and tools based on actual archival resource states and business scenarios.

5. Basic Principles of Archival Semantic Organization

Archival semantic organization may have different practical pathways under different archival resource states and business scenarios, but all should follow basic principles that reflect both distinctions from and connections to semantic organization of other information resources. On one hand, different resource characteristics lead to different semantic source data, requiring archival semantics to completely derive from content, context, and structure at all levels. On the other hand, archival semantic organization inherits common methods from information resource semantic organization, but applying common methods to characteristic resources requires respecting both archival resource characteristics and basic rules of semantic organization.

5.1 Principle of Semantic Integrity

The first step in archival semantic organization is acquiring semantics, during which the principle of semantic integrity should be followed: (1) All levels of archives—including fonds, category, file, and document—should be completely described, with inheritance relationships in content, context, and structure between upper and lower levels fully considered. (2) Description items and formats at all levels should follow certain standards to fully structure archival content, context, and structural data, and to fully semantically describe structured data, including semantic content, structure, format, and relations of all data to support annotation and abstract definition of data meanings. (3) When analyzing archival semantic associations, both inward and outward semantic organization should fully consider relations within content, context, and structure semantics respectively, as well as relations between content, context, and structure semantics. (4) Depending on different service objects and purposes, the center of archival semantic organization frameworks can be either content semantics or context semantics, with structural semantics generally associated around content or context semantics. For example, when oriented toward historical or humanities research, the center of archival semantic organization should be content semantics, with context and institutional semantics providing associative support for understanding content semantics and establishing relevant relations. When oriented toward business support, the center should be context semantics, with content and structural semantics providing support for institutional, business, procedural, and documentary contexts.

5.2 Principle of Chain Association

The principle of chain association means that when establishing associations between archival content, context, and structural semantics, the chain logic within archival content, context, and structure should be respected and followed. The "chains" in this principle include: (1) Content logic chain: Associations of archival content semantics mainly follow content logic chains, including changes in and relationships between the five elements of time, place, person, event (subject), and object involved in archival content, and changes in these relationships. Common examples include single-level logic such as timelines, location changes, person networks, event narratives, and object changes, as well as double-level or multi-level logic such as location changes based on timelines, person networks based on timelines or location changes, and object changes based on event narratives. These content logics often conform to scientific logic in historical or humanities research. Content logic chain-based archival semantic organization centers on content semantics. (2) Institutional function chain: Outward semantic organization between different institutions and different archival collections can analyze relationships between archival collections based on functional relationships between institutions. Therefore, semantic association based on institutional function chains mainly occurs at the archival collection level, centering on context semantics—especially institutional context semantics—to find more relevant relations between content, context, and structural semantics across different archival collections. (3) Business process chain: Semantic associations between documents generated in the same business process can follow the business process chain—that is, the position and role of business activities in which documents were generated within the business flow determine relationships between documents. In increasingly complex business environments, a business process chain may be completed within one department, may span multiple departments of one institution, or may even span multiple institutions. Therefore, semantic association based on business process chains may occur between different documents within the same file series, between different file series within the same archival collection, or between certain documents in different archival collections. Business process chain-based archival resource semantic association centers on context semantics—especially business context semantics—to find more relevant relations. (4) Records lifecycle chain: The records lifecycle endows records with dynamism, manifested as a procedural chain including record creation, transfer, archiving, long-term preservation, and utilization. This procedural chain explains from a management perspective the position of records within institutional functions and business processes, and electronic information systems can leave data traces for these procedures. The records lifecycle chain is crucial for establishing relationships between content, context, and structural semantics of a record, especially relationships between records, responsible parties, and system environments. (5) Combinations of the above four chain logics: For archival semantic organization, both inward and outward organization are important. Combining content logic chains, institutional function chains, business process chains, and records lifecycle chains for

semantic organization can discover more semantic centers and relationship networks, achieving multi-level networked associations across archival collections, between file series within collections, between documents within file series, and between data within documents.

5.3 Principle of Network Multidimensionality

The principle of network multidimensionality means that the archival data network formed through archival semantic organization should be a multi-dimensional network without a single center, capable of supporting multi-angle, multi-dimensional retrieval queries and intelligent services. While inheriting and developing traditional hierarchical organization methods that emphasize hierarchical relationships, archival semantic organization places greater emphasis on network organization of correlation relationships. In this network, there is no single center; instead, any semantic unit in content, context, or structure at different levels can become a center, radiating outward according to chain association principles to establish associations with other semantic units. The "multidimensionality" in this principle includes two basic dimensions and one cross dimension: (1) Horizontal data networks based on associations between content, context, and structural semantics at the same level; (2) Vertical data networks based on associations between content, context, and structural semantics across upper and lower levels; (3) Cross-level data networks based on associations between content, context, and structural semantics at different levels.

Establishing multi-dimensional data networks can provide users with different retrieval points, support expansion, reduction, and modification of retrieval across same-level, upper-lower-level, and skip-level archives, and enable flexible and varied visualization of scope and structure of archival resources. Multi-dimensional, decentralized network organization is the basic concept of archival semantic organization and the fundamental characteristic of the semantic-rich data network ultimately constructed by archival resource semantic organization.

Archival semantic organization is not a completely new process starting from scratch. Existing archival information organization tools, methods, and achievements can provide a good foundation for archival semantic organization. For example, existing archival description rules can support semantic description after being ontologized; existing archival description and annotation data can support the establishment of archival linked datasets after formalized expression. These are concrete pathways for archival semantic organization at the practical level. However, overall, archival semantic organization must still undergo the processes of meaning clarification, formalized encoding, and associative linking. More semantic technologies that may emerge in the future will bring new methods to archival semantic organization, but the principles of archival semantic organization should always be observed. Only by respecting the characteristics of archival resources and the professionalism of archival management can more implementation solutions for archival semantic organization be explored.

References

[1] Qian Yi. Preliminary Exploration of the Evolution of Archival Object Management Space Under Technological Change[J]. Archives Science Bulletin, 2018(2): 10-14.

[2] Zhao Yue. Prospects for Archival Datafication in the Big Data Era: Significance and Challenges[J]. Archives Science Study, 2019(5): 52-60.

[3] Negroponte Nicholas. Being Digital[M]. New York: Vintage Books, 1996.

[4] Jiang Hao. Datafication: Intelligence from the Inside Out[M]. Beijing: Communication University of China Press, 2017.

[5] Tao Jun. The Evolution of Vocabulary Semantic Organization Research (1998-2018)[J]. Library and Information Service, 2018(21): 140-148.

[6] Tan Guoxin, Hou Xilong, Zhuang Wenjie. Research on Semantic Organization of Intangible Cultural Heritage Multimedia Resources[J]. Library Science Research, 2017(24): 44-54.

[7] Wang Yuefen, Xing Mengting. Research on Semantic Organization of Social Public Opinion Information for Government Decision-Making Needs[J]. New Technology of Library and Information Service, 2016, 32(7): 21-31.

[8] Ma Yumeng, Guo Jinjing, Wang Chuang. Research on Semantic Organization Model Framework for Scientific Data in e-Science Environment[J]. New Technology of Library and Information Service, 2015(7): 48-57.

[9] Yuan Yan. Research on Knowledge Discovery in Scientific and Technical Reports[J]. Library World, 2017(5): 82-84.

[10] Ding Heng, Lu Wei. Design and Implementation of Standard Literature Knowledge Service System[J]. Data Analysis and Knowledge Discovery, 2016, 32(7-8): 120-128.

[11] Bi Chuanlong. Digitization of Folk Culture Resources in the Big Data Era[J]. Ethnic Arts Research, 2016(3): 87-93.

[12] Ma Yinyuan. SWOT Analysis and Strategies for Linked Data Application in Archival Knowledge Services[J]. Archives and Construction, 2017(2): 17-20.

[13] Qiu Li. Exploration of Building Archival Knowledge Service Models in the Post-Custodial Era[J]. Yunnan Archives, 2015(9): 52-55.

[14] Lin Zhoujia. Research on Semantic-Level Retrieval Technology for Archives[J]. Archives and Construction, 2007(9): 26-27.

[15] Ren Yan, Pang Yufei, Jing Xin. Research on Semantic Organization and Service of Omnimedia Archival Information Resources[J]. Archives Management, 2019, 237(2): 37-38.

[16] Volchinkov. Organization of Soviet Archival Work (Report at the National Archival Work Conference on December 22, 1956)[J]. Archives Work, 1957(2): 5-9.

[17] He Baorong, Chen Zhaoyu, Song Shiqin. Lecture on Basic Knowledge of Archival Work (Outline)—Chapter 4: Arrangement of Archives[J]. Archives Work, 1980(4): 27-33.

[18] Zhou Ming. Different Paths to the Same Goal: Research on Archival Classification and Subject Methods[J]. Sichuan Archives, 2000(1): 12-14.

[19] Deng Shaoxing. The "Chinese Archives Classification" is an Archival Classification System with Chinese Characteristics[J]. Shanxi Archives, 2003(4): 15-19.

[20] Zeng Na. Research on Network Archival Information Resource Organization[J]. Archives Science Bulletin, 2010(1): 45-49.

[21] Zhao Yi. Organization Methods for Network Archival Information Resources[J]. Scientific and Technical Documentation Management, 2007(1): 28-30.

[22] Li Haijun. Discussion on Technical Framework for Transforming Archival Information into "Archival Knowledge"[J]. Shanxi Archives, 2007(9): 20-23.

[23] Wang Lancheng. Development of Archival Information Organization and Retrieval Under Knowledge Integration Environment[J]. Archives Science Study, 2008(5): 45-50.

[24] Lv Yuanzhi. Research on Knowledge "Association" Organization of Digital Archival Resources[J]. Archives Science Study, 2012(6): 46-50.

[25] Lv Yuanzhi. Research on Semantic Interoperability Implementation of Digital Archival Resource Systems[J]. Archives Science Bulletin, 2013(5): 53-57.

[26] Lv Yuanzhi. Research on Implementation Framework and Key Issues of Cross-Media Semantic Retrieval for Digital Archival Resources[J]. Archives Science Study, 2014(2): 65-70.

[27] Lv Yuanzhi. Research on Implementation Strategies for Cross-Media Semantic Association Aggregation of Digital Archival Resources[J]. Archives Science Study, 2015(5): 60-65.

[28] Proceedings of the Thirteenth International Congress on Archives[C]. Beijing: China Archives Press, 1997.

[29] Liang Menghua. Research on Cross-Media Knowledge Linking of Digital Archival Resources Based on Open Linked Data[J]. Archives Science Study, 2015(4): 111-116.

[30] Feng Huiling. 100 Questions on Electronic Records Management[M]. Beijing: Renmin University of China Press, 2014.

[31] Rong Juntao. Research on Semantic Description Model for Knowledge Elements in Academic Literature Content[J]. Information Science, 2019(7): 30-35.

[32] Du Xiechang, Xiang Jie. Taiwan History Digital Library[EB/OL].[2021-02-17]. http://doi.org/10.6681/NTURCDH.DB_{THDL}/Text.

[33] He Defang, Zeng Jianxun. Research on Deep Aggregation of Collection Resources Based on Semantics[J]. Journal of Library Science in China, 2012, 38(4): 79-87.

[34] DURANTI L. The archival bond[J]. Archives and Museum Informatics, 1997, 11: 213-218.

[35] Du Xiechang, Xiang Jie. Taiwan History Digital Library[EB/OL].[2021-02-17]. http://thdl.ntu.edu.tw/THDL/RetrieveSVG.php?filename=ntu-2252926-0080500806-0000840.txt.

[36] OLIVER W H. Archival arrangement-Five different operations at five different levels[J]. The American archivist, 1964, 27(1): 21-37.

[37] Zhang Zhengqiang, Bian Gang. Principles and Theory of Modern Archival Description[J]. China Archives, 1999(10): 39-41.

[38] ISAD(G): General international standard archival description[S]. Second edition. ICA, 1999: 36.

[39] Ma Yinyuan. Comparative Research on Multi-level Archival Description at Home and Abroad[J]. Archives Science Study, 2017(02): 53-58.

[40] DAMA International. DAMA Guide to the Data Management Body of Knowledge[M]. Translated by DAMA China Branch. Beijing: China Machine Press, 2020.

[41] National Archives. Citizen archivist dashboard[EB/OL].[2021-02-27]. https://www.archives.gov/citizen-archivist.

[42] Citizen archivist dashboard. Citizen contribution policy[EB/OL].[2021-01-27]. http://www.archives.gov/citizen-archivists/resources/tagging-policy.

[43] ANDREW W. Citizen archivist dashboard/Improving access to historical records through crowdsourcing[EB/OL].[2021-01-27]. https://www.citizenscience.gov/citizen-archivist/#.

[44] MARKUS[EB/OL].[2021-01-27]. https://dh.chinese-empires.eu/markus/beta/.

[45] HUGO M. Europeana semantic enrichment framework[EB/OL].[2021-02-01]. https://docs.google.com/document/d/1JvrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEEx0Y/edit.

[46] HYVÖNEN E, VILJANEN K, TUOMINEN J, et al. Building a national semantic Web ontology and ontology service infrastructure-the FinnONTO approach[A]//The semantic Web: research and applications. Berlin: Springer, 2008: 95-109.

[47] WIKIPEDIA. Archives hub[EB/OL].[2021-01-28]. https://en.wikipedia.org/wiki/Archives_{Hub}.

[48] Linking Lives. About locah[EB/OL].[2021-01-28]. http://linkinglives.archiveshub.ac.uk/about-locah/.

[49] ADRIAN S. Final product post: Archives Hub EAD to RDF XSLT stylesheet[EB/OL].[2021-01-28]. http://locah.archiveshub.ac.uk/tag/linkeddata/.

Author Contributions:

Qi Tianjiao: Responsible for paper writing;
Feng Huiling: Paper supervision.

The Connotation, Characteristics and Principles Analysis of Semantic Organization in the Process of Archival Datafication

Qi Tianjiao, Feng Huiling
School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [Purpose/significance] In the stage of archival datafication, archival utilization and services need to meet user needs at the data level, breaking through the limitations of page-level reading and file-level utilization. This requires building a new semantic-level archival organization model during the organization phase to achieve fine-grained analysis and mining of archival content, context, and structural data, and to prepare resources, methods, and technologies for archival value-added development and intelligent knowledge services. [Method/process] Using literature investigation and case analysis methods, and based on the characteristics of the archival datafication stage, this paper analyzes the basic connotation of archival semantics, semantic relations, and semantic organization, compares the differences and characteristics between archival and other information resources in semantic organization processes, and explores the theoretical framework of inward and outward archival semantic organization under the principles of semantic integrity, chain association, and multidimensional networks. [Result/conclusion] Archival semantic organization is carried out based on the meaning and association of data, aiming to discover semantic relations from the content, context, and structural data of archives. Archival semantic organization is the key link to realizing archival datafication and the crucial step toward making archives machine-understandable and machine-operable. Through archival semantic organization, originally scattered, distributed, and field-dependent archival content, context, and structural data can be clearly defined, formally expressed, and associatively linked. Archival data thus becomes machine-understandable and machine-operable, enabling automated associative organization, preservation, and utilization of archives, which ultimately supports intelligent acquisition, utilization, and services of archival resources based on human-machine and machine-machine interaction.

Keywords: archives; archival datafication; semantics; semantic relation; semantic organization

Statement of Academic Integrity for Authors Submitting to *Library and Information Service*

Library and Information Service has always upheld the mission of publishing excellent academic paper achievements and promoting academic exchange in the industry, and is committed to purifying the academic publishing environment and creating a good academic ecology. In 2013, we took the lead in formulating, publishing, and implementing the "Joint Statement of Library Science Journals on Abiding by Academic Ethics and Purifying the Academic Environment" (hereinafter referred to as the "Statement") (see: http://www.lis.ac.cn/CN/column/item202.shtml). Subsequently, we also took the lead in formulating and publishing the "Joint Action Plan for Chinese Library and Information Science Journals to Resist Academic Misconduct" (hereinafter referred to as the "Joint Action Plan") (see: http://www.lis.ac.cn/CN/column/item247.shtml). To implement and realize this philosophy, this journal hereby solemnly declares that from this date forward, all submitting authors must commit that: papers submitted to this journal must comply with the above "Statement" and "Joint Action Plan," consciously uphold academic ethics, and resolutely resist academic misconduct. *Library and Information Service* maintains zero tolerance for all forms of academic misconduct such as plagiarism and appropriation, and will adopt corresponding punitive measures.

Library and Information Service Magazine

Note: Figure translations are in progress. See original paper for figures.