VISIONARY AND THEORY PAPER

Metadata as a Methodological Commons: From Aboutness Description to Cognitive Modeling

Wei Liu[1], Yaming Fu[1,2†], Qianqian Liu[1]

[1]Shanghai Library (Institute of Scientific and Technical Information of Shanghai), No. 1555 Huaihai Middle Road Xuhui District, Shanghai 200031, China

[2]School of Information Management at Nanjing University, No. 163 Xianlin Avenue, Nanjing, Jiangsu 210093, China

Keywords: Metadata; Methodological Commons; Content Architecture; Data Modeling; Ontology; Semantic Web; Semantic Formalization; Web3.0; Metaverse

Citation: Liu, W., Fu, Y.M., Liu, Q.Q.: Metadata as a Methodological Commons: From Aboutness Description to Cognitive Modeling. Data Intelligence 5(1), 289-302 (2023). doi: doi.org/10.1162/dint_a_00189
Received: October 10, 2022; Revised: November 15, 2022; Accepted: December 12, 2022

ABSTRACT

Metadata is data about data, which is generated mainly for resources organization and description, facilitating finding, identifying, selecting and obtaining information[①]. With the advancement of technologies, the acquisition of metadata has gradually become a critical step in data modeling and function operation, which leads to the formation of its methodological commons. A series of general operations has been developed to achieve structured description, semantic encoding and machine-understandable information, including entity definition, relation description, object analysis, attribute extraction, ontology modeling, data cleaning, disambiguation, alignment, mapping, relating, enriching, importing, exporting, service implementation, registry and discovery, monitoring etc. Those operations are not only necessary elements in semantic technologies (including linked data) and knowledge graph technology, but has also developed into the common operation and primary strategy in building independent and knowledge-based information systems.

In this paper, a series of metadata-related methods are collectively referred to as 'metadata methodological commons', which has a lot of best practices reflected in the various standard specifications of the Semantic Web. In the future construction of a multi-modal metaverse based on Web 3.0, it shall play an important role, for example, in building digital twins through adopting knowledge models, or supporting the modeling of

† Corresponding author: Yaming Fu (E-mail: fuyaming_uk@yahoo.com; ORCID: 0000-0001-9736-4916).

① The four FRBR user tasks (https://www.librarianshipstudies.com/2017/09/functional-requirements-for-bibliographic-records-frbr.html).

the entire virtual world, etc. Manual-based description and coding obviously cannot adapted to the UGC (User Generated Contents) and AIGC (AI Generated Contents)-based content production in the metaverse era. The automatic processing of semantic formalization must be considered as a sure way to adapt metadata methodological commons to meet the future needs of AI era.

1. FROM METADATA TO METADATA METHODOLOGICAL COMMONS

1.1 The Emergence and Development Of Metadata

Metadata arises from the need to describe things. It was long before the emergence of the term that the concept had been existed, which can be traced back to five or six thousand years ago when clay tablets were used in the Mesopotamia region to record transactions. Around 280 BC, a labeling and classification system called 'Pinakes' was used by the Library of Alexandria to describe information of scrolls, which then evolved and developed into the library card catalog [1]. Those early and traditional metadata applications are based on manual descriptions on physical carriers, and can only be managed and utilized manually, which is the metadata in the 'pre' digital age, referring to as 'Metadata 1.0' era by Tom Baker, CIO of the Dublin Core Metadata Organization (DCMI).

The term "metadata" was emerged with the development of database technology, as it is normally required to describe data and tables in database application. Lines of records (usually rows in a table) are used to document the various attributes of the object being described. This is in fact moving the contents of the clay tablets or card catalog into the computer and realizing the goal of 'machine readable'. In addition, by applying certain structure to objects description, it is achieving the goal of 'machine computable' to improve the efficiency of querying and managing metadata with the power of computing. For both tape file and relational databases that appeared later on, metadata has been used to record information about data structure and other annotated information. At this stage, it was limited to be used in a closed system, and the MARC bibliographic data that is widely used in the library was born at this stage, which is the most typical form of metadata, also known as the 'Metadata 2.0' era.

It is in the Digital Age that metadata got developed and became a prominent subject, especially in the context of digital library construction. In this phase, the 'information explosion' makes it very difficult to find and sift information, and there is an increasing need to obtain useful information in the global network, so the need of describing and structuring information. The primary goal is to annotate data and objects, and to support the retrieval need on a large number of commercial and non-commercial, semi-structured and unstructured library databases. By doing such, metadata is supposed to achieve the goal of organizing, finding, locating, selecting, and to facilitate the transformation of information into knowledge, which is the core value in the Digital Age, and this stage can be regarded as the 'Metadata 3.0' era.

With the development of technology, the structural unit (e.g. string, list, data set, etc.) that can be processed by computing technology evolves from information to 'knowledge'. The 'computability' of machine refers to not only the management of digital information, semantic description and coding, metadata descriptions and coding specifications. Through standardized coding, 'knowledge' can be delivered and integrated among machines, enabling the goal of 'machine-understandable' and 'machine-interoperable'. This can be

understood as the process of facilitating the transformation of knowledge into 'intelligence', and further accomplishing the function of 'smart data' by supporting data mining or machine learning, even supporting the automatic construction of knowledge systems. The manual way of standardizing metadata in 2.0 and 3.0 era is obviously unsustainable, therefore it is crucial to promote the application of metadata standard specifications to achieve its automatic generation, coding, linking and visualization to the full extent. The technology of artificial intelligence (AI) can be applied to the entire metadata generation process, and this is the anticipation of 'Metadata 4.0' era.

1.2 The Formation of Metadata Methodological Commons

As more databases become accessible via network, metadata descriptions and applications need to follow a unified standard for consistent understanding and data interoperability. Looking back at the history of metadata, it is not only as simple as providing objective description, the purpose of description (why to describe) and the method (how to describe it) are all of significance. Many factors such as convenience and costs need to be considered and balanced, as well as the consideration about extensibility and compatibility. It should be noted that there is always a trade-off in implementing metadata, and a reasonable balance needs to be found between various factors.

Metadata methods are often reflected in metadata standards and their best practices, as well as various metadata models related to application systems. The Dublin Core Metadata Initiative (DCMI) is a typical representative of the standardization of metadata applications, and its development in the past three decades also represents the main process of the 'metadata methodology commons'. DCMI was born shortly after the advent of the World Wide Web. The original intention was to explore the standard specification for the description and coding of web resources. The word 'core' in its name is based on the foresight of this group of pioneers, who insisted on the basic rule of "small is beautiful", committed to providing a universal and simplest set of 'core' elements. For the complex things, they recommended a set of extensions, which are led by methodology, rather than directly defining the set of elements, promoting the fulfillment of needs based on this core.

DCMI not only proposed the Qualified Element Set of resource descriptions, but also put forward the DCAM Abstract Model (DCAM) which declares the composition of metadata records, and regulated that Dublin Core Application Profile (DCAP)[②] should be implemented as an extension rule of metadata schema. Their application achieved a huge success and has been applied on the Internet. In recent years, DCMI has also been committed to the development of a Tabular specification for the formal encoding of DCAP, and proposed an abstract model and coding specification for the basic structure of resource description. DCAP and DCMI Core Package structure provide a basic technical guarantee for the standardization and functional implementation of metadata applications, so that the metadata records can be self-describing and self-interpreting, and be independent of technologies.

[②] The creation of a new metadata schema from available open vocabularies creates what can be called an application profile. See more at: https://www.dublincore.org/groups/application_profiles_ig/dctap_primer/

Metadata application has gradually become prevalent, especially in the field of scientific research databases and cultural heritage management, where many standardization specifications and best practice of metadata have designed and put into practice. On the Internet space, the knowledge graph technology based on Schema.org was proposed by search engines such as Google, which made metadata a common method in supporting Search Engine Optimization (SEO). In addition to this, there is a notable increase in the use of DC metadata tags (such as dc.title, dc.subject, etc.), which is also a crucial step for reveal semantics to search engines. The prevalence of metadata applications has effectively improved the recall and accuracy rate of information systems, helping to reveal the relationship between information and data from multiple angles, as well as providing solutions for interoperability between different systems. However, at this stage, most of the metadata applications works can only be carried out manually, such as indexing, extracting, organizing, linking, proofreading, etc., and therefore the cost is huge. What is more, the strict standards often cause inefficiency and varied data quality, and even simple standards might have problems such as insufficient content description, and lacking usability, etc. Despite the initial formation of metadata methods, its methodological commons construction is still in the preliminary stage and has not been widely used. It also lacks systematic summary of its achievements and research outputs.

In this paper, we offer a straightforward definition of the metadata methodological commons. It can be considered as a systematic methodological framework and specifications for the formal description of content architecture of a specific knowledge system (such as various types of digital libraries), rather than merely providing a structured description of the library resources in the local network. The purpose of metadata methodological commons is to meet the evolving needs of the information system on resource disclosure, interoperability and long-term preservation in the network environment. The specific functions may include querying, searching, browsing, accessing, sorting, and even analyzing, visualizing and so on.

2. METADATA AND MACHINE INTELLIGENCE

2.1 From Aboutness Description to Semantic Coding

Same as the function of cataloging in traditional libraries, metadata is the foundation in building digital libraries. It was put forward under the need of cataloging web and digital resources. Before the development of semantic technology, metadata was simply used to generate structured indexes to support the keyword retrieval, where knowledge was implicit in structured textual information. The knowledge transfer of such information systems can only be processed by human, who need to interpret the content after obtaining information, while semantics cannot be transferred or inter-operated between machines, and not to mention knowledge integration. Therefore, the early digital libraries usually lacked a holistic, macroscopic knowledge description system, and did not establish a correlation relationship between entities within the knowledge system; instead, it only provided a partial or microscopic description of information resources, and the function was also limited.

The emergence of the semantic web technology proposes for the first time a complete set of ways to encode and formalize semantics, supporting the grasp of knowledge through the calculation of semantic

data. The basis of semantic web technology is metadata, which is able in achieving 'machine-processing'; the generation of metadata coding specifications is capable in solving the computational problems, which will achieve the goal of 'machine-understandable' issue.

Based on the 'knowledge representation' technique, semantic technology implements a formal representation of First-Order Predicate Logic (i.e., descriptive logic), and generally provides machine-processable encoding by schema. Specific metadata records also need to be instantiated, that is, the serialized encoding (which is what we usually refer to as metadata encoding) [12]. Strict metadata encoding specifications should be based on the Resource Description Framework introduced by the World Wide Web Consortium, also known as RDF, which specifies the triple structure for describing any entity, as well as the formal encoding specifications, namely RDFS and OWL [11]. In this framework, each piece of metadata is a basic judgment about the properties of things, and the triples are the most proper formal expression of metadata, which can be expressed via natural language, and thus has better readability. This makes it the basic specification of semantic encoding, and its implementation can be any formal language, such as RDF/XML or N3, or JSON-LD, etc.It could be regarded that the triples are the smallest structural unit of human cognition, are bricks and tiles of the knowledge skyscraper and the elementary particles of the entire knowledge universe. Formal representations of metadata enables machine-understanding and strong computability, providing possibilities in a range of metadata services, such as registry, query, mapping, discovery, extension, navigation and so on. Additionally, it provides ways for computers to break through the limitation of only presenting information, and to manage and work on semantics directly. It is not only designed for human understanding, but also for interoperability between machines, which can be further applied to the Internet of Things (IoT), sensor networks and semantic interaction between servers. Till this point, a set of metadata methods aimed at providing semantic architecture and content architecture for information systems got established.

The metadata methodological commons can be seen as a process of modeling the whole domain knowledge system through human analysis. The knowledge ontology itself is a conceptual model of domain knowledge, so the construction of the knowledge ontology (shorten as ontology) is also an important part in the metadata methodological commons. Ontology should be built based on concepts, rather than words, which is similar to the concept-based knowledge system in traditional information retrieval research. Ontology is a normalized description of the domain concepts and the relationship between them, which need to be standardized and explicit as much as possible to support machines processing, and support sharing and reuse. 'Explicit' here means that the types of concepts employed and the constraints to which they apply are clearly defined. The metadata model from DCMI recommends the use of URIs to annotate entity objects such as resources, terms, etc., which in fact can be seen as a kind of 'conceptualization'. To be noted, generally, metadata application does not emphasize this point, and take the current knowledge graph applications as an example, many of them use words as points. Although it is hard to establish a strict concept-based knowledge system, through certain software, the entities could be extracted automatically to build relationships at a lower cost.

Ontology is able to provide an overall description of the underlying resources' metadata. Generally, according to design requirements, the domain knowledge system or resource base or knowledge base at a

macro level is described, and it contains various upper-level entities and interrelationships, which can be treated as a combination of related ontology. The complex part of ontology modelling is the vocabulary design and relationships as it adopts different modeling language such as OWL, etc.; apart from this aspect, in essence, ontology modelling is similar to metadata encoding. OWL itself can be seen as an extension of RDFS, so this makes it available to adopt the same application schema of metadata to integrate different coding patterns, which will not be covered in this paper. It can be referred that metadata schema and formal ontology together constitute the domain knowledge system or the content architecture of the knowledge base and repository.

In library, the usage of metadata methods has expanded to the relevant aspects of information system application, including the description scheme of resource content, management, technology and other processes. The most important function of it is to provide a structure and index based on 'content' for digital document which is the main body of digital collections, by which it breaks through the traditional way of describing document unit as an object, but goes deep into the content. It not only describes and manages the theme, person, transactions, events, etc. in a direct way, but also is capable of further supporting the construction of content models [8]. It is put into practice in digital humanities (DH) platforms and even in the application of digital twins and metaverse to better support digital transformation in areas such as scientific research, education, and publishing (in other words, to support the transition to the fourth paradigm of scientific research).

2.2 The Challenges Brought by Big Data

Over long periods of time, all the knowledge organization and expression have been built on the understanding of human cognition process. Metadata structures information, the link among data forms a meaningful network, and the description and acquisition of the linked relationships benefit from metadata. The larger the network, the more necessary it is to have a mechanism to derive knowledge from data and store and utilize it to distill so-called "artificial intelligence" (AI). The use of semantic technology to encode knowledge (which is the encoding of the objective content of knowledge - 'semantics') so that computers shall 'understand' knowledge, process it, and achieve the acquisition, transmission and processing of knowledge in a distributed environment. At present, the metadata method has technically provided a way for cognitive computing, which is knowledge representation and description. This at least provides a solution, that is, through structuring information, along with domain knowledge and the use of ontology design, knowledge in certain domain gets conceptualized and understood semantically, leading to the construction of domain knowledge model and supporting content-based knowledge association for information. In this way, the massive storage and high-speed computing ability of computer systems can be fully made use of, and application systems of domain knowledge base in various fields can be further developed.

There are two important assumptions behind the above solution: firstly, there should be enough resources to complete the semantic work of the growing massive information in a timely manner; secondly, all the knowledge could be coded, formalized, and modeled through formal knowledge representation tools, so

to be recognized and calculated by computers. Obviously, these two assumptions do not always hold: the information explosion has brought about the geometric growth in data, and it is now generally accepted that manual annotation and organization of information is an impossible task; what is more, not all the knowledge can be coded or expressed symbolically.

Data, computing power and algorithms are the three key elements of AI. At present, machine learning is heavily dependent on big data and available training data. Due to the lack of enough or quality training data for machine learning models, fully automated metadata generation models have only achieved limited success in the field of digital libraries. Many practitioners daunted when applying semantic technologies because of the complexity of ontology, the ambiguity of semantic encoding, the inconsistency in technical implementation and the difficulty of process automation, which usually requires a lot of manual processing during the construction process. Although the recent introduction of big data management technology, especially the graph database technology, is expected to alleviate the problems of efficiency, scale and scalability to a certain extent, the fundamental problems have still not been solved, and it is urgent for researchers to find an engineering method that can be rapidly promoted to big data sets.

2.3 Metadata Methodological Commons and Machine Intelligence

Metadata as methodological commons can be incorporated with machine learning in two ways: certain level of automation of metadata methods can be achieved by machine learning technology, and meanwhile, the knowledge system with labelled metadata can also be data source for machine intelligence.

On the one hand, the acquisition and modeling of metadata can be supported by machine learning; even if it is not yet possible to meet the needs of fully automated processing, it can alleviate the efficiency and cost problems to a great deal. Before the advent of general artificial intelligence (AGI), it is hard to achieve full-scale machine cognition and semantic understanding, and the semantic description of contents generally requires manual work; however, increasing tasks can be completed by computer vision, natural language processing (NLP) and other technologies that are well developed. For example, breakthroughs have been achieved in character and entity recognition, relationship extraction, automatic annotation, text translation and conversion, and there is also discussion about the possibility of automatic generation of metadata on large-scale content [2]. Concept of 'smart data' with self-describing, computable and actionable characteristics was proposed. Triple set that represented by RDF and contains certain causes and effects, sequence or expression axioms is regarded as the simplest form of smart data. With the help of various metadata specification vocabularies and ontological patterns for automatic verification, its management can also be automatically processed, auto-generate semantic data and even construct knowledge graphs.

On the other hand, the description of domain knowledge by metadata can be seen as semantic labels, which can be applied to neural networks and other machine learning algorithms to train model [13], by which the limitation caused by small data set could be overcame, and a more precise learning model or classifier could be created, bringing about more machine learning applications [8]. Currently, digital libraries and many knowledge systems (such as Wikipedia) can be viewed as knowledge buildings

constructed by metadata methods; whether they provide human intelligence directly through digital libraries or provide rich knowledge repositories as training or testing data for machine learning, they are of great value. Some technical experts classify semantic web represented AI research classified into the 'symbolicism', believing that symbolicism should work closely with 'connectionism' which is represented by deep neural networks, as it has made breakthrough but now began to stagnate. This kind of collaboration has in fact begun, and a large number of studies have effectively demonstrated that metadata has a direct impact on the effect of using deep neural networks or other machine learning methods [8, 9, 10]. At present, symbolic neural networks/systems[3] that make machines smarter have become an active field of research, for example, Graph Neural Networks (GNN), Markov Logic Networks (MLN), Knowledge Graph (KG) and Deep Neural Networks (DNN) combined research, are different machine learning tribes trying to collaborate with each other, where metadata can be found all around. This should be a way to break through the bottleneck of machine learning at present.

Intelligence is expected to empower metadata management in the era of AI, by which knowledge creation could be supported by intelligent systems. Worthy of noting, the premise of making machines smart is the manpower: ontology and its specification should be first compiled and generated, and reference rules needs to be encoded. Almost all the traditional methods plus computational thinking require the use of data modeling and resource description, therefore the metadata method is universal. Nevertheless, machine learning, especially neural networks, can only achieve a small scale success in certain areas before it has evolved to general AI [7]. Currently, traditional method is still the primary approach, and metadata even serves as a premise for machine learning. Definitely, we hope that eventually the machine learning approach will succeed, and can take over the role of metadata methods completely, instead of only reinforcing it.

3. METADATA AND METAVERSE

3.1 Metaverse and Web 3.0

The metaverse is a universe of data and technology, just like the Pyramids and the Great Wall, all of which are the products of large-scale collaboration between human being and the latest technology. It is more suitable to say that the Internet is the preceding form of metaverse, rather than saying that the metaverse is the next generation of the Internet. Most of the key characteristics for building metaverse have in fact existed with the development of web 2.0: including a large amount of annotated systems constituting a large-scale space with entities and objects, various types of relationship descriptions linking entities of the entire network, and some extent of interoperability being achieved, supporting the real-time and continuous activities for global users.

Compare with the previous two generations of Internet technologies: Web 1.0 is the content network, in which users consume information; Web 2.0 is the social network, in which users create information; while Web 3.0 is the value network, in which users can possess information assets. From the perspective of

[3] See more at https://www.jiqizhixin.com/articles/101402

metadata, it originated in Web 1.0, succeeded in Web 2.0, and will continue to achieve great value in the Web 3.0. In the Web 1.0 era, metadata has played a role in content disclosure and discovery, where digital libraries rely on metadata, and the rise of search engines such as Google has also used metadata to obtain a considerable commercial scale. In Web 2.0, UGC is on the rise, and through the adoption of the mobile Internet, there emerged many giant company that provide centralized services. At this time, wikis, blogs, podcasts, microblogs, e-commerce, etc. are blooming, and universal relationship descriptions (e.g social networks are relationship computing) have become a necessity, and Really Simple Syndication (RSS) is a typical representative of them. In the Web 3.0 era, it is expected to achieve the real decentralization, where the underlying architecture should be built on blockchain technology, and so any activity and information generated during that can leave traces, which shall bring about decentralized trust (or de-trusting). Through the decentralized identity (DID) which is supported by personal digital wallet, anyone can live a life in metaverse. At the same time, it is also compatible with the centralized model that formed in Web 1.0 and 2.0 era. All the activity in the virtual world is in essence the activity of information, and there is almost no friction; therefore, the relations of production that shaped in the real world would often requires a high cost to manage and operate the digital world. For example, the cost of maintaining the copyright of e-books is much higher than the cost of letting it circulate freely, making its original business model unsustainable, resulting in new models such as traffic patterns and advertising models. Property rights protection in the Web 3.0 era occurs simultaneously with information production, which is an innate attribute of the platform. Certainly, this does not mean that people who own the rights should oppose to open movement, on the contrary, this shall make it more effective in promoting 'open, sharing and free' activities.

3.2 Metadata in Metaverse/Web 3.0

The metaverse consists of multiple layers of data and interconnected protocols [14]. To put it in a simple way, there are content metadata, technical metadata, and management metadata (see Figure 1) [6]. Content metadata is generally the same as that in Web 2.0 era, but it is more complex in Web 3.0 era in the way that there emerges many new data types and formats, which support the new interaction ways such as 3D rendering and spatial computing. For example, there are many interesting properties designed in NFT technique, and the value of it is closely related to combination of those different properties. The modelling of metaverse also relates to ontology construction and attribute description, etc., all of which involves a large number of metadata models. Technical metadata mainly includes system architecture, response mode, interaction specification, service registration and discovery, and specific design of the protocol. Management metadata refers to the life cycle management of all independent digital objects, which is used to ensure that users establish a trust mechanism in each scenario and case, as well as various constraints and settings for the smooth operation of the trust mechanism. Generally speaking, the rules for the operation of metaverse (i.e., DAO) and smart contracts belong to the management metadata.

The metadata in Web 1.0 usually describes offline entities; metadata in Web 2.0 is able to obtain online object files through metadata in a direct way; and metadata in Web3.0 and the objects it describes are all essentially data, which is often difficult to distinguish as they describe each other.

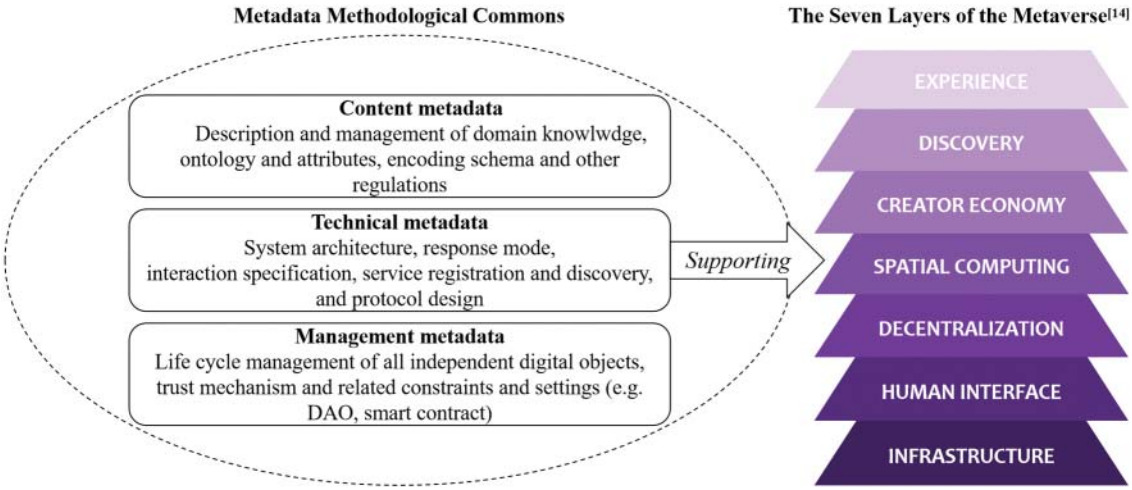

Figure 1. Metadata methodological commons in the metaverse/web 3.0 environment.

Almost all schema innovations in Web 3.0 require metadata. A typical example is the RDF Site Summary or Really Simple Syndication (RSS) which is originated in Web 2.0 era, a content distribution and aggregation standard. In Web 3.0 era, it promoted into a distributed version: RSS3.

As is shown in Figure 2[④], in the traditional RSS2 model, the content created by users is based on a centralized platform, and the platform is then recommended to users through channels or algorithms; in RSS3, the creators themselves control the content, decentralized application (Dapp) pulls user content, and other users subscribe to the creator's content. RSS3 not only automatically aggregates information through metadata, but also guarantees that each content creation can be tracked and continue to receive revenue (if the author has set the corresponding ownership; certainly, many authors share content for free using the Creative Commons License or Creative Commons).

The design of RSS3 is to some extent troublesome, yet very clever. It is proposed by Natural Selection Labs, and it implemented three application components: 'RE:ID', 'Web2Pass', and 'Revery'[⑤]. Respectively, they are implemented to solve problems of user's unique identity (which is bundled with digital wallets, but can obtain an RNS global domain name), personalized home page, and local content asset aggregation and subscription problems, which is suitable for decentralized environments. Compared with traditional RSS, the biggest feature of RSS3 is that once the content is created, it is stored permanently, the author confirms the rights of contents, and the distribution and aggregation are completely controlled by the creator, instead of handing over to the platform. At present, the application of RSS3 is built on the Solana public chain, the cost for content creation is very low, and it also supports low-cost Arweave storage, thus has a very promising application prospect.

④ Figure source: https://blog.rss3.io/platformless-media
⑤ https://blog.rss3.io/

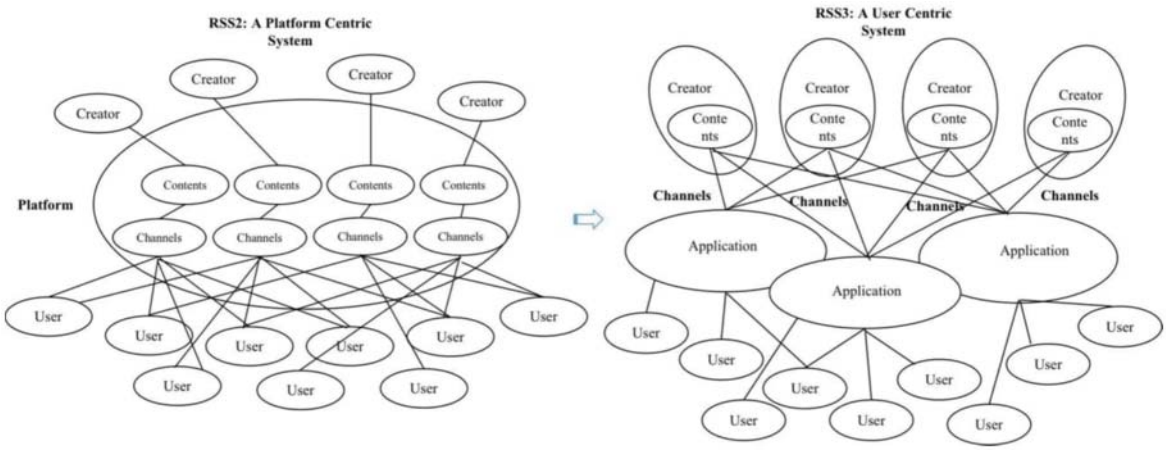

Figure 2. Comparison between RSS2 and RSS3.

The value of metadata got highlighted in the digital age as the need to explicitly describe attributes and provide rules for semantic interoperability of the information body in the distributed information environment, which enriched the function of metadata. Entering to the era of metaverse, the decentralized data architecture of Web 3.0, IoT, edge computing, the ubiquitous machine learning, intelligent services and the interaction of information entities enables the full operation of metadata. The entire building of the metaverse is built on the basis of service discovery and resource discovery, and there shall design a set of description, registration, publishing, discovery, and interoperability mechanisms to assist this basis, which relies on the function of metadata.

4. CONCLUSION

Based on the review of metadata methodological commons, this paper discusses the development of metadata from structural description to semantic coding in the big data environment. It systematically analyzes how metadata is incorporated with machine learning to empower knowledge system construction and knowledge creation with intelligence. As a systematic description specification for knowledge systems, metadata will continue to play an important role in the construction of the Web 3.0 distributed information environment and in the multi-modal metaverse. It is also predictable that this important role will also work further in Web 4.0 (the symbiotic web where human and machines connect and communicate in a symbiotic way) and even Web 5.0, where all forms of data shall be fully grasped by users with the support of sensory systems. Of course just as Web 3.0 did not become a purely semantic Web, it is very hard to make a definitive judgment as to whether Web 4.0 and 5.0 will develop in the direction that people predict. But in any case, the role of metadata may be increasingly hidden behind the scenes, but it will also be increasingly fundamental and therefore increasingly important.

Compared with the large amount of unstructured and semi-structured data in Web 2.0 applications that lack detailed description, which leads to the deficiency of value in data applications, the meaningful fine-grained and ubiquitous linked data and semantic descriptions in Web 3.0 environment makes it an all-inclusive and all-powerful neural network. The value of metadata is fundamental in this context. During this process, metadata needs to be combined with machine learning algorithms to develop a set of methodologies that automatically implement semantic formalization, which is the inevitable way for it to adapt to the needs of the future intelligent era.

AUTHORS' CONTRIBUTIONS

Wei Liu (w.liu@libnet.sh.cn, 0000-0003-2663-7539) contributed to the conception and design of the work, drafting the manuscript, and critical revision of the article. Yaming Fu (ymfu@libnet.sh.cn, 0000-0001-9736-4916) contributed to the writing and revision of the manuscript. Qianqian Liu (qqliu@libnet.sh.cn, 0000-0002-8111-5154) contributed to the writing and revision of the manuscript.

ACKNOWLEDGEMENTS

This work is supported by the National Social Science Foundation (Grant/Award Number: 21&ZD334)

CONFLICT OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

REFERENCES

[1] Garlinghouse, T.: The rise and fall of the Great Library of Alexandria. Livescience (2022). Available at: https://www.livescience.com/rise-and-fall-of-the-great-alexandria-library. Accessed 2 December 2022

[2] Han, H., Giles, C.L., Manavoglu, E., et al.: Automatic document metadata extraction using support vector machines. In: Joint Conference on Digital Libraries, pp. 37–48 (2003)

[3] Schelter, S., Boese, J.H., Kirschnick, J., et al.: Automatically tracking metadata and provenance of machine learning experiments. In: Machine Learning Systems Workshop at NIPS, pp. 27–29 (2017)

[4] Leipzig, J., Nüst, D, Hoyt, C.T., et al.: The role of metadata in reproducible computational research. Patterns 2(9), 100322 (2021)

[5] Ulrich, H., Kock-Schoppenhauer, A.K., Deppenwiese, N., et al.: Understanding the Nature of Metadata: Systematic Review. Journal of Medical Internet Research 24(1), e25440 (2022)

[6] Gong, C.: White Paper of China metaverse development (2022). Available at: https://www.healthit.cn/wp-content/uploads/2022/01/2022中国元宇宙白皮书-龚才春.pdf. Accessed 2 December 2022

[7] Marcus, G.: The next decade in AI: Four Steps Towards Robust Artificial Intelligence (2022). Available at: https://arxiv.org/vc/arxiv/papers/2002/2002.06177v2.pdf. Accessed 2 December 2022

[8] Greenberg, J.: Big metadata, smart metadata, and metadata capital: toward greater synergy between data science and metadata. Journal of Data and Information Science 2(3), 19–36 (2017)
[9] Leipzig, J., Bakis, Y., Wang, X., et al.: Biodiversity image quality metadata augments Convolutional neural network classification of fish species. In: Research Conference on Metadata and Semantics Research, pp. 3–12. Springer, Cham (2020)
[10] Greenberg, J., Zhao, X., Monselise, M., et al.: Knowledge Organization Systems: A Network for AI with Helping Interdisciplinary Vocabulary Engineering. Cataloging & Classification Quarterly 59(8), 720–739 (2021)
[11] Liu, W., Lou, X., Zhao, L.: The History, Present and Future of Dublin Core Metadata (2005). Available at: http://eprints.rclis.org/6077/. Accessed 2 December 2022
[12] Duval, E., Hodgins, W., Sutton, S., et al.: Metadata principles and practicalities. D-lib Magazine 8(4), 1–10 (2002)
[13] Ostendorff, M., Bourgonje, P., Berger, M., et al.: Enriching bert with knowledge graph embeddings for document classification. arXiv preprint arXiv:1909.08402 (2019)
[14] Radoff, J.: The Seven Layers of the Metaverse (2021). Building the Metaverse. Available at: https://medium.com/building-the-metaverse/weekly-kickoff-april-5-2021-ea39a4e22e3. Accessed 2 December 2022

AUTHOR BIOGRAPHY

Wei Liu (1966-), male, Ph.D. Deputy Director of Shanghai Library (Institute of Scientific and Technical Information of Shanghai), adjunct professor of Fudan University, East China Normal University and Shanghai University. Research direction: knowledge organization, digital library management, digital humanities. E-mail: w.liu@libnet.sh.cn.
ORCID: 0000-0003-2663-7539

Yaming Fu (1993-), female, Ph.D. graduated from the University College London (UCL), postdoctoral researcher at Shanghai Library (Institute of Scientific and Technical Information of Shanghai) and the School of Information Management at Nanjing University. Research direction: digital library management, digital humanities. E-mail: ymfu@libnet.sh.cn.
ORCID: 0000-0001-9736-4916

Qianqian Liu (1985-), female, data librarian of Shanghai Library (Institute of Scientific and Technical Information of Shanghai). Research direction: digital humanities, data processing and platform construction. E-mail:qqliu@libnet.sh.cn.
ORCID: 0000-0002-8111-5154