# Construction of Fine-grained Aggregation Ontology for Web Academic Documents (Postprint)

**Authors:** Ma Cuichang, Cao Shujin

**Date:** 2023-07-26T00:00:00+00:00

## Abstract

[Purpose/Significance] This study aims to explore the theories and methods for constructing fine-grained aggregation ontologies for web-based academic documents.

[Method/Process] Based on a review of relevant theories and methods, this paper first clarifies fundamental theoretical issues such as the basic types, granularity characteristics, and definitions of fine-grained aggregation ontology concepts. Subsequently, taking the "citation analysis" topic corpus in the field of library and information science under the web environment as the data source, it conducts a detailed discussion on the construction of fine-grained aggregation unit ontology from the perspectives of concepts, attributes and relationships, and instances, and performs evaluation and discussion of the ontology.

[Results/Conclusions] This paper proposes for the first time the idea and method of constructing fine-grained aggregation ontology based on the knowledge system of aggregation units, which can provide reference for the construction of knowledge organization system tools in fine-grained organization, retrieval, and navigation based on aggregation units.

## Full Text

## Preamble

**Research on the Construction of Fine-grained Aggregation Ontology for Web-based Academic Documents**

*Ma Cuichang*[1], *Cao Shujin*[2]
[1]Sun Yat-sen University Library, Guangzhou 510275
[2]School of Information Management, Sun Yat-sen University, Guangzhou 510006

## Abstract

[**Purpose/Significance**] This study aims to explore the theories and methods for constructing a fine-grained aggregation ontology for web-based academic documents. [**Method/Process**] Building upon relevant theories and methods, we first clarify fundamental theoretical issues including the basic types, granularity characteristics, and definitions of fine-grained aggregation ontology concepts. Using the "citation analysis" topic corpus in library and information science under the web environment as our data source, we systematically examine the construction of fine-grained aggregation unit ontology from the perspectives of concepts, attributes, relationships, and instances, followed by ontology evaluation and discussion. [**Result/Conclusion**] This paper proposes, for the first time, an approach and methodology for constructing fine-grained aggregation ontology based on an aggregation unit knowledge system, which can inform the development of knowledge organization system tools for fine-grained organization, retrieval, and navigation based on aggregation units.

**Keywords:** information aggregation; ontology; web documents; library and information science
**Classification Number:** G250
**DOI:** 10.13266/j.issn.0252-3116.2019.24.012

## 1 Research Background

Current search engines still primarily organize massive web resources at the level of websites, web pages, and various file carriers as the main units of control and organization. However, functions for searching and locating local information units within web pages have emerged, such as Baidu's ability to refine encyclopedia entry search results down to specific knowledge points. Nevertheless, for domain users, their needs for web documents are often dispersed across different genre types depending on search task scenarios. Consequently, we still face the challenge of achieving fine-grained aggregation of multi-type web academic documents—namely, how to enable network information systems to present screened, extracted, and sequenced multi-type web resources (either whole or partial) to users based on their explicitly expressed information needs and retrieval contexts (such as tasks and user preferences). This requires controlling aspects like aggregation unit types, granularity, relationships, and attributes to more flexibly and accurately present target information that meets users' needs for web academic resources.

From an information organization perspective, implementing fine-grained aggregation of web academic documents necessitates constructing a domain-approved knowledge organization system that reflects the hierarchy and types of information units within multi-type web documents and their relationships with user needs. This system must formally define the concepts, relationships, and inference rules between concepts within the knowledge organization system, enabling web resources to effectively express machine-recognizable semantic concepts and

providing a semantic foundation for the extraction, organization, association, and retrieval matching of fine-grained aggregation units.

In view of this, this paper aims to explore the theories and methods for constructing a fine-grained aggregation ontology for web academic documents based on the concept of aggregation units. An aggregation unit, as the basic object of fine-grained aggregation, refers to the fundamental text content unit controlled and processed by the system when fine-grained aggregation serves as the method of information organization and retrieval. It is a collective term for language function units at different levels divided according to the genre structure of web academic documents. Aggregation units for web academic documents can be either entire documents or parts of web resources, such as paragraph units in the conclusion/discussion sections of research papers, or sentence group units like "proposing suggestions for future research" within conclusion/discussion sections.

To achieve our research objectives, this paper first reviews relevant theories and methods, proposes a theoretical framework for fine-grained aggregation ontology, constructs the ontology through empirical research, and conducts evaluation and discussion.

## 2 Theoretical and Methodological Foundations

Knowledge unit theory is closely related yet distinct from the construction of fine-grained aggregation ontology. Knowledge unit theory provides a theoretical basis for organizing information units within resource carriers and offers foundational theories and methods for constructing knowledge organization systems based on knowledge elements. However, knowledge unit theory defines knowledge organization objects only as information content containing knowledge units, whereas fine-grained aggregation ontology defines knowledge organization objects as aggregation units defined by language functions. The organization of aggregation units involves not only knowledge unit organization but also semantic relationships arising from these units' language functions and their associations with user search forms. Therefore, knowledge unit theory cannot cover all issues required for constructing knowledge organization systems for fine-grained aggregation and must be combined with genre theory and methodology—the theoretical foundation of aggregation unit construction—to establish a theoretical and methodological basis for building aggregation unit knowledge systems. Additionally, the construction of fine-grained aggregation ontology relies on fundamental theories and methods of ontology. Consequently, this study synthesizes relevant theories on knowledge units, genre structure rules, and ontology construction to build a theoretical framework adapted to the needs of fine-grained aggregation ontology construction.

### 2.1 Research on Knowledge Unit Theory

Research on knowledge units within documents has attracted attention across various disciplines, but academic consensus on knowledge units remains unestablished. Definitions of knowledge unit concepts vary across different periods and disciplines, encompassing terms such as "knowledge gene," "knowledge concept," "knowledge node," "knowledge factor," "knowledge point," "knowledge element," "knowledge link," and "knowledge unit." All refer to knowledge content with independent meaning within a certain unit, which can be a document, a fragment of a document, or a conceptual knowledge point contained within a document [1]. Research on knowledge units can be categorized into three types: domain-oriented knowledge unit research for knowledge organization, knowledge unit research for knowledge extraction and utilization, and curriculum-oriented knowledge unit organization research in education.

**2.1.1 Domain-Oriented Knowledge Unit Research for Knowledge Organization** According to Wen Tingxiao, knowledge units can be divided into three main forms based on the development stages and depth of knowledge organization: document units, information units, and knowledge units. Knowledge unit practice and research began with document units. As natural carriers containing knowledge, document units naturally became the initial units for knowledge management, gradually forming a complete knowledge system. Thus, document units, as early forms of knowledge units, contain knowledge units, while knowledge units ultimately attach to certain forms of document units and manifest as document units [1].

Related research emphasizes that knowledge organization objects should penetrate to the level of knowledge units within documents, thereby compensating for the deficiency of existing knowledge organization theories in adequately reflecting document content. However, no further theoretical or methodological guidance exists regarding the granularity and hierarchy of knowledge units within documents, failing to meet the requirements of aggregation unit division for fine-grained aggregation knowledge organization systems.

**2.1.2 Knowledge Unit Research for Knowledge Extraction and Utilization** Wen Youkui and colleagues systematically proposed "knowledge element" theory from the perspective of knowledge organization: "It is assumed that the organizational arrangement of text content consists of a logical ordering structure of independent knowledge elements, which are called knowledge elements, and the logical dependency relationships are called knowledge chains. Knowledge elements are the basic units for constructing knowledge structures. Knowledge elements and their structures form different knowledge units" [2]. Knowledge element types include descriptive types (information reports, term explanations, numerical values, literature citations, etc.) and process types (procedures, methods, definitions, principles, experiences, etc.) [2]. Wen Youkui and others conducted systematic research on knowledge element theory and its

knowledge organization methods [2-9], upon which CNKI built a search system for knowledge elements such as definitions, numbers, and charts in academic papers [10].

Evidently, the knowledge element concept in knowledge element theory originates from the refinement of knowledge organization object granularity, which is similar yet different from the aggregation unit concept in this study's fine-grained aggregation ontology. The similarity lies in that both organize objects from the text level to the text content level. The difference is that knowledge element theory's classification of knowledge element types (such as definitions, numerical values, charts, etc.) focuses on knowledge organization and utilization, aiming to build knowledge bases based on knowledge instances and their relationships. In contrast, this study's classification of aggregation unit types focuses on the organization and utilization of useful information fragments, aiming to construct associations between information fragments and between information fragments and user task scenarios. Therefore, research on knowledge element definition, extraction, ontology construction, and organization can provide numerous methodological foundations for aggregation unit extraction and organization, as well as references for aggregation unit ontology construction.

**2.1.3 Curriculum-Oriented Knowledge Unit Organization Research**
This research area is mainly concentrated in educational technology, where knowledge units are often called "knowledge points"—basic units for transmitting teaching information during teaching activities, including theories, principles, concepts, definitions, examples, and conclusions. Knowledge points can be further decomposed; those that are structurally indivisible are called atomic knowledge points. Related groups of knowledge points integrate into knowledge units. The basic principle of knowledge point division is to ensure local integrity of knowledge content, and their sizes can vary greatly depending on needs. For example, a chapter can be divided into a large knowledge point, a section within it can be subdivided into smaller knowledge points, and definitions or theorems within a section can be divided into even smaller knowledge points. Some scholars have built educational technology discipline resource libraries based on knowledge units using the discipline's knowledge classification system [11]. Additionally, some scholars have proposed knowledge element description models for knowledge organization and sharing of educational resources based on knowledge element theory, exploring related methods and technologies for knowledge abstraction and fusion [12].

Knowledge system construction in the education field actually aims at curriculum knowledge organization and education rather than resource organization. Therefore, its knowledge element instances are the knowledge itself, not resources corresponding to the knowledge concepts. However, methods and technologies for knowledge element ontology construction can provide references for constructing ontology concepts and relationships based on aggregation unit knowledge systems.

### 2.2 Research on Genre Structure Rules and Knowledge Systems

Genre presents certain formal features and structural rules according to its communicative goals. Although most genre theory research uses typical genres such as scientific papers, news, or short stories as case studies, genre structures are now recognized as existing in all communication networks, and most professional communication typically relies on genre structures rather than external connections to accomplish their shared work [13]. Research on genre structure rules can be summarized in two aspects: exploration and development of Swales' model in linguistics, and utilization and exploration of genre structures in library and information science.

**2.2.1 Exploration and Development of Swales' Model in Linguistics** Representative genre structure theory includes J.M. Swales' "move-step" analysis model for academic papers. Based on the introduction-methods-results-discussion structure characteristic of research papers, this model further conducts move-step analysis of the "introduction" section content according to research paper objectives, thereby dividing research papers into information units composed of different granularity levels: component-move-step [14].

Building upon Swales' initial model, numerous scholars have applied genre analysis theory and methods to test natural sciences, biomedicine, social sciences [14-16], wildlife behavior research, and ecological conservation fields [17]. The Swales model has also been extended to other components of research papers, including studies on abstracts [18], methods [19], results [20], discussions [21], and all components [21-22]. B.A. Lewin et al. conducted comprehensive studies on moves and steps in introduction and discussion sections of social science corpora [23], testing and enriching the genre structure knowledge system of research papers.

Additionally, domestic scholars have researched genre structures of discourses. For example, Zhao Fuli studied the move structure of TV news leads referring to Bhatja's move model [24]; Ge Dongmei and Yang Ruiying researched academic paper abstracts referring to Bhatja's move model [25]; Cui Yanyan and Wang Tongshun studied the structure of English academic lectures referring to the Swales model [26]; Yang Ruiying conducted move and step analysis of English for Academic Purposes academic paper components using the Swales model and proposed components, moves, and steps for theoretical research academic papers [27].

**2.2.2 Utilization and Exploration of Genre Structures in Library and Information Science** Since the emergence of digital library projects in the late 1990s, issues of digital document deconstruction and reorganization, identification and utilization of information units in digital resources have attracted scholarly attention [28]. Most research on digital document division and utilization is based on genre structure theory, combined with users' information acquisition tasks for division and association analysis.

For example, A. Dillon tested people's cognitive understanding of the genre structure of online news, providing a basis for genre-based understanding and within-text navigation [29]. A. Dillon and colleagues also conducted a series of studies from a user cognition perspective on user disorientation problems and navigation needs around discourse logical and semantic structures [29-34]. L. Zhang constructed a functional unit classification system for psychology journal papers, identifying minimal information units within journal article components (e.g., introduction, methods, results, and discussion) [35-36] and tested the efficiency and effectiveness of reading information acquisition [37]. C.C. Ma and S.J. Cao constructed a cross-genre aggregation unit classification system for the library and information science field [38].

Existing research shows that genre structure division oriented toward user cognition and needs plays an important role in improving information utilization efficiency and effectiveness, providing theoretical support for fine-grained aggregation. More importantly, both linguistics and library and information science research on genre structure division and utilization have formed a series of knowledge systems about language function units under genre structures. However, current research on information unit utilization remains exploratory, without further consideration from a knowledge organization perspective, and without constructing corresponding knowledge organization systems to support practical applications.

### 2.3 Ontology Construction Theory and Methods

Existing ontology research provides direct theoretical and methodological foundations for fine-grained aggregation ontology construction. The following aspects are examined: ontology types, construction principles and methods, construction tools, and evaluation methods.

**2.3.1 Ontology as Formal Specification of Shared Conceptual Models** Numerous ontology studies exist, particularly abroad, with many research organizations and institutions establishing various types of ontologies according to their needs. Based on application scope and level, ontologies can be divided into general ontologies, domain ontologies, and application ontologies. General ontologies are not targeted at specific domain knowledge and can be reused across domains; domain ontologies express knowledge systems of specific disciplines; application ontologies are created for specific applications and can include cross-disciplinary knowledge. General and domain ontologies serve as upper-level ontologies for application ontologies [39-41].

More specifically, R. Mizoguchi et al. proposed classifying ontologies according to their application purposes into domain ontologies, top-level (general) ontologies, and task ontologies. Task ontologies express concept classes, attributes, and relationships specific to particular tasks through top-level concepts, describing conceptual systems in specific tasks or behaviors and providing concept sets that can answer questions related to specific tasks or behaviors [41]. N. Guarino

proposed classifying ontologies based on the specificity of ontology concepts and their independence from domains. According to the level of detail of ontology concepts, they can be divided into more detailed reference ontologies and more concise shared ontologies; according to the independence of ontology concepts from disciplines, they can be divided into four categories: top-level ontologies, domain ontologies, task ontologies, and application ontologies [42-43]. Additionally, some scholars classify ontologies into five types based on application: domain ontologies, general (common sense) ontologies, knowledge ontologies, linguistic ontologies, and task ontologies [44-45].

**2.3.2 Mature Theories and Methods for Ontology Construction** Among these, typical ontology construction principles include T. Gruber's five principles of ontology construction: clarity, coherence, extensibility, neutrality, and minimal ontological commitment [46]. Typical construction methods include: skeleton method, IDEF5 method, seven-step method, five-step cycle method, METHONTOLOGY method, TOVE method, KACTUS method, SENSUS method, and cyclic acquisition method [45].

**2.3.3 Protégé as an Important Ontology Construction Tool** Protégé features a visual user interface, supports DAML+OIL and OWL languages, enables modular design [47], and allows external modification using ontology description languages. Due to its open-source code, Chinese version availability, and other advantages, Protégé has been widely adopted in China [47].

**2.3.4 Diverse Ontology Evaluation Methods** These include user evaluation, application evaluation, corpus evaluation, expert evaluation, composite indicator evaluation, and gold standard evaluation. While these methods have their applicability and operability, research indicates they also have limitations, with less-than-ideal cross-domain applicability and difficulty in large-scale application. Currently, constructing indicator systems is the most common evaluation method [47]. Additionally, scholars note that ontology evaluation method research remains exploratory both domestically and internationally, lacking widely recognized evaluation theoretical and methodological systems. Evaluation focuses on concepts, attributes, and relationships, without creating comprehensive ontology evaluation systems or authoritative evaluation standards [47].

## 3 Construction of Fine-grained Aggregation Ontology

### 3.1 Theoretical Framework of Fine-grained Aggregation Ontology

**3.1.1 Basic Types of Fine-grained Aggregation Ontology Concepts** Ontology construction generally involves two stages: the first stage aims to determine the ontology's concept set and establish a core terminology set; the second stage aims to determine relationships between concepts. Based on the objectives and framework of web resource fine-grained aggregation, this paper

defines four basic types of concepts for fine-grained aggregation ontology: web documents, aggregation units, domain concepts, and task scenarios, as shown in Figure 1 [Figure 1: see original paper].

Based on this, the approach to constructing fine-grained aggregation ontology is: determine the aggregation unit concept set through the aggregation unit knowledge system; determine the domain concept set through methods for constructing domain knowledge organization systems; determine the task scenario concept set through existing research. Since domain ontology construction has extensive research and task scenario ontologies are relatively simple, this paper focuses on exploring ontology construction methods based on aggregation unit knowledge systems.

**3.1.2 Granularity Characteristics of Fine-grained Aggregation Ontology** From existing knowledge organization theory, we know that the level of detail in knowledge organization systems affects web resource retrieval and utilization efficiency—the finer the knowledge granularity, the higher the descriptive accuracy; the smaller the information fragment, the higher the retrieval relevance. Therefore, based on the two important knowledge bases in the web resource fine-grained aggregation ontology framework—the domain knowledge system and aggregation unit knowledge system—this paper defines granularity characteristics for web resource knowledge organization systems by dividing them into domain concept granularity and controlled unit granularity, as shown in Table 1 .

The clarity of granularity levels between domain knowledge systems and aggregation unit knowledge systems helps improve the efficiency and effectiveness of web resource fine-grained aggregation from two aspects: domain knowledge accuracy and web academic document accuracy/relevance.

**3.1.3 Definition of Fine-grained Aggregation Ontology Concepts** For fine-grained aggregation ontology, since previous research has established an aggregation unit knowledge system [38] and investigated the associations between tasks and information units [36,49], its core terminology and attributes can be relatively easily determined and formally specified qualitatively.

This paper defines the domain concept C as a quadruple:

$C = \{C_0, PC, RC, SynC\}$

where $C_0$ represents the domain concept; PC represents general attributes of the domain concept; RC represents the set of domain concept relationships; SynC represents the synonym set of domain concept $C_0$.

This paper defines the task scenario T as a triple:

$T = \{T_0, CT, PT, RT\}$

where $T_0$ represents the task type concept; CT represents the domain concept corresponding to the task theme; PT represents general attributes of the task; RC represents the set of task concept relationships.

This paper defines the complete concept of aggregation unit concept A as a quintuple:

$A = \{A_0, CA, U(A_0,T), RA, SynA\}$

where $A_0$ represents the aggregation unit concept; CA represents the domain concept corresponding to the aggregation unit theme; $U(A_0,T_0)$ represents the task scenario relevance of the aggregation unit, determined by the task type $T_0$ and its relevance degree to the aggregation unit $A_0$. For example, the aggregation unit "web encyclopedia" has encyclopedia genre attributes at the genre level, and its semantic function is to introduce various aspects of corresponding concepts, with high perceived usability under the "learning background" task. U is the specific numerical value of perceived usefulness in instances; RA represents the relationship set; SynA represents the synonym set of concept $A_0$.

This paper defines the complete concept of web document concept D as a nonuple:

$D = \{D_0, CD, Tit, Cont, Auth, Inst, S, G, Time\}$

where $D_0$ represents the web document concept; CD represents the domain concept corresponding to the aggregation unit theme; Tit represents document title; Cont represents document content; Auth represents document author; Inst represents document institution; S represents document source; G represents document genre; Time represents publication time.

### 3.2 Construction and Formalization of Fine-grained Aggregation Ontology

Using 81 types of web documents on the topic of "citation analysis" from a previously constructed library and information science corpus [38] as the data source, this paper establishes the four concept sets, attributes, relationships, and corresponding instances of the fine-grained aggregation ontology through empirical research, constructing and formalizing the ontology.

The experimental corpus includes four genre types: open access research papers, online bibliographic records, web encyclopedia entries, and academic blog posts.

**3.2.1 Fine-grained Aggregation Ontology Concept System** According to the basic types of fine-grained aggregation ontology concepts, its concept system includes: aggregation unit concept system, domain concept system, task concept system, and document concept system.

**(1) Aggregation Unit Concept System.** Based on the aggregation unit knowledge system, we construct the aggregation unit concept set and inter-concept relationships [38], determining aggregation unit ontology concepts and

their attributes in a top-down manner: according to the aggregation unit classification system, we identify aggregation unit concepts at different levels, such as research papers, introductions, topic background introductions, and other level concepts, as shown in Table 2 .

**(2) Domain Ontology Construction.** Domain resource ontology construction has a long research history with relatively mature theoretical and methodological foundations. To explore methods for constructing fine-grained aggregation ontology, this paper uses a dictionary-based, machine-assisted approach to construct the topic concept knowledge system. Specifically, we use the knowledge system about citation analysis from the Baidu Baike entry as the foundation for the "citation analysis" corpus knowledge system. We employ the ROSTCM software developed by Wuhan University to segment the corpus text and calculate word frequencies, obtaining high-frequency keyword concepts reflecting domain characteristics. After discussion among research team members, all meaningful new words are added to the citation analysis knowledge system for refinement, thereby constructing the "citation analysis" ontology using a "top-down" approach. Ultimately, we construct a domain concept system comprising 6 levels and 100 concepts.

**(3) Web Document Ontology and Aggregation Unit Ontology.** Web document ontology aims to construct ontology concepts and conceptual relationships regarding various dimensional information units of web documents, thereby cooperating with the aggregation unit concept system to support fine-grained aggregation of web documents. Referencing Zhu Jiaxian et al.'s research on Web resource ontology and Qiu Junping et al.'s research on collection resource semantic ontology [40,50], this paper constructs main concepts of the web document ontology. Thus, the document ontology includes concepts such as genre, content, creator, institution, source, and title, as shown in Figure 2 [Figure 2: see original paper].

The task scenario concept set defines concepts about task scenarios, providing a foundation for constructing associations between tasks and aggregation unit concepts and supporting fine-grained aggregation of web documents. Based on L. Zhang and L. Freund et al.'s definitions of task ontology [36,49], this paper constructs main concepts of the task ontology, as shown in Figure 3 [Figure 3: see original paper].

**3.2.2 Fine-grained Aggregation Ontology Attributes and Relationships** Based on the various ontology concept sets identified in the previous stage, the second stage clarifies main attributes of ontology concepts, including: perceived usefulness of aggregation units, semantic relationships between aggregation unit classes and instances formed by aggregation units' unique language functions, relationships between web document ontology concepts, and relationships between themes and aggregation unit ontologies and task ontologies. Therefore, fine-grained aggregation ontology mainly contains 11 major attributes, as shown in Table 3 .

Based on the class-attribute relationships in the fine-grained ontology, the following reasoning can be performed:

**(1) Usefulness of Aggregation Units.** If an aggregation unit's "perceived usability" score under a specific task scenario is higher than the threshold, then the aggregation unit has high usefulness under that task scenario.

**(2) Task Relevance Relationships of Aggregation Units.** If certain aggregation units' perceived usability scores under a specific task scenario are all higher than the threshold, then these aggregation units have task relevance.

**(3) Themes of Aggregation Units.** Lower-level aggregation units inherit the attribute "domain concept is…" from higher-level aggregation units. If a higher-level aggregation unit's domain concept is X, then the lower-level aggregation unit's domain concept is also X. An equivalent inverse relationship exists between higher and lower-level aggregation units.

**(4) Document Ontology-Related Attributes of Aggregation Units.** Aggregation units at all levels obtain document ontology-related attributes: if a document's author/institution/source/title/genre is X, then the aggregation unit's author/institution/source/title/genre is X. An equivalent inverse relationship exists between document ontology and aggregation unit.

Based on Table 2 and the formalized description of semantic reasoning based on class attributes, this paper provides explicit formalized descriptions of these semantic relationship types, determining from top to bottom that fine-grained aggregation ontology mainly contains five major relationship types:

**(1) Inheritance Relationships and Inverse Relationships.** These include subclass-to-class inheritance, instance-to-class inheritance, class-attribute relationships, and their inverses. Semantic relationships between concepts in aggregation unit ontology are relatively clear and can be established based on relationships between upper and lower aggregation units in the aggregation unit knowledge system. For inheritance relationship attributes, definitions can be made according to relationship sources and natures. Semantic relationships between instances contained in classes can be obtained through relationships between their belonging classes.

**(2) Progression Relationships and Inverse Relationships.** These are semantic progression relationships and their inverses between subclasses of the same parent class and their instances in aggregation unit ontology, which can be established based on relationships between upper and lower aggregation units in the aggregation unit knowledge system.

**(3) Task Relevance.** Since specific aggregation units have high perceived usability under specific task types, associations based on task scenarios are formed between aggregation units at the same or different levels. These can be qualitatively described and quantitatively formalized through numerical relationships of aggregation unit perceived usefulness.

**(4) Semantic Relationships of Domain Concepts.** These are semantic associations formed between domain concepts; semantic relevance between instances contained in classes at the same or different levels can be calculated through weighted combination based on domain ontology concept relevance and this class's aggregation unit scenario association attributes.

**(5) Relationships in Web Document Concepts.** Associations between web documents and their authors, institutions, etc., can be directly obtained from web document metadata or acquired and integrated from public information sources such as institutional websites.

**3.2.3 Instances of Fine-grained Aggregation Ontology** Using the "citation analysis" corpus as the source for fine-grained aggregation ontology instances, we extract and count relevant instance information of web documents in the corpus according to the fine-grained aggregation ontology, obtaining instance quantity distributions for ontology classes as shown in Tables 4 and 5 .

Table 4 shows instance statistics for ontology classes in the aggregation unit knowledge system, while Table 5 shows instance statistics for document ontology and domain concept ontology classes. As shown in Tables 4 and 5, aggregation unit ontology instances come from 81 instances across four genres, containing 254 component units under genre units, and 1,247 language function units under component units. The domain concept ontology includes 100 instances, while the document ontology contains 89 authors, 50 institutions, and 38 sources.

For fine-grained aggregation ontology classes, instance attributes can be defined according to the class attributes in the ontology. Here, we focus on exploring the numerical attribute of aggregation unit instances obtained through calculation—perceived usefulness of aggregation unit instances. During aggregation unit attribute definition, perceived usability serves as a numerical attribute of aggregation unit class, influenced by aggregation unit type and user task scenario, acting as an indicator of aggregation unit task relevance.

**3.2.4 Formalization of Fine-grained Aggregation Ontology Based on Protégé** Based on clarified concepts and attributes of fine-grained aggregation ontology, we provide formalized specifications to form a formalized ontology. Using the "citation analysis" dataset as the corpus source, we employ the ontology editing and visualization tool Protégé to add semantic tags, perform coding and formalization according to OWL language specifications, thereby establishing the fine-grained aggregation ontology for web documents, with its full view shown in Figure 4 [Figure 4: see original paper].

From the perspective of major ontology classes, fine-grained aggregation ontology includes aggregation unit ontology, domain ontology, web document ontology, and task ontology. Therefore, viewing the upper-level ontology concepts according to hierarchical structural relationships reveals the main composition

and relationships of fine-grained aggregation ontology, as shown in Figure 5 [Figure 5: see original paper].

As shown in Figure 5, aggregation unit ontology forms the foundation of fine-grained aggregation ontology, providing hierarchical relationships, semantic relationships, and task relevance relationships for resource fine-grained aggregation, thereby supporting fine-grained aggregation of web documents and visual navigation and retrieval methods based on relationships between aggregation units. The hierarchical overview of aggregation unit ontology is shown in Figure 6 [Figure 6: see original paper].

As shown in Figure 6, aggregation units belong to part of the content attribute in document ontology, containing the attribute of perceived usability, described by domain concepts, and including network document genres such as online bibliographic abstracts, OA papers, academic blog posts, and online encyclopedia entries. Relationships exist between aggregation units at various levels, including whole-part relationships and progression relationships and their inverses between aggregation units in the same group at the same level.

Based on the "citation analysis" web document corpus, the hierarchical structure of concept relationships in the domain ontology is shown in Figure 7 [Figure 7: see original paper].

The relationships between document ontology, task ontology, and domain and aggregation unit ontologies are shown in Figure 8 [Figure 8: see original paper]. As shown, domain concepts are themes of document ontology, task ontology, and aggregation unit ontology. Aggregation units originate from the content attribute of document ontology.

## 4 Evaluation and Discussion of Fine-grained Aggregation Ontology

Yue Lixin and Liu Wenyun synthesized various foreign evaluation indicators to propose ontology evaluation standards of completeness, clarity, consistency, extensibility, and compatibility [51]. This paper evaluates the constructed fine-grained aggregation ontology around these criteria.

### 4.1 Completeness

Since the constructed fine-grained aggregation ontology derives from experimental corpora, it can largely cover concepts and relationships of various ontologies in the corpus, especially achieving 100% coverage of document ontology. Compared with the initial classification system of aggregation units proposed by C.C. Ma and S.J. Cao [38], since this study's aggregation unit ontology does not adopt the two semantic function units with lower scores in the initial classification system, the coverage of aggregation unit ontology is 96.5%.

However, since this study's task ontology concepts derive from task types pro-

posed by L. Zhang and L. Freund [36,49], its completeness and systematicity are insufficient compared with the task faceted classification system proposed by Y. Li [52]. Therefore, future research can reference Y. Li's task system [52] to construct associations between tasks and aggregation units from more facets and types, establishing a more complete task ontology. Additionally, for more genre types of web documents in the library and information science field and even more disciplinary fields, this study's fine-grained aggregation ontology requires not only supplementing and improving aggregation unit types and relationships but also constructing domain concept systems based on dictionaries or large-scale corpora to ensure completeness of fine-grained aggregation ontology.

### 4.2 Clarity

This study constructs a fine-grained aggregation ontology for four genre types of web documents in library and information science and user needs. However, since aggregation unit division for multiple genres remains exploratory, is difficult, time-consuming, and lacks stable automatic classification methods, this study uses small-scale corpus samples with manual division as the primary data source. Although the small corpus size limits the application of fine-grained aggregation ontology, it enables exploratory research on construction methods and ontology utility while ensuring conceptual system accuracy.

All four types of concepts in the fine-grained aggregation ontology can be constructed into ontologies with clear hierarchies, clear concept boundaries, and explicit attribute and relationship definitions based on existing knowledge systems. Among them, the aggregation unit knowledge system is constructed according to genre structure theory in linguistics [38], making aggregation units have relatively clear definitions in concepts, attributes, and relationships. Task ontology references definitions by L. Zhang and L. Freund to clarify meanings and attributes of different tasks [36,49]. Document ontology concepts are more universally understood. Domain concepts are determined through automatic word segmentation, reference to online encyclopedia entries, and discussion among library and information science researchers, ensuring clear concept meanings and inter-concept relationships, thereby guaranteeing clarity of fine-grained aggregation ontology.

### 4.3 Consistency

Since the aggregation unit concept classes, attributes, and relationships that serve as the main body are far fewer in number than the domain concept system, and the aggregation unit knowledge system is generated through manual content analysis and move-step analysis, the classification system construction itself has undergone internal consistency investigation of divided aggregation units [38], thus no noise data from semi-automated/automated construction processes exists, resulting in high consistency.

### 4.4 Extensibility

Although this study's fine-grained aggregation ontology remains at the method exploration and manual construction stage, its basic framework—especially the aggregation unit concept system based on genre structure theory—allows maintenance and updates for ontology evolution. It can continuously improve hierarchical structures and semantics, expand newly emerged terms, concepts, and relationships through automated or semi-automated methods according to actual conditions.

### 4.5 Compatibility

The aggregation unit knowledge system and task knowledge system included in fine-grained aggregation ontology have unified and clear theoretical foundations, enabling compatibility across aggregation unit knowledge systems and task systems in multiple disciplines. Domain concepts selected based on encyclopedia entries combined with corpus word frequencies provide a foundation for compatibility and mapping of concept systems.

This study aims at constructing fine-grained aggregation ontology for web academic documents, establishing the ontology after clarifying its theoretical framework. It proposes, for the first time, the idea and method of constructing fine-grained aggregation ontology based on aggregation unit knowledge systems, clarifies basic types of fine-grained aggregation ontology concepts, clarifies the relationship between aggregation unit granularity and domain concept granularity, defines ontology concepts, and thereby constructs the theoretical framework for fine-grained aggregation ontology construction. Additionally, it clarifies ontology construction approaches and methods from three aspects: ontology concept system, attributes and relationships, and instance construction.

This study's experimental corpus samples and corresponding ontology will be applied to a fine-grained aggregation prototype system for systematic exploration in aggregation unit automatic classification, organization and indexing, and interaction research, aiming to more comprehensively explore and grasp aggregation unit utility and provide a foundation for larger-scale, automated, and intelligent exploration and application.

## References

[1] Wen Tingxiao, Luo Xianchun, Liu Xiaoying, et al. Review of knowledge unit research[J]. Journal of Library Science in China, 2011, 37(5): 75-86. [2] Wen Youkui, Jiao Yuying. Knowledge discovery based on knowledge elements[M]. Xi'an: Xidian University Press, 2011. [3] Wen Youkui. Knowledge organization and retrieval based on "knowledge elements"[J]. Computer Engineering and Applications, 2005, 41(1): 55-57, 91. [4] Wen Youkui, Xu Guohua. Knowledge element linking theory[J]. Journal of the China Society for Scientific and Technical Information, 2003, 22(6): 665-670. [5] Wen Youkui, Wen Hao, Xu Duanyi,

et al. Text knowledge indexing based on knowledge elements[J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(3): 282-288. [6] Wang Yan, Wen Youkui. Research on transforming text units into knowledge units[J]. Information Studies: Theory & Application, 2007, 30(3): 409-411, 362. [7] Wen Youkui, Jiao Yuying. Research on knowledge unit organization and retrieval based on category theory[J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(3): 387-392. [8] Wen Youkui, Jiao Yuying. Research on Wiki knowledge element semantic graph[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(6): 870-876. [9] Wen Youkui, Jiao Yuying. Research on knowledge element semantic linking model[J]. Library and Information Service, 2010, 54(12): 27-31. [10] Zhou Xiuhui. Knowledge element search engine: CNKI knowledge search platform[J]. Modern Information, 2007, 27(5): 220-222. [11] Tao Shanju, Liu Qingtang, Wang Fan, et al. Construction of educational technology discipline resource library based on knowledge units[J]. Modern Educational Technology, 2011, 21(5): 115-120. [12] ZOU J, LIU Q. A knowledge element model for knowledge abstract and fusion system[C]//2009 International conference on new trends in information and service science. Washington, DC: IEEE Computer Society, 2009: 23-26. [13] TRACE CB, DILLON A. The evolution of the finding aid in the United States: from physical to digital document genre[J]. Archival science, 2012, 12(4): 501-519. [14] SWALES JM. Aspects of article introductions[M]. Birmingham: the University of Aston in Birmingham, 1981. [15] CROOKES G. Towards a validated analysis of scientific text structure[J]. Applied linguistics, 1986, 7(1): 57-70. [16] HOPKINS A, DUDLEY-EVANS T. A genre-based investigation of the discussion sections in articles and dissertations[J]. English for specific purposes, 1988, 7(2): 113-121. [17] SAMRAJ B. Introductions in research articles: variations across disciplines[J]. English for specific purposes, 2002, 21(1): 1-17. [18] POSTEGUILLO S. The schematic structure of computer science research articles[J]. English for specific purposes, 1999, 18(2): 139-160. [19] BRUCE I. Cognitive genre structures in methods sections of research articles: a corpus study[J]. Journal of English for academic purposes, 2008, 7(1): 38-54. [20] BRETT P. A genre analysis of the results section of sociology articles[J]. English for specific purposes, 1994, 13(1): 47-59. [21] KANOKSILAPATHAM B. Rhetorical structure of biochemistry research articles[J]. English for specific purposes, 2005, 24(3): 269-292. [22] NWOGU KN. The medical research paper: structure and functions[J]. English for specific purposes, 1997, 16(2): 119-138. [23] LEWIN BA, FINE J, YOUNG L. Expository discourse: a genre-based approach to social science research texts[M]. London: Continuum, 2001. [24] Zhao Fuli. Move structure analysis of English TV news leads[J]. Foreign Language Teaching and Research, 2001, 33(2): 99-104. [25] Ge Dongmei, Yang Ruiying. Genre analysis of academic paper abstracts[J]. Modern Foreign Languages, 2005, 28(2): 138-146, 219. [26] Cui Yanyan, Wang Tongshun. Macrostructure and microstructure of English academic lectures: application of genre analysis in academic discourse analysis[J]. Shandong Foreign Language Teaching Journal, 2004(5): 27-30. [27] Yang Ruiying. Application of genre analysis: structural analysis of

applied linguistics academic articles[J]. Foreign Languages and Their Teaching, 2006(10): 29-34. [28] BISHOP AP. Document structure and digital libraries: how researchers mobilize information in journal articles[J]. Information processing & management, 1999, 35(3): 255-279. [29] DILLON A. Designing usable electronic text[M]. Boca Raton FL: CRC Press, 2004. [30] DILLON A, SCHAAPD. Expertise and the perception of shape in information[J]. Journal of the American Society for Information Science and Technology, 1996, 47(10): 786-788. [31] VAUGHAN MW. Identifying regularities in users' conceptions of information spaces: designing for structural genre conventions and mental representations of structure for Web-based newspapers[D]. Indiana: Indiana University, 1999. [32] VAUGHAN MW, DILLON A. Learning the shape of information: a longitudinal study of Web-news reading[C]//Proceedings of the fifth ACM conference on digital libraries. New York: ACM, 2000: 236-237. [33] DILLON A, SCHAAPD. Expertise and the perception of shape in information[J]. Journal of the American Society for Information Science and Technology, 1996, 47(10): 786-788. [34] DILLON A. Spatial-Semantics: how users derive shape from information space[J]. Journal of the American Society for Information Science, 2000, 51(6): 521-528. [35] ZHANG L, KOPAK LR, FREUND L, et al. A taxonomy of functional units for information use of scholarly journal articles[C]//Proceedings of the American Society for Information Science & Technology. Somerset, NJ: John Wiley & Sons, 2010, 47(1): 1-10. [36] ZHANG L, KOPAK LR, FREUND L, et al. Making functional units functional: the role of rhetorical structure in use of scholarly articles[J]. English for specific purposes, 2011, 31(1): 21-29. [37] ZHANG L. Grasping the structure of journal articles: utilizing the functions of information units[J]. Journal of the American Society for Information Science and Technology, 2012, 63(3): 469-480. [38] MA C, CAO S. Identifying structural genre conventions across academic web documents for information use[C]//Proceedings of the Association for Information Science & Technology. Somerset, NJ: John Wiley & Sons, 2017: 260-267. [39] Ma Yumeng, Liu Fenghong, Huang Jinxia. Research on domain ontology model framework in STKOS[J]. Library and Information Service, 2015, 59(3): 119-125, 139. [40] Qiu Junping, Yang Qiang, Lou Wen. Theoretical and empirical research on resource ontology construction[J]. Information Studies: Theory & Application, 2014, 37(5): 1-6. [41] MIZOGUCHI R. YAMATO: Yet another more advanced top-level ontology[EB/OL]. [2019-10-28]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=00B4895D3EF153E0F74DC6B248D307FB?doi=10. [42] Zhang Nannan. Research on semi-automatic domain ontology construction method for semantic web[D]. Dalian: Dalian Maritime University, 2008. [43] GUARINO N. Semantic matching: formal ontological distinctions for information organization, extraction and integration[C]//PAZIENZA MT. Information extraction: a multidisciplinary approach to an emerging information technology. Berlin: Springer Verlag, 1997: 139-170. [44] Guo Jiaqi. Research on domain ontology construction and its application in information retrieval[D]. Beijing: Beijing University of Posts and Telecommunications, 2007. [45] Wang Xiangqian, Zhang Baolong, Li Huizong. Overview of ontology research[J]. Journal of Intelligence, 2016, 35(6): 163-170. [46] GRUBER T.

Towards principles for the design of ontologies used for knowledge sharing[J]. International journal of human-computer studies, 1995, 43(5/6): 907-928. [47] Li Jing, Meng Xianxue, Su Xiaolu. Research on domain ontology construction methods and applications[M]. Beijing: China Agricultural Science and Technology Press, 2009. [48] DANIEL LR, NATALYA FN, MARK AM. Protégé: a tool for managing and using terminology in radiology applications[J]. Journal of digital imaging, 2007, 20(S1): 34-46. [49] FREUND L. A cross-domain analysis of task and genre effects on perceptions of usefulness[J]. Information processing & management, 2013, 49(5): 1108-1121. [50] Zhu Jiaxian, Bai Weihua, Li Jigui. Research on multi-granularity semantic annotation and application technology for Web resources[J]. 2011, 38(8): 83-87. [51] Yue Lixin, Liu Wenyun. Comparative study of domain ontology construction methods at home and abroad[J]. Information Studies: Theory & Application, 2016, 39(8): 119-125. [52] LI Y, BELKIN NJ. A faceted approach to conceptualizing tasks in information seeking[J]. Information processing & management, 2008, 44(6): 1822-1837.

**Author Contributions:** Ma Cuichang: research design, implementation, and paper writing; Cao Shujin: research objectives and ideas.

**Study on the Construction of Fine-grained Aggregation Ontology for Academic Documents in the Internet Environment**

*Ma Cuichang*[1], *Cao Shujin*[2]
[1]Sun Yat-sen University Library, Guangzhou 510275
[2]School of Information Management, Sun Yat-sen University, Guangzhou 510006

**Abstract:** [Purpose/Significance] Fine-grained information aggregation has become a focus in knowledge organization. This paper aims to explore the construction of fine-grained aggregation ontology for web academic documents. [Method/Process] The study clarifies the types, granularity characteristics, and definitions of fine-grained aggregation ontology concepts. Using "citation analysis" documents in library and information science as corpus, the ontology is built through concepts, attributes, relationships, and instances, followed by evaluation and discussion. [Result/Conclusion] This paper first proposes constructing fine-grained aggregation ontology using aggregation unit concepts, informing knowledge organization system development for fine-grained organization, retrieval, and navigation.

**Keywords:** information aggregation; ontology; web documents; library and information science

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