Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **7**, Number **2**, May **2022**

A Novel Network Security Data Resource Description Standard

Wei Han

Department of Electronic and Information Engineering
Laiwu Vocational and Technical College
No. 1 Shancai Street, Jinan, Shandong, China
bluesky_han@126.com

Xiu-Yan Sun

Department of Mechanical and Electrical Engineering
Laiwu Vocational and Technical College
No. 1 Shancai Street, Jinan, Shandong, China
sunxiuyan0634@163.com

Chen He

Media and Communications College
WeiFang University
No. 5147 Dongfeng East Street, Weifang, Shandong, China
61604638@qq.com

Lin-Lin Tang*

Department of Computer Science and Engineering
Harbin Institute of Technology, Shenzhen
Taoyuan Street, Shenzhen, China
Corresponding author: hittang@126.com

Saru Kumari

Department of Mathematics
Chaudhary Charan Singh University
Meerut, Uttar Pradesh 250004, India
saryusiirohi@gmail.com

Received August 2021; revised October 2021

Abstract. *Current network security situational awareness field has problems of poor communication and low efficiency of data resources and information. Therefore, in order to solve these two problems, this paper has established a better network security data resource description standard (metadata standard) for communication of data resource information, so that communication of data resource information has a unified standard basis. Then it provides a more efficient channel for circulation of data resource information, and improve efficiency of data resource information flow through publishing and subscription technology. Finally, a data resource sharing platform was developed, which applied the above metadata standards.*
Keywords: Network Security, Metadata, Ontology, Publish/Subscribe

1. **Introduction.** Data is expression form and carrier of information. It is a broad concept. Any digital resource carrying information can be called data in a broad sense. Therefore, data resource description language is also a broad concept. Description of data resources has been going on since data is stored by computer. Where there is data, there is metadata. Network security data resource description language studied in this paper is a narrow concept. Firstly, object described is data resources involved in network security research. Secondly, main goal of this description language is to achieve better data resource publishing and subscription, and serve participants of machine and data exchange at the same time. Based on this, related research of data resource description language mainly includes data and intelligence knowledge base [1, 2], situational awareness architecture [3], metadata management in smart grid field [4], library information management and scientific data warehouse management [5]. These research work are based on metadata, So, establishment of data resource description language in this paper starts from metadata standard.

Metadata is an important concept in information organization and management. There are many existing metadata standards. It is applied to different types of resource descriptions, which can be summarized as follows:

(1) Network resources: DC [6] (Dublin Core), ROADS, Web Collections [7], Zcollection [8] and CDF [9] (Channel Defination Format).

(2) Literature: MARC [10], TEL Header(Text Encoding Initiative Header) and EAD [11] (Encoding Archival Description).

(3) Digital museum resources: CDWA [12] (category description of works of Art),VAR (core category of visual resources).

(4) Educational resources: ADL/SCORM [13] (a content aggregation specification), DCED [14], gem [15], etc.

(5) Geographic Information Resources: FGDC / csdgm [16].

(6) Scientific database: HCLs [17], dats [5], etc.

Most widely used metadata standard is DC (Dublin Core). Its 15 elements. DC is mainly used for resource discovery and is compatible with a large number of metadata standards. Marc (machine-readable directory format) standard is mainly used for resource description. Usually, in a metadata application scenario, multiple standards will be applied and integrated at the same time to achieve their application objectives.

As resource collection, network security data set needs metadata suitable for resource collection. At present, there are many metadata standards that can be used to describe resource collection, such as Zcollection, RSS [18] (RDF/Rich Site Summary, website information aggregation), RSLP Collection Description Schema [19] (RSLP collection description mode), ISAD (g) [20] (international standard file description standard) Gils (government information location service), DC CD AP (Dublin core description configuration file) and webml [21] (website modeling language), etc. the establishment idea of these standards is mainly to expand on the existing standards.

This paper analyzes a large number of existing metadata standards and metadata management methods in data systems, establishes a metadata model of network security data resources, adds semantic information to data resources, and aligns with OWL ontology, so that this standard has better re-usability and interoperability, and provides standard data resource description for data producers and consumers, establishes good communication standards [22-24].

2. **Relate Work.**

2.1. **Data of Network Security.** Network security situational awareness depends on the widest possible security event information in network environment. Collecting, integrating, merging, aligning and analyzing information can dynamically reflect security status of the network and predict potential threats. It should deal with rapidly changing network environment to reduce probability of network security accidents, Goal of timely response to network security events and timely stop loss. Situation awareness is not only an important means of network security monitoring, but also a research hotspot.

Data resource aggregation method adopted in this paper is different from yhssas system. It is not highly integrated, but a loose form similar to data mart and data lake, but does not store data resources. Compared with state perception system, life cycle of data resources should be integrated in metadata rather than agreed by process of platform. Therefore, metadata of data resources should include life cycle of data, such as how the data has been preprocessed, whether data is in deep processing state or original data? whether data has been desensitized, followed by granularity of data? whether the represented data exists in the form of structure and data (schema + data) or just a data record, or is it just an attribute value of a record?

2.2. **Semantic Relationship of Network Security Data.** Combined with classification idea in the research on establishment of situation knOWLedge base, this paper extends system class to more general context environment class, consequence class and network event class are promoted. In addition, vulnerability data is introduced, that is, data resources can be divided into: Context data (also known as asset data), vulnerability data (also known as vulnerability data), attack data, event data (also known as log and traffic data). And their relation can be shown in the following Figure 1.

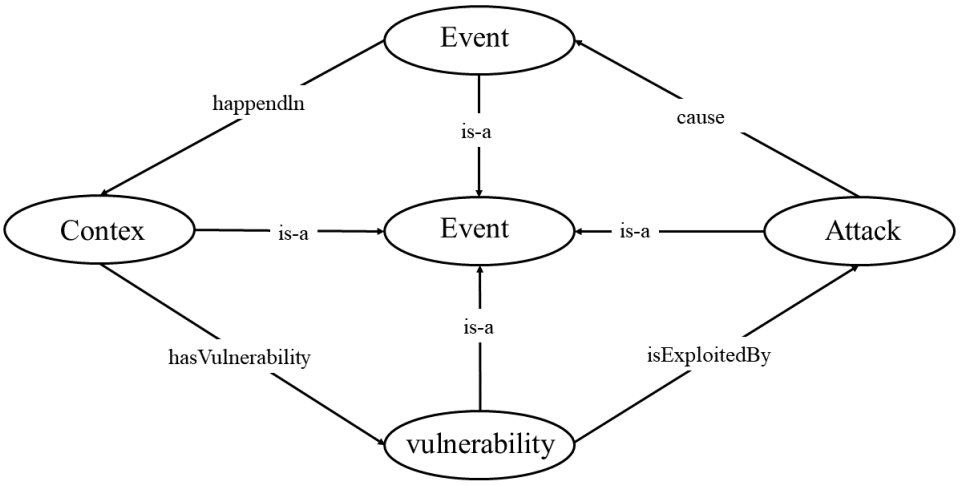

Figure 1. Semantic Relationship of Four Types of Data

Attack data includes attack attributes, tools or software used by the attack, attack results, security status of attack object and attacker information. They have a semantic relationship with attack data, which attack attribute information attack has, what attack is carried out with, impact of the attack, what security state the attack needs, and what attacker caused the attack, as shown in Figure 2.

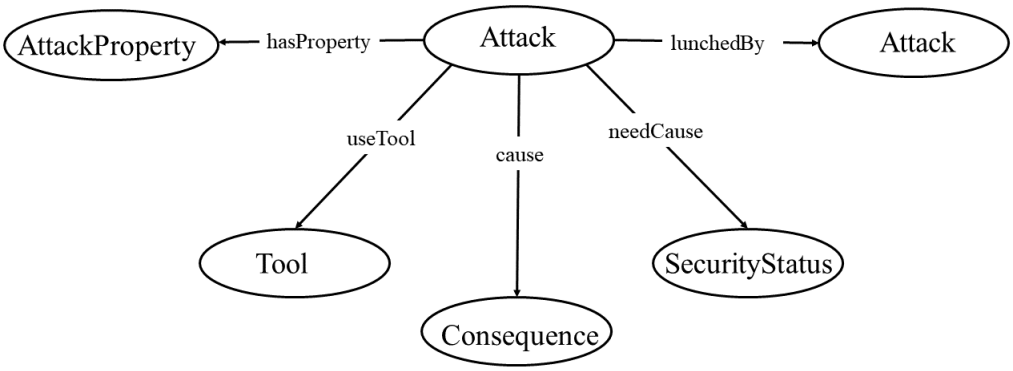

FIGURE 2. Semantic Relationship Between Attack Class and Its Subclasses

Some status data can be obtained by statistics or conversion of log data, so there is semantics of generating sampling data from log data. As shown in Figure 3.

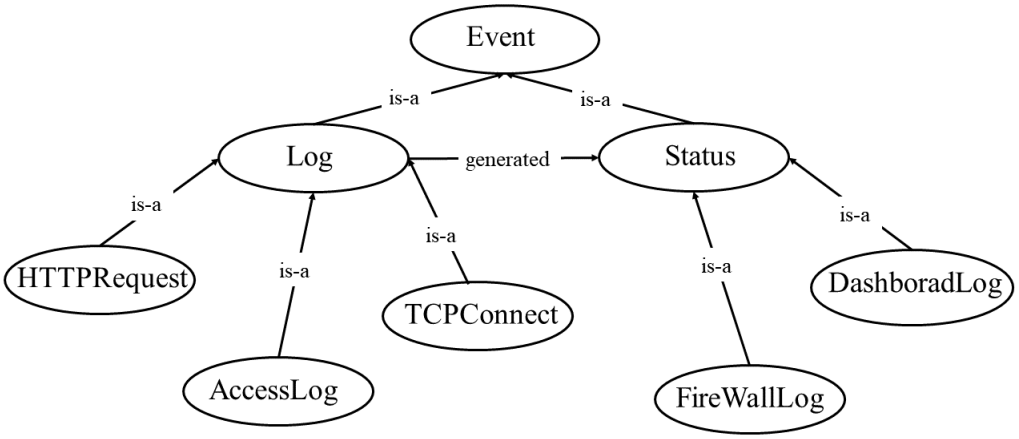

FIGURE 3. Semantic Relationship Between Event Class and Its Subclasses

3. **Our Proposed Method.**

3.1. **Classification and Enumeration of Metadata Elements.** There are many classification standards. According to generation method, metadata can be divided into the fully automatic generation, the semi-automatic generation and the fully manual generation; According to necessity, it can be divided into mandatory metadata, recommended metadata and optional metadata; According to the requirements hierarchy, it can be divided into storage/access information, description/index information and use case analysis information; According to the function classification of metadata elements, they are

divided into index type, management type, storage type and professional technology type. Among them, functional classification of metadata elements is the most mainstream. This section lists the metadata of various categories in this way.

Managed metadata is used to manage life cycle of data resources in data sharing platform and clarify ownership of data resources, mainly including:

(1) Data source: the context in which the data is collected.

(2) Data producer: the person responsible for the reliability of data content, which can be software.

(3) Data publisher: the person who provides data intelligence to the data sharing platform, which is different from the producer.

(4) Release time: the time when the data resource information can be found by others after being completed and submitted successfully.

(5) Version number: data resources may be updated periodically and non incrementally, and the version number is recorded.

(6) Responsible department: the enterprise, organization and department to which the publisher belongs represent its authoritative information.

(7) Privacy permission: it is used to specify which role can be authorized to access data.

(8) Copyright notice: similar to the open source software protocol, it specifies the extent to which data resources can be used.

(9) Data resource type: whether data resources are provided in the form of files, data warehouse links, or service interfaces.

(10) Dataset format: whether data resources exist in structured format or in the form of pictures, XML documents, etc.

(11) Unique identifier: similar to ID card, it corresponds to data resources one by one on the data sharing platform.

Descriptive metadata is used to provide data subscribers with detailed data resource information and facilitate subscribers to understand functions of data resources, mainly including:

(1) Data resource name: a high summary of data resource information.

(2) Key words: technology, field and application scenario vocabulary involved in data resources, which serves the retrieval of data resources.

(3) Introduction: similar to abstract of paper, it mainly introduces the characteristics of data resources.

(4) Data abstract structure: what kind of abstract structure is data: one-dimensional sequence, two-dimensional table, tree or graph.

(5) Dataset size: how many records does the dataset contain.

(6) Data technical documents: technical documents that facilitate data users to apply data resources.

Storage metadata describes how data resources are stored and determines how data users obtain data resources, mainly including:

(1) Data source URL: such as database access URL, data file download address, data interface address, etc.

(2) Access user: a specific user name or visitor.

(3) Access authentication: password or access token, etc.

(4) Data structure information: This is a composite structure, similar to the schema of relational database. It defines the table / partition, column, index and constraint of data.

(a) Table name, table remarks, table code and table serial number (applicable to the vertical or horizontal splitting of the table).

(b) Column name, column comment, data type, constraint condition.

(c) Index definition: reflects the index type and the columns involved in the index, such as unique index (ID).

Technical metadata is oriented to technicians and has many contents. It is usually divided into different subcategories. The elements involve metadata in specific fields, as listed below:

(1) Physical metadata: metadata describing physical resources, including but not limited to the following metadata.

(a) Server address: server IP, domain name, etc.

(b) Device platform: the computing power and vulnerabilities of devices on different platforms are different, such as x86, PowerPC and arm.

(c) Machine room location: context data relates to geographical location of the server.

(d) Time span: difference between the end time and start time of data resource collection or size of time window.

(e) Spatial span: geospatial of data collection.

(2) Evaluation metadata: evaluation information of data.

(a) Timeliness of data: different types of data have different timeliness.

(b) Data confidence: determined by confidence and history of publisher's organization.

(c) Subscription amount of data: Statistics of data sharing platform.

(d) Data heat: set a weight inversely proportional to current time length. The sum of the product of the weight and the number of subscriptions in the time period can be used as the data heat.

(e) Scoring of data resources.

(3) Statistical metadata: statistical summary of data resources, applicable to columns of relational data, including but not limited to the following metadata.

(a) Frequent item sequence of data values: applicable to non numeric data attributes.

(b) Degrees of freedom: applicable to non numeric attributes, and counts the number of optional cases of label.

(c) Frequency histogram sequence: applicable to numerical data attributes.

(d) Quantile: applicable to numerical type, including 0, 25, 50, 75100 quantiles.

(e) N-order moment sequence: numerical mean, variance, skewness and kurtosis.

(f) Information entropy: the information density of reaction attributes.

Since the user is not only data producer but also data consumer on data sharing platform, there are a lot of interactions between the user and the whole system, which should not be just a point, but a face. Therefore, it is necessary to describe the user's attributes and the interaction records of the user's data resources in detail. The elements describing the user are defined as:

(1) User name.

(2) User ID: this ID is only visible to the machine and is used to uniquely identify the user.

(3) Gender.

(4) Age.

(5) Department: used to better discover the relationship between data subscribers or publishers.

(6) Department level: used to mark the hierarchical relationship between departments and restore the organization structure.

(7) Address: user's region information.

(8) Role: the user's role in the system.

There is a many to many relationship between users and data. The following metadata is used to record the interaction between users and data resources on the data sharing platform:

(1) Unique identifier of the dataset.

(2) User ID.

(3) Whether the user subscribes to the data resource: as the user's subscription record.

(4) How often users view the resource information: clicking on it many times means they are interested.

(5) Total time for users to view the resource information: the time from the interface being clicked to out of focus (mouse click outside the interface or jump out of the interface).

(6) The integrity of the user's view of the resource information is similar to that obtained by monitoring whether the user slOWLy scrolls the product details interface to the end of the page in e-commerce.

3.2. **Compatibility of Metadata with Existing Metadata Standards.** According to the interoperability and reusability in fair principle, it can be seen that metadata standard of new invention should be compatible with existing authoritative metadata standards, and it is best to expand or reuse on existing metadata standards. Reuse principles are as follows:

(1) The principle of being reused should be a mature and stable metadata standard.

(2) The semantics of the reused metadata element shall be consistent with the original, and there shall be no semantic ambiguity, incomplete reference and semantic intersection.

(3) You can reuse only part of the elements, not all of them.

(4) Try to use the element name consistent with element in the reused standard, but do not restrict renaming or adding modification.

(5) Different data granularity is common, so the element mapping during reuse can be one to many (one element corresponds to multiple multiplexed elements) and many to one (multiple elements and one multiplexed element pair).

Most influential element set in metadata field is DC (Dublin Core), which itself serves description and publication. Data resource publishing and subscription studied in this paper is equivalent to non strict description and publication. Therefore, all DC elements can be reused in theory. It is verified that elements contained in metadata listed in previous section can correspond to all 15 elements in DC core. For the mapping of data, especially the elements with different granularity of the same concept, attribute mapping function is added in the system development to complete the mapping from the original data resource attributes to the standards in this paper.

3.3. **Alignment of Metadata Standard and OWL Ontology.** This section attempts to organize metadata according to OWL syntax. Basic idea is shown in Figure 4 below. Data resource class and metadata elements classified according to purpose in previous section are in middle object layer, metadata of specific data resources is in data layer, and OWL semantic layer at top specifies ontology semantics of objects and relationships between objects in the object layer.

Ontology is a concept in the field of philosophy, which is used to refer to objective things. In field of artificial intelligence, ontology is used to formally describe the concepts in a certain field, the relationships between concepts, and the boundaries and constraints of concepts, so that these concepts can be processed and reused by computers. In OWL, ontology is composed of classes, properties and individuals, and instances are also called individuals. An individual represents a specific object in the domain, such as "Ma Yun is a specific object in the Chinese domain". An object may have multiple names. Whether the objects are the same should be explicitly declared in OWL, such as "Jack Ma 'OWL:

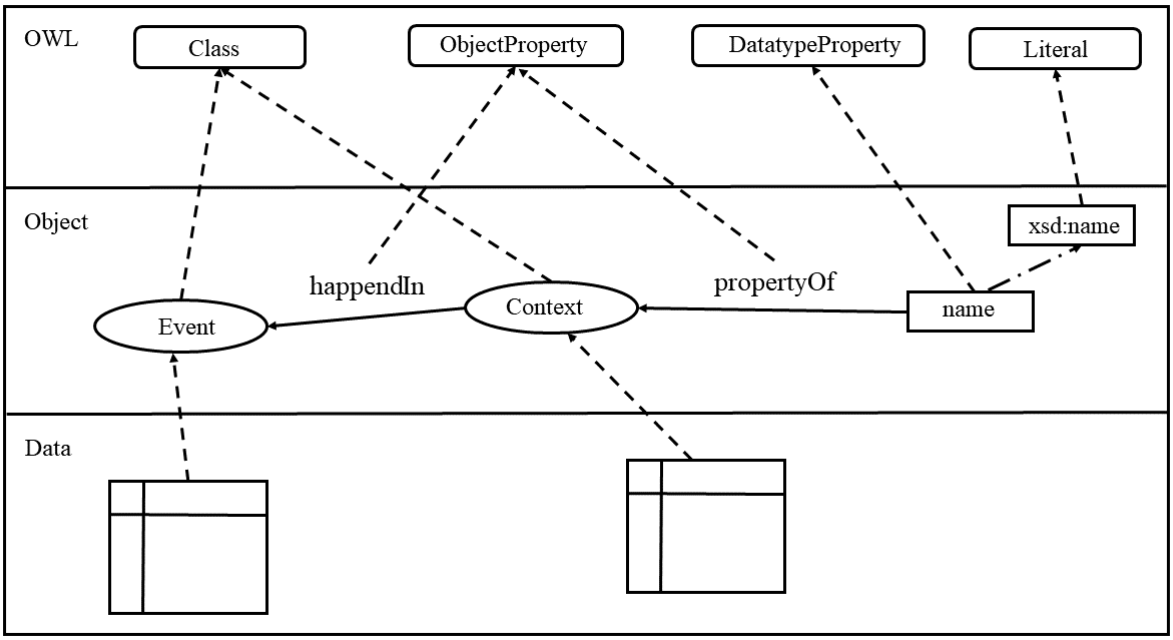

FIGURE 4. Metadata Standardization Semantics

sameindividual' ma Yun". An individual is an instance of a class, and an individual may belong to multiple classes. Property is the relationship between instances and is used to connect two instances. It means the same as attributes. It can be seen from OWL definition document that properties are also divided into datatypeproperty and objectproperty. Datatypeproperty indicates that property is a literal or value type, while objectproperty indicates that property is a complex structure, and objectproperty connects instances, relationship between instances is an axiomatic concept in OWL DL. A class is a collection of instances. The attributes of a class highly summarize structure and meaning of its member instances. There are a variety of relationships between classes, such as parent subclass relationship (OWL: subclassof), equivalent class relationship (OWL: equivalentclasses), and non intersecting (orthogonal) class relationship (OWL: disjointclass).

3.4. **Metadata Model Analysis.** OWL can support multiple representations, including RDF / XML (RDF described by XML syntax), Turtle (N-Triples, Turtle, Trig and N-Quads), JSON-LD (JSON based format to link data) and RDFA (nested format of XML and HTML). These methods can generally be divided into three categories: XML, turtle, and JSON. Turtle representation method is a line based plain text representation. Each line is an RDF triplet (subject, assertion and object). The three parts are separated by spaces, and the end of each triplet ends with ".". This method has the advantages of simple processing, no dependency between lines, fast reading and writing speed, and no nested structure, Human reading can not directly feel relationship network, and its readability is poor. XML has the most mature method and the most common support. XML has the strongest expression ability and fast parsing speed. Disadvantage is that there are a large number of structural text in the format and high redundancy. JSON based method is slightly weaker than XML in expression ability, much stronger than turtle, and better than XML in readability. It is convenient to convert it into common data structures "dictionary" and "array". In other words, JSON is composed of nested

dictionaries and arrays, and has been widely used in web application data transmission in recent years, with good compatibility, Therefore, metadata persistence in this paper is based on JSON-LD.

Important difference between JSON-LD and JSON is that it predefines some keywords, such as "@ context" defines context of JSON-LD document, which can be used to define specification of document, "@ base" can define relative IRI of document, which can make JSON-LD document concise and easy to read, "@ include" can refer to other node objects, and "@ version" represents the version number.

Metadata specification, alignment with OWL and persistence to JSON-LD proposed in this section follows the corresponding standards, while OWL full and JSON-LD have good scalability. In addition, ontology in field of network security will change with the development of the field of network security. This paper does not limit addition of subclasses to created network security data classes. There is no clear definition of required fields of various data resources. Such work can only be formulated by experts in corresponding field, such as data producer or data demander. Data sharing platform will provide such operability to data sharing participants. Relationship between network security data classes and mapping of owl objects designed in our proposed method is shown in the following Table 1.

TABLE 1. Relationship Between Network Security Data Classes and Mapping of OWL Objects

Attributes	Domain	Range	Meaning
hasVulnrability	Contex	Vulnerability	There are loopholes in the network environment.
isExploitedBy	Vulnerability	Attack	What are the vulnerabilities exploited?
cause	Attack	Event	Events caused by the attack.
happendIn	Event	Contex	Network environment in which the event occurred.
lunchedBy	Attack	Attacker	The initiator of the attack.
needCause	Attack	SecurityStatus	Conditions required for attack.
useTool Attack	Tool	Tools	used for attack.
hasProperty	Attack	AttackProperty	Attack properties.
isObjectOf	VulObject	Vulnerability	What kind of vulnerability does the vulnerability subject belong to?
isProperty	VulProperty	Vulnerability	Vulnerability to which the vulnerability attribute belongs.
exploited	ExploitMethod	Vulnerability	What vulnerabilities can be exploited by vulnerability exploitation methods?
isEndnodeOf	Endnode	Network	Network to which the network terminal belongs.
runIn	System	Endnode	Network terminal for system operation.
instanceOf	Service/Process	Program	Which program file runs the service or process?
generated	Status	Log	What logs can be generated for status?
control	Endnode/System	Sensor/Peripherals	Terminal or system controlled peripherals.

4. **Conclusions.** Firstly, this paper studies the use methods of data in the field of network security. Then, combined with the existing research on network security data, we classify semantic relationships between network security data resource classes, subclasses and different types of data. And we also list the required elements in detail, then reuse OWL semantics, and finally briefly discusses its persistence scheme Metadata retrieval scheme and scalability.

Acknowledgment. This work is supported by Shenzhen Foundational Research Funding JCYJ20180507183527919 and Shenzhen Fundamental Research Funding JCYJ20180306171938767.

REFERENCES

[1] F. Ravat, Y. Zhao, Metadata management for data lakes, *European Conference on Advances in Databases and Information Systems*, pp.37–44, 2019.

[2] D. Chaves, E. Malinowski, Document Data Modeling: A Conceptual Perspective, *European Conference on Advances in Databases and Information Systems*, pp.19–27, 2019.

[3] S. Liu, H. Liu, V. John, Z. Liu, E. Blasch, Enhanced situation awareness through CNN-based deep multimodal image fusion, *Optical Engineering*, vol.59, no.5, pp.053–103, 2020.

[4] T. H. Nguyen, V. Nunavath, A. Prinz, Big data metadata management in smart grids, *Big data and internet of things: A Roadmap for Smart Environments*, pp.189–214, 2014.

[5] S. A. Sansone, A. Gonzalez-Beltran, P. Rocca-Serra, G. Alter, J. S. Grethe, H. Xu, I. M. Fore, J. Lyle, A. E. Gururaj, X. Chen, H. Kim, N. Zong, Y. Li, R. Liu, I. B. Ozyurt, L. Ohno-Machado, DATS, the data tag suite to enable discoverability of datasets, *Scientific Data*, vol.4, no.1, pp.1–8, 2017.

[6] D. C. M. Initiative, Dublin core metadata element set, version 1.1, 2017, http://hdl.handle.net/10421/3401.

[7] J. Cho, N. Shivakumar, H. Garcia-Molina, Finding replicated web collections, *Acm Sigmod Record*, vol.29, no.2, pp.355–366, 2000.

[8] J. Zhao, Y. Song, X. L. Wang, Study on the resource collection metadata standard system based on" Zcollection", *Manufacturing Automation*, vol.2, 2005.

[9] C. Ellerma, Channel definition format (CDF), *WorldWideWeb Consortium*, pp.66–84, 1997.

[10] D. Ivanović, D. Surla, Z. Konjović, CERIF compatible data model based on MARC 21 format, *The Electronic Library*, pp.46–74, 2011.

[11] E. H. Dow, Encoded Archival Description as a Halfway Technology, *Journal of Library Metadata*, vol.7,no.3, pp.108–115, 2009.

[12] M. Baca, P. Harpring. Categories for the Description of Works of Art, *The Getty Research Institute*, pp.98–104, 2016.

[13] M. Rey-López, R. Redondo, A. F. Vilas, J. Pazos-Arias, J. G. Duque, A. Gil-Solla, M. Cabrer, An extension to the ADL SCORM standard to support adaptivity: The t-learning case-study, *Computer Standards and Interfaces*, vol.31, no.2, pp.309–318, 2009.

[14] M. M. Jalil, Practical Guidelines for Conducting Research - Summarising Good Research Practice in Line with the DCED Standard, *Social Science Research Network Electronic Journal*, pp.123–144 2013.

[15] T. Lalor, S. Vale, A. Gregory, Generic Statistical Information Model (GSIM), *The North American Data Documentation Initiative Conference*, pp.24–45, 2013.

[16] D. A. Ignizio, M. S. O'Donnell, C. B. Talbert, Metadata Wizard: an easy-to-use tool for creating FGDC-CSDGM metadata for geospatial datasets in ESRI ArcGIS Desktop, *The United States Geological Survey Open-File Report*, pp.45, 2014.

[17] A. J. Gray, J. Baran, M. Marshall, M. Dumontier, Dataset Descriptions: HCLS Community Profile, *Semantic Web in Health Care and Life Sciences Interest Group*, pp.21–47, 2015.

[18] P. R. Babu, Measuring Research in RSS Feed Literature: A Scientometric Study, *Measuring and Implementing Altmetrics in Library and Information Science Research*, pp.74–86, 2020.

[19] S. Giannoulakis, N. Tsapatsoulis, N. Grammalidis, Metadata for Intangible Cultural Heritage, *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp.634–645, 2018.

[20] E. Shepherd, C. Smith. The Application of ISAD(G) to the Description of Archival Datasets, *Journal of the Society of Archivists*, vol.21,no.1, pp.55–86, 2000.

[21] S. Ceri, P. Fraternali, A. Bongio, Web Modeling Language (WebML): a Modeling Language for Designing Web Sites, *Computer Networks*, pp.137–157, 2009.

[22] P. Wang, C. M. Chen, S. Kumari, M. Shojafar, R. Tafazolli, Y. N. Liu, HDMA: Hybrid D2D Message Authentication Scheme for 5G-Enabled VANETs, *IEEE Transactions on Intelligent Transportation Systems*, vol.22,no.8, pp.5071–5080, 2021.

[23] C. T. Li, C. C. Lee, C. Y. Weng, C. M. Chen: Towards Secure Authenticating of Cache in the Reader for RFID-based IoT Systems, *Peer-to-Peer Networking and Applications*, vol.11,no.1, pp.198–208, 2018.

[24] C. M. Chen, W. Fang, K. H. Wang, T. Y. Wu, Comments on An improved secure and efficient password and chaos-based two-party key agreement protocol, *Nonlinear Dynamics*, vol.87,no.3, pp.2072–2075, 2017.