
# The Semantic Web and expert metadata: pull apart then bring together

Gordon Dunsire

Presented at Archives, Libraries, Museums 12 (AKM12), Poreč, Croatia, 2008

Published: Glasgow, Scotland : Centre for Digital Library Research, 2009

**Title**

The Semantic Web and expert metadata: pull apart then bring together.

Presented at Archives, Libraries, Museums 12 (AKM12), Poreč, Croatia, 2008

**Author**

Gordon Dunsire

## A problem

Human beings are very good at processing information, in terms of its creation, analysis, synthesis, and communication. This facility is often proposed as a defining characteristic of consciousness, of what makes us human [1]. We are so interested in information that we have used our cognitive abilities to invent machines to process data. But these computers and their networks do it much faster than we can, on a global basis, and without needing to pause for sleep, food, and all the other necessities for sustaining our organic hardware.

The result is the information eruption that is the World Wide Web. The term “information explosion” was created in the early 1960s to describe a rapid increase in published information and its impact on information management and use. If we accept that putting something on the Web is an act of publication, and make no distinction in terms of utility or quality, then the explosion shows no sign of ending, even after 15 years; this is more like a volcano than a meteor. Information professionals trying to create structured, accurate and comprehensive metadata cannot keep up using “traditional” methods. Instead, we need to be clever and seek the solution to our problem in its cause; we need to get our machines to process metadata as effectively as they process data.

## Semantic Web

The Semantic Web is “… an evolving extension of the [WWW] in which the semantics of information and services on the web is defined.” [2] In a computing environment, the use of the term “semantic” is better understood in the concept of “semantic integration ... the process of interrelating information from diverse sources ...” [3]; that is, it is similar to the idea of functional metadata.

The basic building block of the Semantic Web is Resource Description Framework (RDF) [4]. RDF supports the creation of simple metadata statements in the form of subject-predicate-object expressions, called triples. An example of a triple is “This presentation” - “has creator” - “Gordon Dunsire”. Note that this example, simple though it is to human information processors, requires further refinement if it is to be effectively used by machines: “presentation” is a type of information object and “creator” is a type of relationship between an information object and an agent such as a person, organisation, or computer. So we can extend our simple triple to “This (“information object”-“has type”-“presentation”)”-“has (“object/agent relationship”-“has type”-“creator”)”-“Gordon Dunsire”. Note also that “Gordon Dunsire” is not a type of agent, but a label for a specific agent; the statement does not say whether this label applies to a person, organisation, or computer. Similarly, “This” is a relative label for a specific presentation. In RDF, “information object” and “object/agent relationship” are called “classes” and “presentation” and “creator” are called “properties” of those classes, while “This” and “Gordon Dunsire” are called “instances” or “values” (the former implicit, the latter explicit).

RDF is intended to make it easier for machines to process these metadata statements. This requires a machine-processable language for representing RDF statements. Extensible Markup Language (XML) possesses the necessary characteristics, and was a natural choice for RDF. There are other ways of representing RDF, but XML is well-established, familiar to information technologists, and retains a degree of human readability. Strictly speaking, “RDF/XML” is the proper label for the syntax [5], but it is often shortened to just “RDF”. Another requirement is a system of machine-processable identifiers for instances of RDF subjects, predicates, and objects. Labels such as “Gordon Dunsire” are not a good choice for identifiers; they can be ambiguous and subject to change. Instead, RDF prefers the Uniform Resource Identifier (URI) [6]. The better-known URL (Uniform Resource Locator) is a type of URI. URIs are not intended to be understood by humans.

So, for full machine-processing, an RDF triple is a set of three URIs embedded in XML. In RDF, the things requiring identification or URIs are the specific classes, properties, and instances associated with RDF subjects, predicates, and objects. In the example triple given above, the subject “This presentation” has an electronic location given by the URL http://cdlr.strath.ac.uk/pubs/dunsireg/AKM2008.pps. The predicate “has creator” uses a property “creator” already defined in the Dublin Core metadata format with a URI http://purl.org/dc/terms/creator. And the object instance “Gordon Dunsire” has an entry in the Library of Congress Name Authority File which has been made available on the Web by OCLC with the URI http://errol.oclc.org/laf/nb2001-72552.html.

Both the predicate and object instances “creator” and “Gordon Dunsire” have URIs because they are entries in vocabularies which have been made available as “namespaces” in Semantic Web applications. A namespace is a device for providing context to a list of controlled terms, along with term definitions, scope, etc.

Semantic Web namespaces assign a URI to every term, and each URI usually starts with the same URI as the namespace itself (anything on the Web can be given a URI).

There are three Semantic Web applications used extensively for maintaining namespaces. RDF Schema (RDFS) [7] expresses the structure of metadata classes and properties, such as the “information object” class and “presentation” property in our example. Simple Knowledge Organization System (SKOS) [8] expresses the basic structure and content of concept schemes such as thesauri and other types of controlled vocabularies. Web Ontology Language (OWL) [9] explicitly represents the meaning of terms in vocabularies and the relationships between them (scope, etc.)

## Library namespaces

There are several initiatives and projects underway to create namespaces for library vocabularies.

In particular, the DCMI RDA Task Group [10] is developing namespaces for metadata structure and content terminologies from Resource Description and Access (RDA) [11], the successor to the Anglo-American Cataloguing Rules. The RDA metadata element vocabulary is being declared in RDFS, while several sets of controlled terms for the content of specific elements are being made available in SKOS. This will result in the standard labels for metadata elements and attributes, for example “Title” and “Content type”, each having its own URI. The terms which are allowed as values for specified elements, for example “spoken word” (a value for content type) and “microform” (an instance of media type), will also have URIs. This will help various metadata encoding formats, such as MARC21 and Dublin Core (DC), to make machine-processable declarations of which RDA elements and values they use, which in turn will improve interoperability between metadata stored in different encoding formats.

The following is an example of what part of the SKOS version of an RDA value term might look like; it is illustrative only, and it should be assumed that the official version will differ in detail. The line numbers are not part of the RDF/XML file, and have been added for clarification.

```
 1    <?xml version="1.0" encoding="UTF-8"?>
 2    <rdf:RDF
 3     xmlns="http://www.w3.org/2004/02/skos/core#"
 4     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 5     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 6     xmlns:skos="http://www.w3.org/2004/02/skos/core#"
 7     xmlns:dc="http://purl.org/dc/elements/1.1/">
 8    <!-- WARNING: This is a single-concept fragment -->
 9    <!-- Scheme: RDA Content Type -->
10    <skos:ConceptScheme rdf:about="http://RDVocab.info/termList/RDAContentType">
11     <dc:title>RDA Content Type</dc:title>
```

```
12      </skos:ConceptScheme>
13      <!-- Concept: spoken word  -->
14      <skos:Concept rdf:about="http://RDVocab.info/termList/RDAContentType/1001">
15       <skos:inScheme rdf:resource="http://RDVocab.info/termList/RDAContentType"/>
16       <skos:prefLabel>spoken word</skos:prefLabel>
17       <skos:definition>Content expressed through language in an audible form.
18        Includes recorded readings,recitations, speeches, etc., computer-generated
19        speech, etc.</skos:definition>
20      </skos:Concept>
21      </rdf:RDF>
```

*Example: RDF/XML file for the SKOS version of the RDA term "spoken word".*

Line 8 contains a comment explaining that this is only part of the complete vocabulary. SKOS distinguishes "concepts" from "labels" to accommodate synonyms and translations of the "term" being described.

The preferred label "spoken word" is given on line 16, with a definition of the term on lines 17 to 19.

The URI of the term is given on line 14: "http://RDVocab.info/termList/RDAContentType/1001" (note that this is a fictional example, but in reality it will also not be a URL and will not be found by a Web browser).

But this file is itself a structured metadata record, with elements and values; "spoken word" is the value of the element "prefLabel". So these elements themselves must be given in a form that can be processed by machine; they must have their own namespaces and URIs. Of course, namespaces for basic Semantic Web components, including XML, RDF, RDFS, SKOS, and Dublin Core (DC), have already been created. Instead of repeating the full URI for each element from these namespaces used in this example, shortcuts are defined in lines 3 to 7 for each XML namespace (xmlns). That is, line 6 tells the computer that any element in this file which starts with "skos" is taken from a namespace where the URI for all elements starts with "http://www.w3.org/2004/02/skos/core#"; the URI for "skos:prefLabel" is therefore "http://www.w3.org/2004/02/skos/core#prefLabel". (The machine does not need the shorthand – it just makes the file more human-readable, by programmers.)

When the official RDA namespace is finalised, any other RDF/XML file containing bibliographic metadata can define, say, "rdact" as the shorthand for the RDA Content Type namespace and then use it to refer to specific values from the namespace. Using the fictional example, "rdact:1001" would be a shorthand URI identifying the value "spoken word".

The International Federation of Library Associations and Organisations (IFLA) is also actively developing namespaces for some of its standard vocabularies for bibliographic control. The FRBR Namespace Project [12] seeks to define

appropriate namespaces for Functional Requirements for Bibliographic Records (FRBR) [13] in RDF and other appropriate syntaxes. This involves creating RDFS representations of FRBR entities, for example “Expression”, and relationships, for example “is expression of” and its reciprocal “is expressed by”. This initiative is related to the RDA work because RDA is based on the FRBR model. Discussions during the World Library and Information Congress 2008 in Québec City, Canada, indicated a significant interest in making other IFLA vocabularies available to the Semantic Web; these may, in the future, include appropriate parts of Functional Requirements for Authority Data (FRAD) [14], International Standard Bibliographic Description (ISBD) [15], Functional Requirements for Subject Authority Records (FRSAR) [16], and UNIMARC [17].

At the same time, the Library of Congress is creating namespaces for Library of Congress Subject Headings (LCSH) and Library of Congress Name Authority File (LCNAF) in SKOS, and for MARC21 and associated metadata structure format in RDFS and OWL [18].

Collectively, these activities include the metadata structure and content vocabularies which are most widely-used in the international library domain. The potential impact of making these compatible with the Semantic Web is very high. Information retrieval service developers will have access to ontologies and terminologies which have been developed with international collaboration over many years. Archives, libraries and museums will see radical changes to their conceptualisations of metadata and their management, as consideration of the development of the library bibliographic record will demonstrate.

## Evolution of the bibliographic record

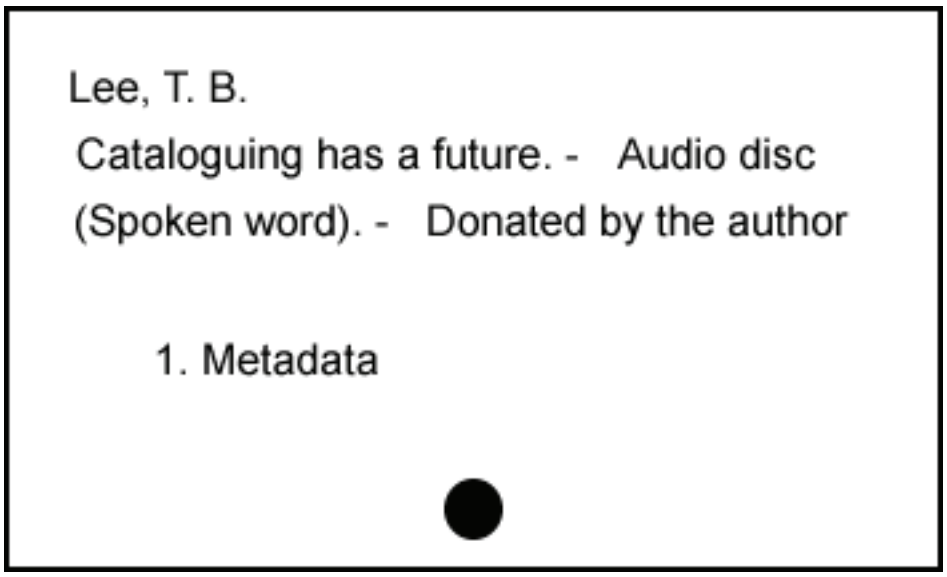


*Figure 1: A simple catalogue card.*

In the beginning was the catalogue card. It is a purely physical carrier for a metadata record, and cannot be processed by machine after it is created (although a computer may be used to produce it). It is intended only for processing by humans. Indeed, the hole in the bottom is an aid to such processing, by allowing a set of cards to be held in place by a rod to prevent them spilling out of order when consulted by the user. The metadata structure is entirely implicit. Fields are not labelled, and are delimited by standard punctuation. The semantic type of each field is implied by its position on the card, its contents, and the context of the card. In Figure 1, “Lee, T.B.” is assumed to be the author because it is the first piece of metadata and the card is filed in the “author index”. Similarly, “Audio disc” is not assumed to be part of the title because it is preceded by something that looks more like a real title, and the punctuation indicates that it is a separate piece of metadata. There are many circumstances where this approach results in ambiguous metadata, for example the titling of the 1986 recording by the group Public Image Ltd is “Album”, “Compact disc” and “Cassette” depending on the format [19].

*Flat-file record*

| | |
|---|---|
| *Author* | Lee, T. B. |
| *Title* | Cataloguing has a future |
| *Content type* | Spoken word |
| *Carrier type* | Audio disc |
| *Subject* | Metadata |
| *Provenance* | Donated by the author |

*Figure 2: A flat-file record.*

The first stage in developing the catalogue record for machine-processing is to label or otherwise identify the different types of field in the metadata structure. In Figure 2, the field labels are English words or phrases, which makes them easy to identify by English-speaking humans, but for machines the only requirement is that the label is unique for each field. For example, the MARC21 metadata format uses three-digit labels; its “245” label is similar to “Title” in Figure 2. The resulting “flat” record is easier to manipulate by computer; a list of titles is generated by listing the contents of the field “Title” from all the records. Furthermore, the program will not break down if it encounters “Title” as part of the contents of a specific record; that is, when the resource described by the record has the title “Title”. It is easy to process the flat-file record to display it like the catalogue card in Figure 1 or the columnar

layout of Figure 2 which is now prevalent in the library online public-access catalogue (OPAC) and in directories and listings in Web-based services.

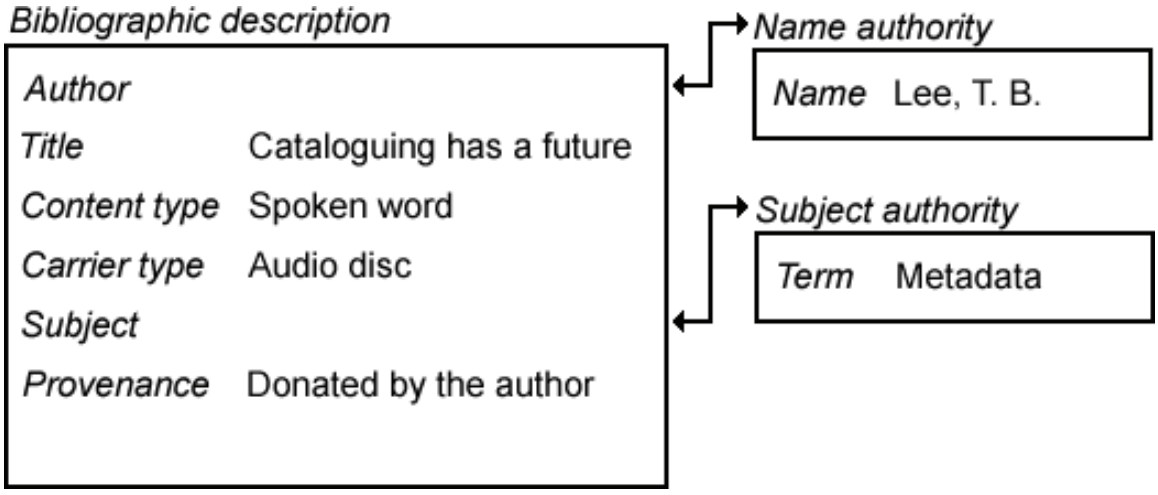


*Figure 3: A typical modern OPAC record, with description and authority heading components.*

Behind the OPAC display, however, there is typically more than one flat-file component record. Relational database management techniques make machine-processable data more efficient to store and maintain by reducing duplication of content. Instead of repeating the same content inside multiple records, a single copy of it is stored in a separate record and then linked to all the relevant main records. In addition to saving data storage space, only the single record requires maintenance and any update is immediately reflected in main record displays. The separate record can also be used to store other data related to the de-duplicated content. The modern OPAC system exploits these techniques by using authority files for the names of personal and corporate names, and subjects. Both types of content are relevant to multiple main bibliographic records; both utilise controlled vocabularies and normalisation requiring separate maintenance; both include additional fields for defining and scoping the vocabulary terms. Figure 3 shows how authority file content is linked to a bibliographic description record. (Note that the terminology “heading” reflects the use of normalised names and subject terms to determine the sorting of catalogue cards in author and subject sets, as shown in Figure 1 where the author name heads the metadata record.) The links are made using the machine-processable identifiers of records in the authority files; these may be local to the system, or derived from external sources. The components of the OPAC record do not have to be maintained locally, and many libraries import authority metadata records from other organisations, for example LCNAF and LCSH. Non-local identifiers improve interoperability, particularly if they are in widespread use.

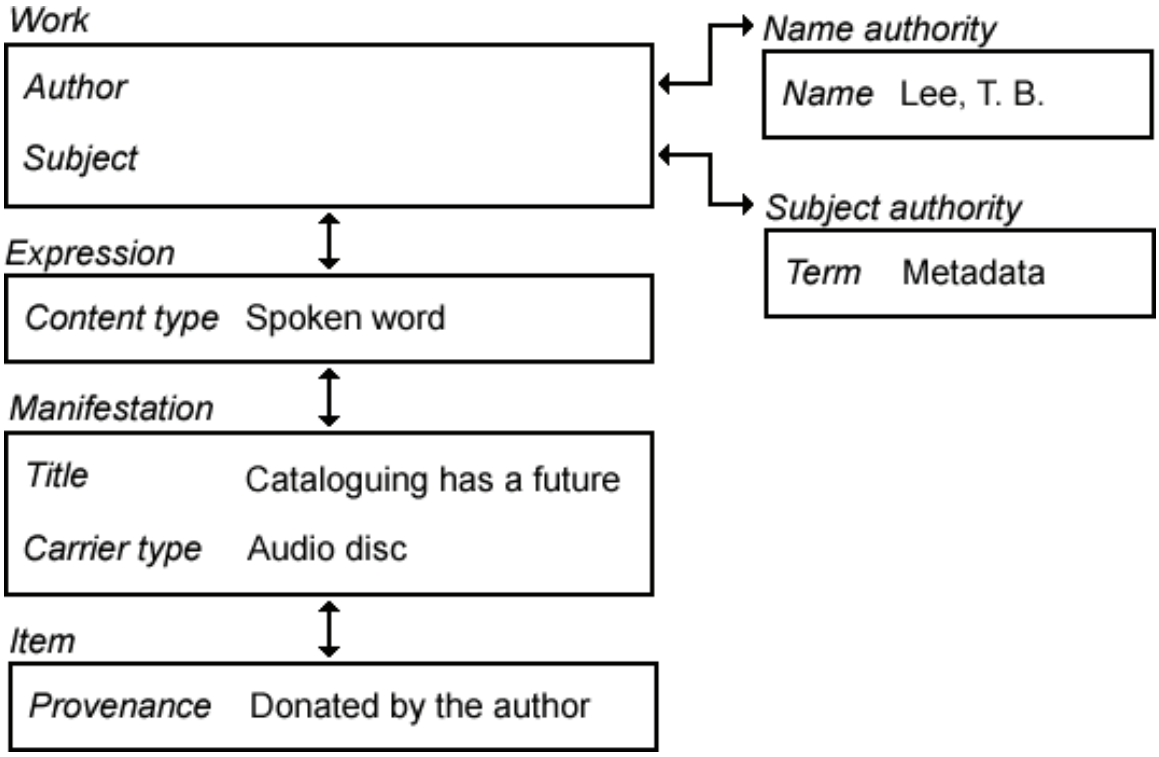


*Figure 4: A FRBR-ised record, with disaggregated description and authority heading components.*

Application of the data model of FRBR results in further disaggregation of the record. The model assigns metadata fields (attributes) to four groups or entities: work, expression, manifestation, and item. “Work” contains fields relevant to the abstract, intellectual components of a resource; “Expression” relates to the content of the resource; “Manifestation” relates to the carrier of the content; “Item” refers to specific copies of the resource. A work can have more than one expression; an expression more than one manifestation; a manifestation more than one item. This structure has similar qualities to the disaggregated bibliographic description and authority file model of the typical OPAC record, where the same sub-set of metadata can appear in multiple main catalogue “records”. The advantages of treating work, expression, manifestation and item metadata as separate records are the same: avoidance of duplication; more efficient and effective maintenance; and integration with external sources. It is reasonable, then, to predict that FRBR-isation of library catalogues will result in bibliographic descriptions being split into three or four discrete records, as shown in Figure 4. (Item metadata are already treated separately in most modern systems because of their interaction with circulation control.)

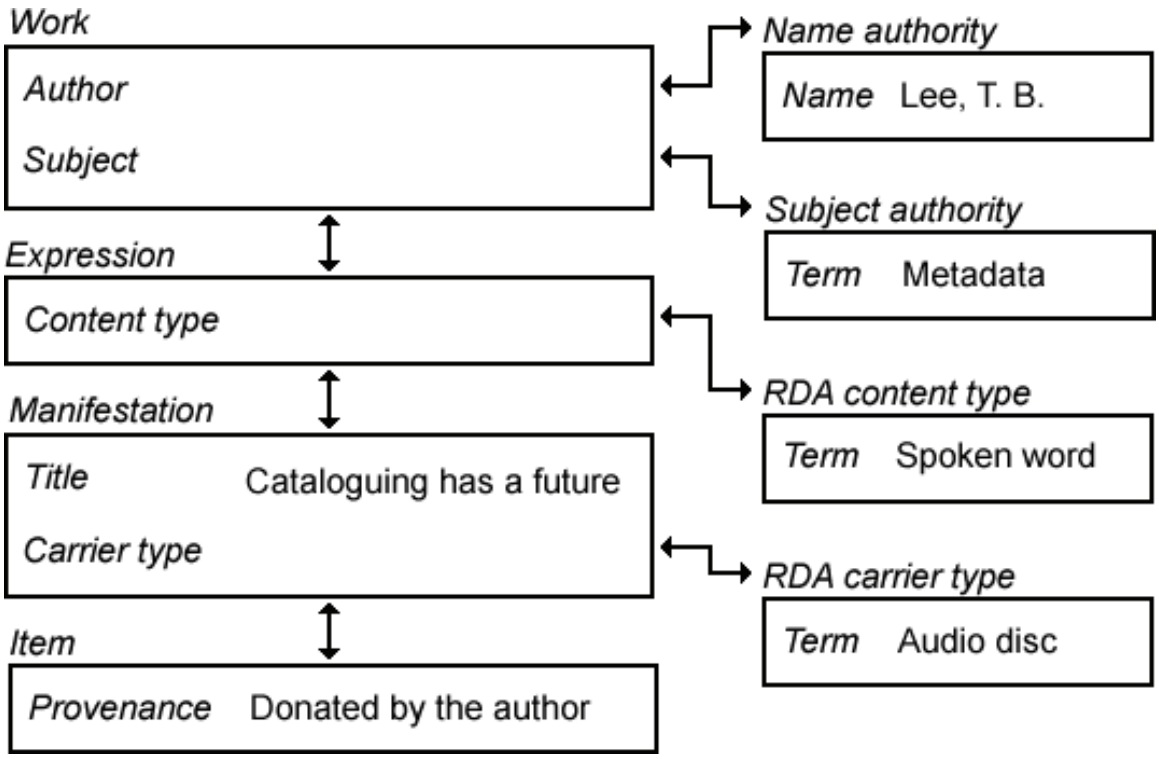


*Figure 5: A FRBR-ised record with RDA vocabulary components.*

The availability of RDA vocabularies will provide further impetus to disaggregation, as they are to all intents and purposes authority files. The controlled vocabulary content of the “Content type” and “Carrier type” fields in the example record can be replaced by URIs, as shown in Figure 5. An additional advantage of doing this is to improve interoperability between metadata in different languages. The Croatian equivalent of “Spoken word” will have the same URI if it is included in the RDA SKOS file, or have a URI which is linked in a machine-processable way using OWL to the URI of the English-language term.

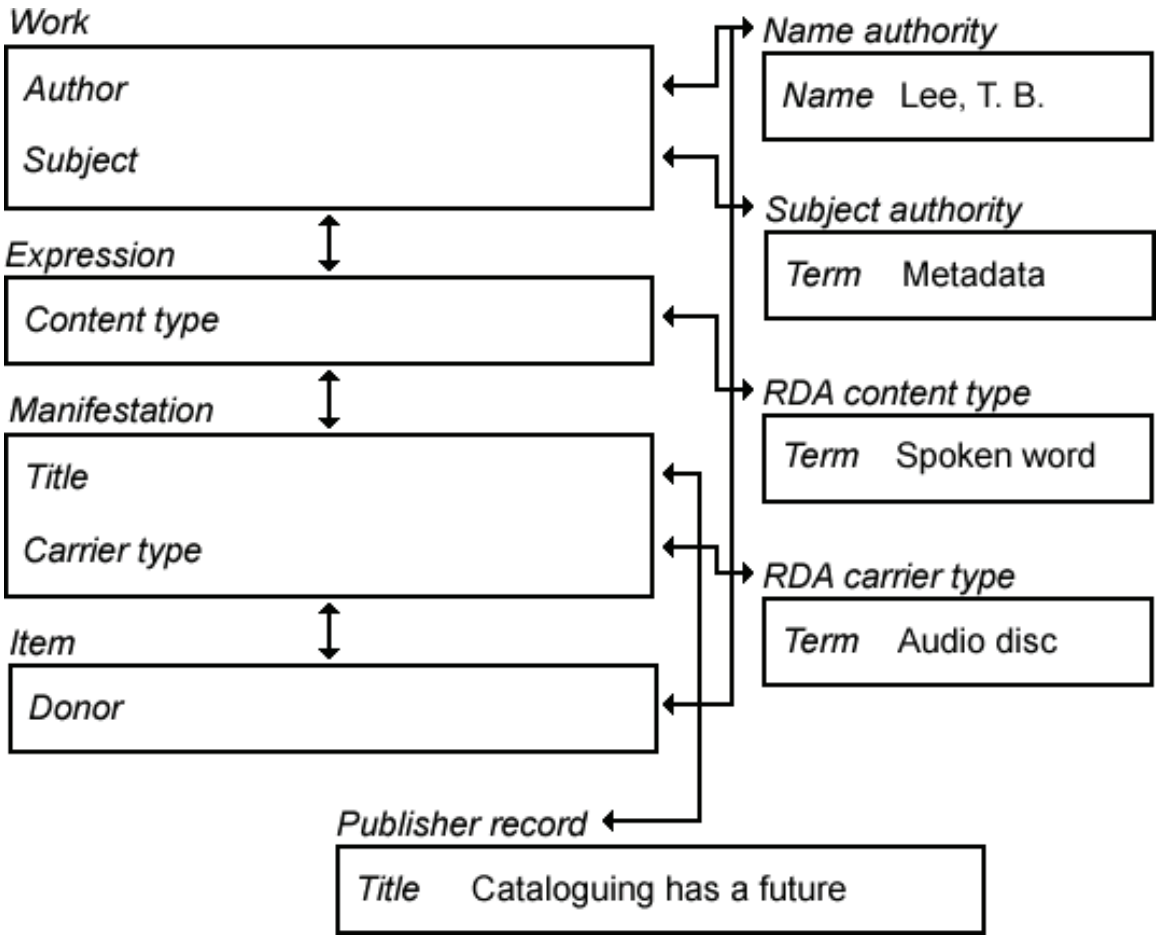


*Figure 6: A completely disaggregated record based on Semantic Web components.*

It is possible to restructure the Item metadata in the example in order to take advantage of the name authority file to avoid duplication, and improve machine-processing. The provenance note becomes an explicit donor field which is then linked to same value of name as the author field, as shown in Figure 6. Finally, publisher metadata containing authoritative title information can be linked.

So the original catalogue card, with explicit local content and implicit structure, has evolved into a multi-record aggregation with explicit structure and distributed global content shared amongst many such “records”. If this is a truly different species, then the traditional library record based on the catalogue card has become extinct.

## Implications for common information environments

The Semantic Web will be a web of metadata, broken-down into simple statements which can be re-aggregated in many different combinations. If all archive, library and museum metadata are processed in this way, the different domains can take advantage of each other’s expertise and output. There will be no metadata records,

only one metadata record covering everything, or a near-infinite number of different metadata records, depending on the point-of-view of the metadata user. The Semantic Web will allow machines to create a metadata record for a particular resource just-in-time and on-the-fly, rather than have static records stored just-in-case. The benefits of metadata creation and maintenance by information professionals will be available to all.

The user will have control over the presentation and detail of metadata. Recombination from the basic building blocks of the RDF triples will allow information retrieval interfaces to display a record in formats familiar to users of archives, libraries or museums (and users of Amazon, Google and Flickr), as well as innovative layouts.

And by avoiding duplication, cataloguers and other metadata creators can devote their efforts to describing new stuff, with considerable assistance from the computer.

## References


[1] Pasternak, Charles (editor). What makes us human? Oneworld Publications, 2007.

[2] Wikipedia. English [edition], 19.50 30 Aug 2008. Available at: http://en.wikipedia.org/wiki/Semantic_web

[3] Wikipedia. English [edition], 10.16 19 Mar 2008. Available at: http://en.wikipedia.org/wiki/Semantic_integration

[4] Resource description framework (RDF). 2004. Available at: http://www.w3.org/RDF/

[5] RDF/XML syntax specification (revised). 2004. Available at: http://www.w3.org/TR/rdf-syntax-grammar/

[6] Joint W3C/IETF URI Planning Interest Group. Uniform resource identifiers (URIs), URLs, and Uniform resource names (URNs): clarifications and recommendations. 2002. Available at: http://www.rfc-editor.org/rfc/rfc3305.txt

[7] RDF vocabulary description language 1.0: RDF schema. 2004. Available at: http://www.w3.org/TR/rdf-schema/

[8] SKOS Simple knowledge organization system - home page. Available at: http://www.w3.org/2004/02/skos/

[9] OWL Web ontology language: overview. 2004. Available at: http://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.1

[10] DCMI/RDA Task Group wiki. Available at: http://dublincore.org/dcmirdataskgroup/

[11] RDA: Resource description and access. Available at: http://www.collectionscanada.gc.ca/jsc/rda.html

[12] FRBR Review Group. Available at: http://www.ifla.org/VII/s13/wgfrbr/

[13] Functional requirements for bibliographic records. Current text. 2009. Available at: http://www.ifla.org/VII/s13/frbr/

[14] IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). Available at: http://www.ifla.org/VII/d4/wg-franar.htm

[15] International Standard Bibliographic Description (ISBD). Available at: http://www.ifla.org/VII/s13/pubs/cat-isbd.htm

[16] IFLA Working Group Functional Requirements for Subject Authority Records

(FRSAR). Available at: http://www.ifla.org/VII/s29/wgfrsar.htm

[17] UNIMARC manual : bibliographic format 1994. Available at: http://www.ifla.org/VI/3/p1996-1/sec-uni.htm

[18] Marcum, Deana B. Response to On the record: report of the Library of Congress Working Group on the Future of Bibliographic Control. Available at: http://www.loc.gov/bibliographic-future/news/LCWGResponse-Marcum-Final-061008.pdf

[19] Wikipedia, English [edition], 21 Mar 2008. Available at: http://en.wikipedia.org/wiki/PiL#Album.2FCompact_Disc.2FCassette