Organizacija znanja, 28 (1–2), 2023, 2328007, https://doi.org/10.3359/oz2328007
Strokovni članek / Professional article

RDA in the Fifth Information Age: entity, identity, authority

RDA v peti informacijski dobi: entiteta, identiteta, kontrola

Gordon Dunsire[1]

ABSTRACT: The Fifth Information Age is characterised by the ubiquitous recording of social transactions in an explosion of persistent cultural memory. A person who contributes to social media is an agent who creates digital cultural artefacts. The internet of things, of personal devices for the capture, creation, transmission, and output of information means everyone is potentially an author, a publisher, a distributor, with a profound impact on traditional methods of authority control. The focus of authority control is shifting from the provision of a unique label for an individual person or group to the description of an individual as an instance of a class, as an entity. This paper will discuss some of the issues in entity-based cataloguing and the tools provided to resolve and accommodate them in *RDA: Resource Description and Access*. These include the implementation of entities and relationships defined in the *IFLA Library Reference Model;* the categorization of entity labels as unstructured, structured, and identifier strings; and the development of implementation scenarios to encompass linked open data. The infrastructure that allows any person to be the creator of a cultural resource also allows any person to describe the resource; everyone is potentially a cataloguer, a metadata creator. The paper describes the general approach to metadata provenance that is provided in RDA and how it can be used to authenticate metadata from a wide variety of sources, including crowd-sourcing, and the application of automatic data inferencing to interoperate metadata from multiple sources.

KEYWORDS: bibliographic control, social media, resource description, standards

IZVLEČEK: Za peto informacijsko dobo je značilno vseprisotno zajemanje družbenih transakcij v eksploziji trajnega kulturnega spomina. Oseba, ki prispeva v družbene medije, je akter, ki ustvarja digitalne kulturne artefakte. Internet stvari, osebnih naprav za zajemanje, ustvarjanje, prenos in izdajanje informacij pomeni, da je vsakdo lahko avtor, založnik, distributer z močnim vplivom na tradicionalne načine normativne kontrole. Osredotočenost normativne kontrole se premika od nudenja enoznačne oznake za posameznika ali skupino k opisu posameznika kot primera razreda, kot entitete. V tem prispevku bomo obravnavali nekatere od problemov, povezanih s katalogizacijo entitet, in orodja, ki so nam na voljo za reševanje in odpravljanje teh problemov v *RDA: Resource Description and Access.* Sem spada implementacija entitet in razmerij, opredeljenih v modelu *IFLA Library Reference Model*; kategorizacija oznak entitet kot nestrukturirani nizi, strukturirani nizi in nizi identifikatorjev; ter razvoj implementacijskih scenarijev za pokritje povezanih podatkov v prostem dostopu. V infrastrukturi, kjer je lahko vsaka oseba ustvarjalec kulturnih virov, lahko tudi vsaka oseba opiše vir; vsak je lahko potencialno katalogizator, ustvarjalec metapodatkov. V prispevku bo opisan splošni pristop k izvoru metapodatkov, ki je na voljo v RDA, in kako se lahko uporabi za preverjanje pristnosti metapodatkov iz širokega nabora virov, vključno z množičnim zunanjim izvajanjem in uporabo samodejnega povzemanja podatkov za interoperabilnost metapodatkov iz več virov.

KLJUČNE BESEDE: bibliografska kontrola, družbeni mediji, opis virov, standardi

The article was created based on the presentation with the same title at the COBISS Conference 2022, 19–22 October 2022.

[1] Gordon Dunsire, Independent Consultant, Edinburgh, United Kingdom, gordon@gordondunsire.com.

1 Introduction

S. R. Ranganathan’s simple model of the function of bibliographic metadata is »Every reader [their] book« (Ranganathan, 1931). That is, the primary purpose of catalogues and other finding aids is to allow anyone to access the information resources that they require for cultural, educational, and entertainment reasons. The human need for access to cultural information recorded by their forebears is evident from the survival of cave paintings and portable objects that seem to have only ritualistic functions. These cultural heritage items seem to be created with future generations in mind, and later human interaction seems to respect that intention. The relationship between the »reader« and the »book« stretches back over millennia, at least 50,000 years (Gerber, 2022). If we accept that these activities are essential to »being human«, the development of metadata for information retrieval services is driven by changes in the characteristics of resources and the means of obtaining them.

From time to time a technology is invented that has a profound and lasting impact on this relationship, as well as on human social activity in general. These technologies are the invention of writing, of movable type printing, of telecommunication and digital information, and of the Internet. Each technology marks a significant jump in the characteristics of recorded information and how it is accessed, and divides the continuum of this relationship into a set of five »information ages« (Dunsire, 2022).

In the current Fifth information age, the »book« is any information resource with any kind of content embodied in any kind of carrier. Types of content include text, images, and music, while types of carrier range from printed volumes and videodiscs to online files. The role of an ordinary person as a mere »reader« is no longer confined to obtaining the content and has expanded to any interaction with an information resource, including creation, amendment, publication, reproduction, and distribution activities, as well as the creation of metadata that describes the resource. There is widespread disintermediation in the lifecycle of information resources with no guarantee that professional standards or ethics are applied. This is a result of the development of the Internet and the Semantic Web, and the availability of personal »smart« devices for capturing and recording information. The quantity of recorded memory has increased but the quality is more difficult to assess. The sources of bibliographic and cultural heritage items for collection, preservation, and access by future readers are more diffuse. All of this presents a challenge for metadata and its application in connecting the reader to their book.

2 RDA entities and elements

Ranganathan’s simple model was formulated in the middle of the Fourth information age, 80 years after the development of Morse code and 50 years before the beginning of the Internet. It now requires extension and refinement to reflect the complexity of the interactions between the »reader« and the »book« in information retrieval services and the metadata that drive them. An integrated set of tools to support the development of new approaches to metadata is offered by the current iteration of RDA: resource description and access (RDA Steering Committee, 2023). RDA is a successor to the Anglo-American Cataloguing Rules that reflects a shift from a specific cultural view of the reader and their book to a more neutral and international treatment of bibliographic and cultural heritage metadata (Dunsire, 2020).

RDA is an implementation of the IFLA Library Reference Mode (LRM) (Riva, Le Bœuf, and Žumer, 2017). The LRM is the »Definition of a conceptual reference model to provide a framework for the analysis of non-administrative metadata relating to library resources«. RDA refines the entities, relationships, and attributes of this framework to reflect the requirements for library and cultural heritage metadata in practical information retrieval systems, including the complexity of the LRM's user tasks to find, identify, select, and obtain an information resource, and to explore the context of a resource with respect to other resources.

The LRM treats the reader as a single person or a group of persons that interacts with a resource. In RDA this is represented by a hierarchy of classes of entities, from the broadest Agent class to the specific Person, Family, and Corporate Body classes. The Family and Corporate Body classes are refinements of the LRM. An important constraint in the LRM is that the persons involved must be »real« and not fictitious, supernatural, or non-human. This reflects an assumption that the recording of and subsequent access to an information resource is an activity unique to humans.

The LRM treats the metadata that describe the book as a set of interconnected entities that reflect various aspects and characteristics of an information resource. This is the so-called »WEMI stack«. The Work, Expression, Manifestation, and Item classes are aspects of a resource that are linked with »primary« relationships that maintain the integrity of the metadata for the resource as a whole. The Work and Expression aspects cover the content and context of the resource. The boundaries of these entities are determined by and essentially represent local culture. The Manifestation and Item aspects cover the physical instantiation of the resource and essentially represent recorded memory. The flexibility of describing an expression embodied in multiple manifestations, and vice-versa, provides efficient accommodation of the diversity of the products of the Fifth information age. RDA implements the resource entity classes directly, without further refinement.

RDA refines the LRM relationships between resource entities and agents to cover the roles of readers in their interaction with books, from the creation of content as author, editor, translator, and performer to the creation of carriers as publisher, distributor, reproducer, and modifier. RDA relationship elements are arranged in hierarchies to support different levels of granularity in metadata. For example, »mixing engineer agent« is the finest relationship in a hierarchy of six levels with »related agent of expression« as the broadest relationship. The hierarchies imply that a relationship recorded at one level is valid for all broader levels; a mixing engineer must also be an audio engineer, a contributor to a recorded performance, a contributor to an amalgamation expression, a creator of an expression, and an agent related to an expression. These implied relationships can be generated automatically to create broader metadata for display and interoperability with external data sources.

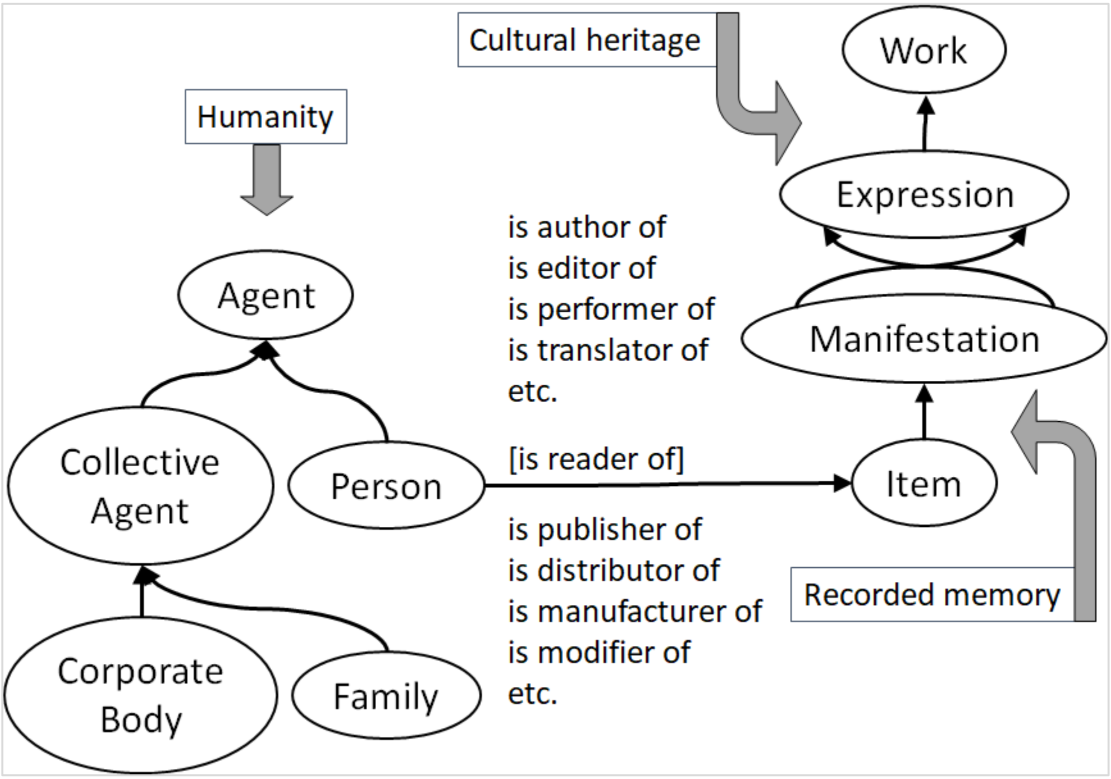

Figure 1: RDA entities for the reader and the book (source: the author)

Figure 1 shows the RDA agent and resource entities that accommodate metadata for the reader and the book, and some of the relationships that describe the interactions between them. The arrows between the agent entities represent super-type or »is a« relationships: a family is a collective agent is an agent. The arrows between the resource entities are primary relationships that ensure the integrity of the description of a whole information resource. The accumulation of resource content as works and expressions is the foundation of cultural heritage and the accumulation of resource carriers as items and manifestations is the recorded memory of humanity.

Note that the simple ur-relationship »is reader of« is out of scope for the LRM and RDA because it is considered to be administrative information that has no relevance to the tasks of information retrieval, but this ignores the impact of »influencers« in social media whose recommendations for sources of information may have a significant impact on the behaviour of information seekers. Metadata on who is reading what are automatically recorded by many Internet search engines and social media sites for business and legal purposes. The addition of such information to the scope of library and cultural heritage metadata may become a future requirement.

RDA also provides refinements of the LRM entity attributes to cover the range of characteristics associated with current information resources. In particular, the RDA content type and RDA carrier type vocabularies for expressions and manifestations respectively cover the whole range of possible kinds of content and carriers. For example, image content is refined in 12 categories such as »still image« and »two-dimensional moving image« and carriers that require access with an audio player are refined in nine categories such as »audio disc« and »audiocassette«. These vocabularies are based on the RDA/ONIX framework for resource categorization (Joint Steering Committee for Revision of AACR, 2006). The Framework identifies fine-grained attributes for the characteristics of information carriers and content that are combined to build a set of basic categories (Dunsire, 2007). The Framework also provides controlled terms for some of the attributes.

All specific things of interest in the bibliographic universe are individuals that are members of one of the LRM classes of entities. The model says that the basic classes of entity are disjoint, so no single specific thing can be a member of more than one class unless the class has a super-class. An individual thing is automatically a member of every super-class in a class hierarchy, in the same way as a relationship hierarchy.

The boundary of a specific individual is a set of characteristics that separate the individual from other individuals in the same class. The characteristics that determine the boundary for each entity within a class are either physical or cultural, but not both. Physical boundaries are measurable and objective in a global environment. Individuals with physical boundaries in RDA include manifestations, persons, places, and timespans. It is worth noting that while RDA states that »The absolute boundary of the [Place] entity is determined by characteristics of the entity that reflect the physical world«, the LRM states that »The entity place, as relevant in a bibliographic context, is a cultural construction; it is the human identification of a geographic area or extent of space«. RDA takes the view that »human identification« is a characteristic of the Nomen entity, and that a physical rather than cultural boundary is better for interoperability with external datasets for places. This does not break the model because any physical boundary can be regarded as a global cultural boundary. The inverse is not valid, so a local cultural boundary cannot be regarded as a global physical boundary. Cultural boundaries are determined locally and are subjective in a global environment. Individuals with cultural boundaries include works, expressions, and families. For example, the LRM states »Bibliographic and cultural conventions play a crucial role in determining the exact boundaries between similar instances of works«.

3 Appellations and identity

An individual entity may be assigned an appellation, a specific label that is used to reference the individual and to distinguish the individual from other individuals. Such a label is a string composed of characters, ideograms, symbols, and other signs that can be recognized by a reader. The signs are typically taken from a specific script, for example a Latin or Cyrillic alphabet, so the utility of an appellation depends on the reader being able to understand the script. Arabic numerals have near-universal recognition, but can be difficult to use when the label is more than a few numbers in length, which is necessary to distinguish between a large number of individual entities. An example is the International Standard Book Number (ISBN) identifier which required an increase from 10 digits to 13 digits in 2007 to cover the increase in the quantity of books being published. Numeric labels are human-readable, but are typically assigned for machine recognition. A specific individual may be assigned a label composed of symbols that provides instant recognition by many readers, for example the unique symbol used by the person otherwise known as »Prince« or the mathematical signs used as titles of music albums by Ed Sheeran, but this is a rare occurrence.

An appellation is categorized in RDA by the kind of agent who assigns it to the individual entity and the context of its use. A name of an agent, place, or timespan is typically assigned by an agent who reflects a local cultural context, while a title of a work, expression, manifestation, or item is usually assigned by an agent who creates the resource. The purpose of the label ranges from recognition and representation of the entity to branding and marketing. The label is assumed to be an uncontrolled and unstructured human-readable string because of the wide range of assigners and contexts. An access point of an entity is assigned by a metadata agency such as an authority control service. The purpose of an access point is unambiguous identification of an individual in a wider national or international context, and to support collocation of individual entities in information retrieval processes. RDA accommodates an access point as a controlled and structured human-readable string. An identifier of an entity is assigned by a metadata agency as a controlled and structured machine-readable string that uniquely identifies an individual.

It is notable that a uniform or internationalized resource identifier (URI or IRI) is not treated as an identifier in RDA. It is not a string in the context of its function, which is to be a unique and permanent direct representation of an individual entity in global linked open data and the Semantic Web. However, an IRI can be »stringified« to create an identifier string for metadata applications that are not based on Resource Description Framework (RDF, the syntax of the Semantic Web).

A label that refers to an individual entity is called a »nomen string« in the LRM and RDA. Nomen is a special entity class that represents a combination of an individual string and the entity it references. Each individual entity combined with a nomen string that is assigned to it is treated as an individual nomen. This allows the appellation to be described separately and to be related to other appellations of the same individual entity. This utility is employed typically in authority control, where one of several appellations of an individual entity is selected for preference in an information retrieval system. An individual nomen is not quite the same as a string on its own that identifies an individual, but the distinction is only significant in linked open data applications. For other applications the RDA Toolkit instructions do not distinguish between an individual nomen and its nomen string.

A nomen string can be related to the agent who assigned the label; for example, an International Standard Serial Number (ISSN) is associated with the ISSN Centre that assigns it. An access point can be related to the method of its construction, known as a »string encoding scheme«, and its language and script. For example, the access point of an individual work may be associated with an encoding scheme that concatenates the name of its creator agent with a title of the work. A nomen string can be related to other nomen strings that are assigned to the same individual entity, such as variant titles and access points. As already noted, it is not necessary to use the Nomen entity when authority control is not required or relevant. A nomen string can be directly related to an individual entity without creating an individual nomen.

RDA provides a set of »appellation« elements that accommodate the distinct categories of name/title, access point, and identifier strings. The categories are aligned with RDA's »unstructured description«, »structured description«, and »identifier« recording methods for element values, and also the »flat file data«, »bibliographic/authority data«, and »relational or object-oriented data« implementation scenarios that are supported by RDA. To complement this, the Nomen entity is aligned with the »IRI« recording method and the »linked open data« implementation scenario. This infrastructure clarifies and supports the complex interplay of humans, nomenclature, and culture in the recording of and access to recorded memory.

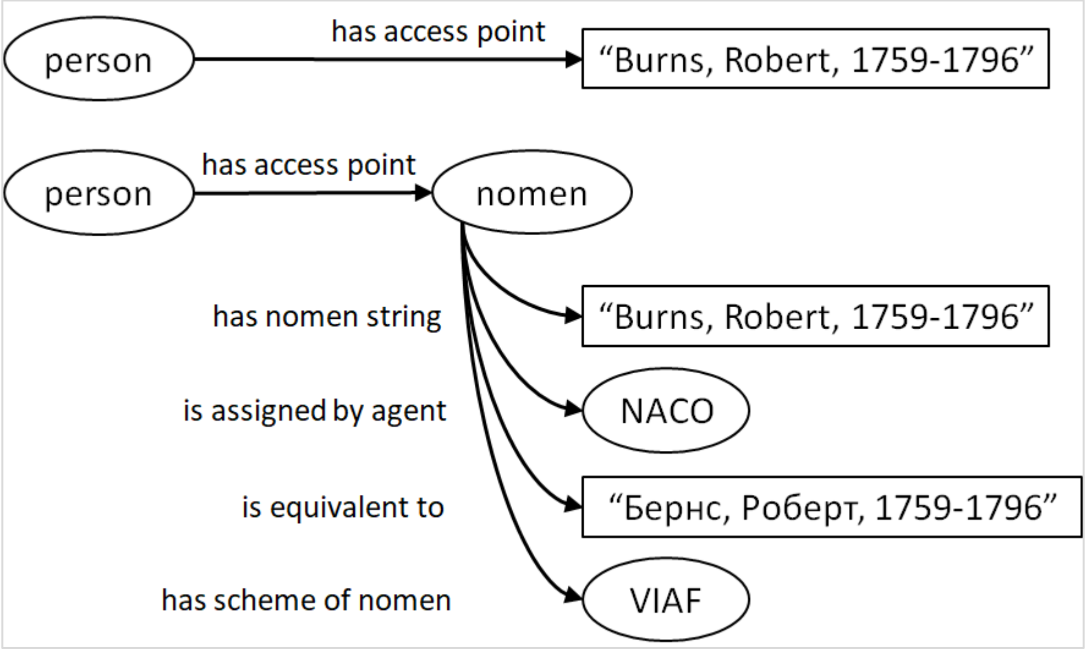

Figure 2: RDA entities and elements for authority headings and authority control (source: the author)

Figure 2 shows the simple and complex data architectures for accommodating a controlled appellation in the form of an access point or authority heading. The top of the figure is an example of an access point used directly as the string appellation of a person in an information retrieval system. The bottom of the figure shows the same access point as a nomen string within an authority control system. NACO is the collective agent »Name Authority Cooperative Program«; VIAF is the vocabulary encoding scheme or dataset »Virtual International Authority File«. Nomen strings and nomen entities can be mixed without losing integrity or semantic consistency. Note that the Cyrillic value of »is equivalent to« is itself a distinct nomen string, so it could be replaced by another instance of a nomen entity, for example if two authority control systems are being merged or interoperated.

An instance of nomen entity is technically a reification (or »thingification«) of a metadata statement that assigns a nomen to its referred entity. The simple statement at the top of Figure 2 is the instance of a nomen in the complex statement at the bottom, which essentially repeats the »person has access point« part of the simple statement. This redundancy allows the simple and complex approaches to be governed by the same set of RDA instructions. While this complication is of interest to system developers, it can remain transparent to the operational cataloguer.

4 Data provenance

The RDA model of data provenance, defined as »Information about the metadata recorded in an element or set of elements« or »metadata about metadata«, extends the process of reification. Reification involves treating metadata as a thing in itself, irrespective of the other things that it describes. A single metadata statement or a set of multiple statements, known as a »description set« in RDA, is reified to create a »metadata work«. This is a kind of Work that can be described like any other work. The expression of a metadata work has the characteristics of the syntax in which the statement is made, such as Resource Description Framework for linked data or UNIMARC for relational databases. This is equivalent to expressions of a work in different languages. The manifestation of a metadata work is the carrier of a metadata statement or description set, such as an online resource, a computer file, or a catalogue card.

This approach allows the full set of RDA attribute and relationship elements to be used in the recording of data provenance at multiple levels of granularity using unstructured or structured descriptions, identifiers, and linked data. The RDA model also has the advantage of re-using existing entities and elements, and does not require any additional entity classes. Some new elements are added to support flexibility in how values for specific data provenance are recorded, such as »note on metadata work« and »recording source«. Some legacy elements are clarified and repurposed, for example »source consulted«. These older elements were the result of earlier attempts to accommodate data provenance in RDA, analysed in the report from the RSC Technical Working Group on RDA models for provenance data (RDA Steering Committee, 2016). That report was published before the LRM, and the current RDA model of data provenance is a development of the analysis for compatibility with the LRM. It is worth noting that the reification approach to provenance is latent in the LRM, for example the »assigned« relationship element between an agent and a nomen (LRM-R14). This is equivalent to a cataloguer who assigns a value to an element of a specific entity to create a metadata statement.

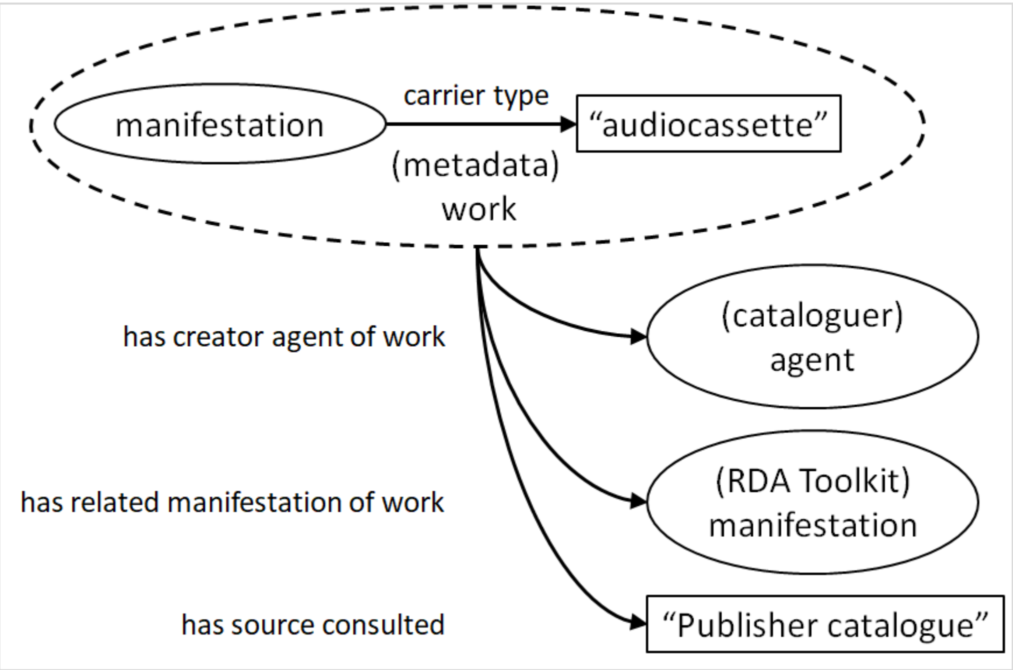

Figure 3: Reification of a metadata statement to accommodate data provenance (source: the author)

Figure 3 gives an example of reifying a metadata statement as an instance of a metadata work that can be assigned data provenance by using Work attributes and relationships. The creator of the statement is a cataloguer or cataloguing agency. The content standard used to create the statement is RDA, embodied in RDA Toolkit and associated with the statement by a broad relationship element. The source of »audiocassette« as the value of the statement is given by a generic title for a publisher's catalogue.

A feature of the 5th information age is a much broader group of agents who generate metadata for information resources. Such interactions include the creation of descriptions of entities in unstructured forms, for example reviews of Amazon products, and in structured metadata, for example descriptions of resources in Wikipedia. Some agents who create metadata can be categorized as trained cataloguers who reflect local cultural norms in the context of national and international bibliographic and cultural heritage standards, and publishers and producers of resources whose aims are to market and promote their products and the agents who create their content. Both of these categories are now vastly outnumbered by the social media crowd with diverse intentions, but the metadata published by the crowd may be of very low quality (Jiang, Liu, Liu, Liu, Chen, and Xu, 2021).

Data provenance is a useful tool for managing metadata from multiple sources. Agents and the metadata they create can be assessed for quality, trust, and consistency, and this information can be used to warn end-users of a dubious source or to filter low-quality metadata from information retrieval systems. RDA provides several methods for reusing and interoperating disparate sources of library and cultural heritage metadata. Data from different sources may vary in the granularity of the entity being described and the granularity of the attribute and relationship elements that are used. The semantic consistency of entity and element hierarchies allows metadata to be automatically broadened, or 'dumbed-up', to a lowest common granularity that allows interoperability. Metadata can be processed in a bulk operation or on-the-fly, while retaining the original metadata for 'smarter' use in the future.

5 Implementation scenarios

RDA supports four scenarios for implementing metadata in an operational information retrieval system. The »flat file« and »bibliographic/authority« scenarios process metadata in packages that describe multiple entities. For example, a bibliographic description usually combines the resource entities into a single record and relates it with authority headings to descriptions of other associated entities. The »relational or object-oriented« and »linked open data« scenarios process metadata in sets that describe a single entity. Metadata for multiple entities is brought together for retrieval and display purposes, but is maintained and stored as separate descriptions; this is entity-based cataloguing.

The semantic integrity of the RDA entities and elements supports the transformation of metadata from one scenario to another. The process is asymmetric: flat file data requires considerable processing to transform it to linked open data, and some data may be lost; linked open data can be flattened automatically with no loss. This implies that the linked open data scenario is the most flexible for operational applications. This is echoed by the LRM, which states »this model is developed very much with semantic web technologies in mind".

The implementation scenarios cover the range of data structures that are used in crowd-sourced metadata. For example, an online review of a movie by a fan may be a flat file description, while an entry for a movie in Wikipedia or IMDb (Internet Movie Database) is more like a bibliographic/authority description. The transformation and granularity of metadata supported by RDA for the scenarios supports the integration of descriptions by the amateur reader with the professional efforts of the trained cataloguer in a cultural heritage organisation.

6 Conclusion

RDA provides a variety of tools for managing bibliographic control in the 5th information age.

RDA implements the full range of LRM entities of interest in the bibliographic universe, and provides a set of elements for each entity class that supports entity-based cataloguing. Elements are arranged in semantic hierarchies to support description at multiple levels of granularity.

The RDA categories of appellation element reflect the contexts in which labels that reference individuals are assigned and used. The Nomen entity and its elements support authority control of access points to reflect local cultural expectations within an international and interoperable infrastructure.

The RDA model for data provenance allows the range of RDA tools to be used to manage multiple sources of metadata in large scale information retrieval applications.

The semantic integrity of RDA entities and elements supports interoperability of metadata that is produced by a wide range of agents for local contexts and scenarios. In the Fifth information age, the reader is the cataloguer, the production of recorded memory is the culture, and the local is the global.

References

Dunsire, G., 2007. Distinguishing content from carrier: the RDA/ONIX framework for resource categorization. *D-Lib Magazine*, 13(1/2). Available at: https://www.dlib.org/dlib/january07/dunsire/01dunsire.html [28. 9. 2023].

Dunsire, G., 2020. *Internationalization of RDA Toolkit during the 3R Project*. Available at: http://www.rda-rsc.org/sites/all/files/247%20Internationalization%20of%20RDA%20Toolkit%20during%20the%203R%20Project.pdf [28. 9. 2023].

Dunsire, G., 2022. Bibliographic control in the fifth information age. In: Bergamin, G. and Guerrini, M. eds. *Bibliographic control in the digital ecosystem*. Rome: Associazione italiana biblioteche. (pp. 25–36). Available at: https://media.fupress.com/files/pdf/24/10612/30806 [28. 9. 2023].

Gerber, H., 2022. *Ten oldest known cave paintings in the world*. Available at: https://www.thearchaeologist.org/blog/ten-oldest-known-cave-paintings-in-the-world [28. 9. 2023].

Jiang, G., Liu, F., Liu, W., Liu, S., Chen, Y., and Xu, D., 2021. Effects of information quality on information adoption on social media review platforms: moderating role of perceived risk. *Data Science and Management*, 1(1), 13–22. Available at: https://doi.org/10.1016/j.dsm.2021.02.004 [28. 9. 2023].

Joint Steering Committee for Revision of AACR, 2006. *RDA/ONIX framework for resource categorization*. Available at: http://www.rda-rsc.org/archivedsite/docs/5chair10.pdf [28. 9. 2023].

Ranganathan, S., 1931. *The five laws of library science*. Madras: Madras Library Association.

RDA Steering Committee, 2016. *RDA models for provenance data*. Available at: http://www.rda-rsc.org/sites/all/files/RSC-TechnicalWG-1.pdf [28. 9. 2023].

RDA Steering Committee, 2023. *Welcome to RDA Toolkit*. (July 2023). Available at: https://access.rdatoolkit.org/ [28. 9. 2023].

Riva, P., Le Bœuf, P. and Žumer, M., 2017. *IFLA library reference model: a conceptual model for bibliographic information.* Den Haag: IFLA. (As amended and corrected through December 2017). Available at: https://repository.ifla.org/bitstream/123456789/40/1/ifla-lrm-august-2017_rev201712.pdf [28. 9. 2023].