Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience

Carl Lagoze
Cornell University
lagoze@cs.cornell.edu

Presented at Seminar on Metadata
Organized by Archiefschool, Netherlands Institute for Archival Education and Research
June 8, 2000

Abstract: The Dublin Core Metadata Element Set (DCMES) grew out of a recognized need for improved resource discovery of web resources. Initial work on the DCMES focused on the requirement of simplicity: "ordinary" users should be able to formulate descriptive records based on a relatively simple schema[1] (fifteen free-text elements). Over the years there has been a movement within the Dublin Core community to use the DCMES for more complex and specialized resource description tasks and, correspondingly, develop mechanisms for incorporating such complexity within the basic element set. This work has generally been called *qualified Dublin Core*. We examine the notion of accommodating complexity in a simple metadata model and argue that the dual requirements are incompatible. We discuss the role of events and processes in more expressive metadata and how simple resource-centric models, such as DCMES, are not equipped to express these semantics.

1 Realities for all occasions

Reality is chaotic. It consists of entities and objects of all types and forms. These entities change over time and sometimes morph into other distinct objects. As a result entities are interrelated in numerous and complex ways. Just limiting our domain to the document world we see relationships such as translations, derivations, editions, versions, and citations, just to name a few.

People try to understand and work with this chaotic reality by simplifying it. Using categorization and classification they create artificial ordered realities in which entities fit into

---

[1] In a rapidly developing and wide-ranging field such as metadata, finding the right common terms is a significant part of the problem. Throughout the remainder of this paper, we use a number of terms in relation to metadata:

1. *Vocabulary* – The set of elements (properties) provided by a specific metadata set.
2. *Statement* – The result of associating a metadata element with a resource and value (e.g., "Romeo and Juliet has a creator William Shakespeare").
3. *Record* – A set of statements that collectively describe a resource.
4. *Schema* – The rules, or data model, for constructing statements in a metadata set.
5. *Metadata Set* – A "standard" for metadata that includes both a vocabulary and schema.

As described in this paper, the DCMES defines a limited vocabulary, the fifteen elements, and a simple resource-centric schema.

convenient slots. As noted by Bowker and Star [9], humans are insatiable classifiers who deeply fixate classification schemes into social, political, and scholarly structures. This categorization allows people to ignore the idiosyncrasies of individual entities and manipulate them via their coarse granularity group characteristics.

The modern library, and the Catalog that is at the core of its operations, is arguably the preeminent example of classification. From Melvil Dewey's conception of the Dewey Decimal Classification system [12] in the 19th century to the now preeminent MARC encoding [14] of AACR2 cataloging rules [15], libraries have been deeply engaged in classifying a variety of physical (and now digital) artifacts. In his excellent study of catalogers and their craft, David Levy described the process of "order making" [20], whereby a veneer of regularity is overlaid on the natural disorder of the artifacts that libraries encounter.

The emergence of the Web and the explosive growth of on-line content has challenged and enriched this order-making task. Interest in *metadata* has evolved in response to this new and challenging context. While it shares many of the same purposes as cataloging, ordering and therefore simplifying the entities that it describes, metadata plays a somewhat different role due to the way the Web differs from the traditional library environment.

The environment within which the Catalog and its standards exist is relatively self-contained and controlled. Creation and maintenance of catalog records is the task of a controlled community of expertise. The interface to Catalog records is generally restricted to specialized Integrated Library Systems (ILS) with little or no published interface to other systems. Finally, exchange of catalog records is regimented, usually in the form of MARC downloads from authorized sources like OCLC or RLG.

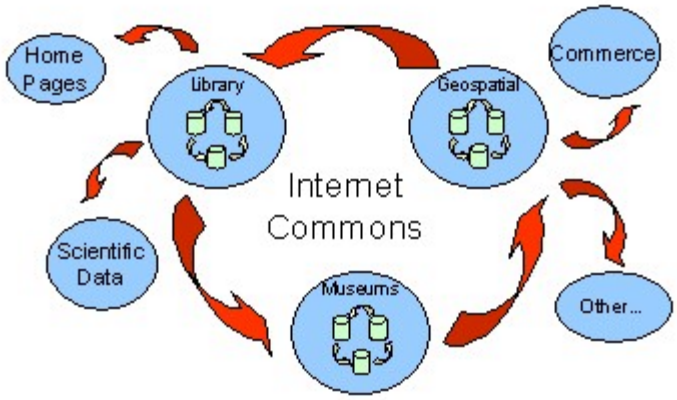

Figure 1 - Mixing information from multiple communities

In contrast to this controlled community, the Web is a bit like the Wild West. The maintenance of information is the purview of diverse communities with a variety of descriptive standards. The content and services provided by these diverse communities coexist in the same common space and their use frequently crosses community boundaries. Stuart Weibel, who has led the Dublin Core Metadata Initiative since the beginning, has characterized this as the *Internet Commons*, where boundaries of control and use of information are blurred or non-existent. This concept is illustrated in Figure 1[2], where the circles indicate domains – the divisions among which are

[2] This figure is adapted from one originally presented by Stuart Weibel in a number of Dublin Core talks.

themselves frequently not distinct – and the arrows represent the exchange of information amongst these domain boundaries.

This environment presents both a challenge and opportunity for the formulators of descriptive metadata standards. The boundary-crossing nature of Web use creates the need for descriptive standards that facilitate usability across domain and community boundaries. This requirement is often described under the rubric of *interoperability.*

While the need for interoperability in the Internet Commons is important, the fabulous diversity of the Web also provides a unique opportunity for customization and specialization. As we noted in an earlier paper [16], metadata on the Web should allow individuals to use and search the global information space conformant to their current roles and needs. We can think of metadata like a database *view*, capable of projecting multiple order-making schemes on content. This then makes it possible to customize the services that consume that metadata, for example search engines, and tailor their functionality to differing needs. This concept is illustrated in Figure 2 where the same content, the painting of the Mona Lisa, may be projected via metadata in three views:

1. *geo-spatial* – that describes the specific location of the object and routes to use to find it. Such metadata might be useful for applications such as museum directories or tours on mobile devices[3].
2. *rights* – that emphasizes the identity of agents and organizations involved in ownership or management of the object. Such metadata might be useful in the production of copies and derivations of the original work.
3. *museum* – that emphasizes facets of the object associated with its exhibition and preservation.

Such multiple descriptive views are possible by using a *modular* approach, where separate metadata *packages* are associated with the resource. This is the approach taken by a number of Web metadata architectures including the Warwick Framework and Resource Description Framework [4, 10, 18, 19], which permit multiple communities of expertise to associate need and domain specific metadata with Web content.

Such modularization has typically been described and justified from the perspective of domain specificity. However, the advantages afforded by such an architectural framework are amenable to other dimensions of specialization. This paper focuses on the simplicity/complexity dimension[4]. This dimension plays an important role in resource discovery because of the manner in which users generally begin the discovery process using basic terms and then "drill-down" with higher levels of specificity in their queries. Simple descriptions, those that have been embodied in a number of "core" metadata sets, are extremely important components in such a strategy. Similarly, complex descriptions play an important role of permitting users to query more granular, often domain-specific, aspects of resources, or by providing information vital to the preservation, administration, and management of access to the resource. We argue that attempting to intermix in a single descriptive schema and vocabulary the simple semantics needed for coarse granularity queries with the complex semantics needed for drill-down queries, and

---

[3] John Perkins of CIMI has described to me a number of interesting applications of mobile devices in the museum environment.

[4] Complexity, as covered by this paper, is in the form of richness of description. We recognize that this is but one facet of complexity. Others, which play an important role in metadata and which also deserve examination, include versioning, multiple languages, and multiple encodings.

other purposes, leads to metadata sets that are not ideally suited for either purpose. As an alternative, the modularization principles embodied in RDF and the Warwick Framework should be exploited to develop and deploy schema tailored for simplicity and others tailored for complexity.

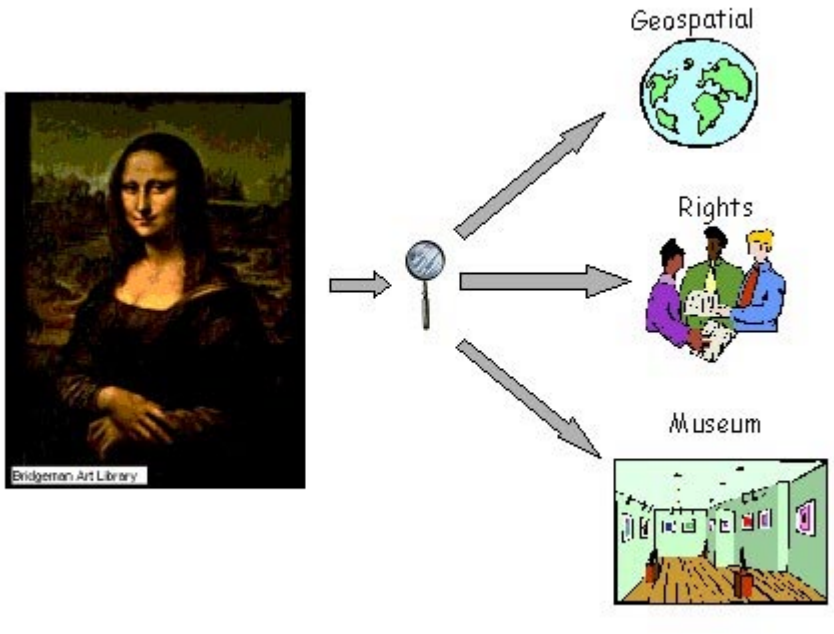

Figure 2 - Multiple views of the same content

The paper will use the Dublin Core Metadata Element Set (DCMES) and its development history to illustrate these points. The purpose here is not to disparage the DCMES, which has proven enormously successful for its original purpose as a descriptive metadata set for coarse granularity resource description. We do mean to raise issue with the efforts in the Dublin Core Metadata Initiative (DCMI) over the past several years to re-purpose the DCMES as the basis for richer descriptions. In our opinion, and that of many others in the metadata community with whom we have discussed this issue, such an effort has interfered with the original goal and failed to provide a basis for such rich descriptions. We encourage the DCMI, and other communities involved in metadata developments, to turn to modularity when faced with a variety of descriptive requirements[5].

2 A world of document-like objects

The history of the Dublin Core has been well documented in a number of workshop reports[13, 22-25]. The purpose of this paper is not to replay this history. What follows is a brief review of the development of the DCMES to provide a context for the remainder of this paper..

---

[5] Throughout this paper we will clearly distinguish between two things: 1) the DCMES that is the 15-element set, and 2) the DCMI that is the organization that is examining metadata for networked resources and has the DCMES as its most visible result. Other tangible results of the DCMI are the Warwick Framework and much of the work on using RDF for descriptive metadata. This distinction corresponds to recent work of the DCMI that has involved refining and broadening scope.

The DCMI began in 1994 in response to the recognized need for better resource discovery tools for the Web. This requirement grew out of dissatisfaction with two extremes:

- The standard cataloging methods in libraries were, and still are, too complex and expensive to provide a reasonable basis for resource description of Web content. Whereas such methods may be appropriate for stable entities such as the physical artifacts that libraries collect, they are inappropriate for the dynamic Web environment. Web content is ephemeral and disseminated by a variety of sources that are often far removed from established publication authorities.
- The simple "one text box" approach used by existing web search engines, while useful, does not permit even the simplest type of search specificity. There are times that users find it desirable to be more specific in their searches. For example, even the simplest queries frequently need to distinguish between *by-ness* (e.g., books by Charles Dickens) and *about-ness* (e.g., books about Charles Dickens).

The product of the early Dublin Core meetings – one that has remained essentially stable and is recognized as the primary result of the DCMI – is the fifteen-element DCMES. These elements include some that are reasonably consistent across all domains – for example, creation of the resource, naming of the resources, subject of the resource – and others that some argue stand on the fringe of "core-ness" – such as, geospatial characteristics and rights management statements. Focusing on the exact composition of the Dublin Core elements is not the purpose of this paper. In fact, any argument about the exact composition of core semantics is rather moot since each community evidently has individual notions of such.

The more relevant task is to use the *view* metaphor mentioned earlier and thereby understand the nature of the "ordering" that DCMES is imposes on content. This understanding can be inferred by looking at the types of resources that DCMES was targeted at; simple Web documents written in HTML. Much of the early literature on the DCMES characterized this type of object as a *document-like object*, or DLO.

The exact nature of a DLO has never been specified and, in fact, lack of specificity about its definition is central to its nature. Drawing from its humble origins, the simple Web page, we argue that the essence of a DLO is simplicity in both structure and lifecycle. That is, a DLO is not composed of compound sub-parts nor is it characterized by complex inter-relationships with other resources, either physical or digital.

This simplicity may not actually correspond to any resource. An analysis of many Web pages, even those from the earliest days of the Web, shows that very few of them are stand-alone items, and most have subtle and unexpected complexity. Exact correspondence of the DLO view to reality, however, is neither an important issue nor is it relevant to our argument. We take the perspective articulated by Borges who noted that "…there is no classification of the universe that is not fictional and conjectural." [8]

The relevance of the DLO is its usefulness as a simple metaphor for describing a mixture of resources and facilitating cross-domain, cross-genre resource discovery. By "pretending" that a cross-section of resources is uniformly simple we thereby make it possible to search for them in a simple manner. Two useful metaphors have surfaced during the development of the DCMES to express this simplicity:

1. *Pidgin Languages* – Tom Baker, who is a member of the Dublin Core Executive Committee, has compared the DCMES to a pidgin language. Such a language ensues when individuals with different language backgrounds are mixed together (often forcibly as in refugee communities or during the time of American slavery). Inevitably, the

members of such communities rapidly develop a crude, syntax-less language for basic communication amongst themselves.

2. *Digital Tourist* – Ricky Erway of Research Libraries Group (RLG) used the digital tourist metaphor for the DCMES. A tourist who travels to a country with another language brings a basic phrase book and develops a set of rudimentary phrases for communication during the visit. An examination of phrase books reveals that the elements of the basic language are quite uniform despite the target language differences (e.g., "I need a doctor", "Where is the train station", etc.).

Key to the simplicity of both of these simple languages is both limited vocabulary and simple structure. This means that the *statements* that can be constructed lack compound syntactic constructions (e.g, sub-phrases and complex clauses) and are mainly stated in the present tense. Events and entities are flattened into simple declarative phrases. Such flattening is central to the simplicity of the DCMES as originally conceived. Basic DCMES descriptions, which are a combination of very simple statements about single resources, project views of the resources that generally hide or obscure the complexity of their origins and derivations.

DC Element Set Record

Title	Mona Lisa in Curlers
Creator	Leonardo da Vinci
Creator	George Castaldo
Date	1506
Date	1994

Figure 3 - Flattening complex reality

We can examine this through an illustrative example shown in Figure 3. The example shows an image called "Mona Lisa in Curlers"[6]. This image, created by George Castaldo in 1994, is clearly a derivation of the famous "Mona Lisa", created by Leonardo da Vinci in 1506. As described later, such derivations have complex histories that consist of a variety of agents, tools, and events. Shown in Figure 3 beside the image is, in a syntax free representation, one possible simple DCMES record for this image[7]. The record demonstrates the nature of flattening, whereby the creators *da Vinci* and *Castaldo* are placed at the same descriptive level, as are the two dates of

---

[6] Mona Lisa in Curlers is © 1994 the American Postcard Co., Inc. as part of the *Misguided Masterpieces* Series. This image was extracted from the Web page at http://www.pipeline.com/~rabaron/MONA08.htm.

[7] The flattening of what are essentially a number of creations into one record as shown in this example is not mandated by the DCMES. In fact, there has been considerable discussion in the DC community about the so-called *one-to-one rule* (due to David Bearman). Briefly stated, this "rule" expresses the notion that creators of DCMES records should recognize the dangers of too much flattening and, where appropriate, should create separate records for what are essentially separate items. Therefore, the single record in this example could be expressed as separate records with linkages between them using the *relation* element. Our view is that either form, the single record or multiple linked records, is not necessarily "correct" and one of the strengths of DCMES is the ability choose the compositions of the records based on what is most deemed most useful for resource discovery. As shown in the remainder of the paragraph, the single record as shown has definite use as a tool for discovery.

"creation". Without a doubt, the description as shown does not stand alone as a complete description of the artifact. On the other hand, the simple document-like view provides data for indexing that is useful for simple resource discovery. For example, users making a "digital tourist" query for resources by "da Vinci" would find the resource. Of course, such simple metadata does not permit queries about the type of software used for digitally processing the image, but such information falls outside the notion of pidgin and ventures into the more complex descriptive languages inhabited by domain experts.

The next section of this paper explores further aspects of this example and describes problems that result from trying to extend the flat model in an open fashion.

3 Confounding the simple model

Agreement on simplicity and the nature of the DLO has certainly not been universal among the parties involved in the DCMI. Indeed, there has been substantial interest from the beginning in schemes to enrich the descriptive power of the DCMES and use it for purposes generally outside the application of cross-domain resource discovery. Rather than think of the DCMES as a simple view of richer descriptions, some have sought to use the DCMES as a mechanism for creating rich cataloging records[8].

The early discussions about this were described as a division of the DC community into the *minimalists* and the *structuralists*. The former advocated that the most valid use of the DCMES was in their simple, unadorned free-text form; the latter were interested in methods of enriching the capabilities of the DCMES through various mechanisms. The arguments by the latter group centered on the fact that the unadorned elements were simply insufficient for "real description" of any resources. We do not disagree with this argument, but maintain that "real description" cannot be done in a generic manner (it is context-specific) and that neither the schema nor vocabulary of the DCMES is sufficient for such description.

Over time, the minimalist/structuralist discussion evolved into the characteristics of *qualification*. Broadly speaking, qualification consists of mechanisms for adding semantic specificity to DCMES descriptions. While the basic fifteen elements have remained almost invariant over the past five years of DCMI, the issue of qualification has been a sea of shifting interpretations and models. Much of the difficulty has been devising a means of accommodating complexity and extensibility with the simplicity of the original DLO view. In theory, individual communities should be able to establish qualifiers to elements, tailored to their domain-specific needs. Furthermore, the DC records enhanced with these qualifiers should be able to interoperate with records containing qualifiers devised by other communities and with DC records that employ just the simple unqualified semantics. The key principal for accomplishing this interoperability is the notion that element qualifiers should *refine* rather than *extend* element semantics[9].

---

[8] As a corollary, there has been a general presumption that DC elements are fixated in physical records. As we note in [15], mechanisms whereby DCMES elements are computationally projected from more complex descriptions stored in databases may be the more sensible and scalable approach.

[9] We are ignoring here another form of qualification informally known as *value qualification*. This form allows the specification of controlled vocabularies or encoding rules for element values. An example of such is the association of the encoding rule "LCSH" to a DC *subject* value to indicate that the value term is described in the Library of Congress Subject Headings. In fact, recent work by Tom Baker using sentence construction metaphors indicates that the distinction between the two "types" of qualification may indeed be a red herring.

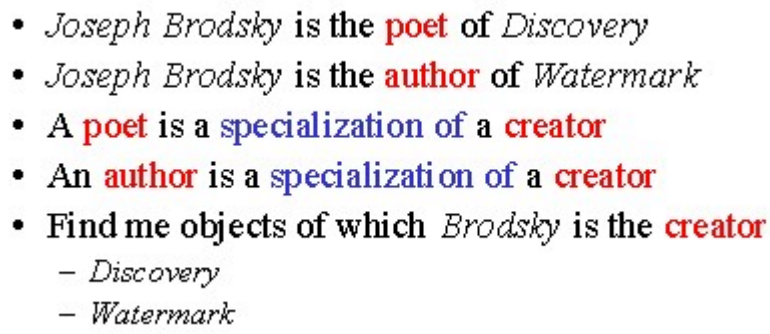

Figure 4 - Employing the "dumb-down" principle

A simple example of such semantic refinement and its use is illustrative. Figure 4 shows, in natural language, qualification of the DCMES creator element, which is defined as "An entity primarily responsible for making the content of the resource"[10]. In the example, one community has defined a qualifier *poet* as a specialization of creator, and another has defined a qualifier *author*. The nature of semantic refinement makes it possible for a search engine, which has processed the two "facts" and knows about the specialization relationship, to answer more general queries such as that shown in Figure 4. This generalizing, stripping off qualifiers and returning to the base element form, should make it possible for diverse communities to essentially ignore qualifiers that are unknown to them, yet make sense of the records. The Dublin Core community has coined the term *dumbing-down* for this process of stripping off semantic refiners and returning to base forms.

We don't doubt that such a qualification model, employed in a controlled fashion, is possible. Control would mean that a central authority, for example the DCMI, defines a fixed and relatively simple qualification set that adheres to the principles outlined above. Yet, this is not the model that has been consistently promoted by the DCMI and therein lays a problem with dumbing-down. Instead, a model that has been frequently proposed by the DCMI is that qualification occurs in a distributed, community-specific manner and can be used for increasing levels of complexity and specificity[11].

Such an approach is flawed in both its motivation and its execution. Establishing a pathway for extensive and essentially unlimited qualification of the DCMES presupposes a broader scope than the original "simple resource discovery". It frames the DCMES a one-stop cataloging standard with which records can be constructed that describe any and all facets of resources and their related entities (their creators, their intended audiences, etc.) Even if such an expanded scope were acceptable, the execution of it within the framework of the DCMES is problematic for number of reasons.

1. *Model* - Building complex descriptions on top of the flat DC data model is fundamentally flawed since the model makes it difficult to distinguish between different entities and their attributes.

---

[10] This definition is taken from the Dublin Core Metadata Element Set, Version 1.1: Reference Description at http://purl.org/DC/documents/rec-dces-19990702.htm.

[11] In fact, lack of clear guidelines for qualification and the lack of a clear definition of scope for the DCMES have led to a proliferation of Dublin Core records that are qualified in ways that defy dumbing-down. An informal study by Sigfried Lundberg documents this at http://www.mailbase.ac.uk/lists/dc-usage/2000-04/0024.html.

2. *Vocabulary* - The DCMES elements themselves were not originally engineered for such complex descriptions. The elements are completely non-normalized, ranging from ones that are essentially data types (e.g., *date*) to ones that are facets of a more general concept (e.g., *creator, contributor,* and *publisher* are all facets of agency). Other elements such as *rights* appear to contribute no information that is actionable by a computer with the context of user queries. Furthermore, qualification uncovers relationships among the elements that should be expressed structurally. If *published* is a qualifier for the *date* element, how does this relate to the *publisher* element? If *scanned* is a qualifier for the *date* element, how does this relate a *format* element that lists *tiff* as one of its values? Such engineering sloppiness is acceptable, and in fact may be the best method, for fulfilling the original DCMES intent, pidgin metadata. It is not appropriate as the basis for a rich descriptive framework.

3. *Process* - As pointed out by John Perkins of CIMI[12], the notion of refinement is implicitly community-specific. A guideline such as "qualifiers shall only refine and not extend element semantics"[13] will inevitably be interpreted by communities in fashions that will make dumbing-down impossible and defeat the interoperability goal. Our fear is a balkanization of the element set with DCMES qualified by community "A" incompatible with that qualified by community "B" and neither compatible with those who wish to use the element set in its simple unqualified form.

Discussions within the DCMI over qualification of the agent elements (*Creator, Contributor, Publisher)* demonstrate the nature of the problem. One of the qualifiers that was suggested for common use was *affiliation,* indicating the organization with which the individual is affiliated. (This qualifier was subsequently rejected based on the principles described in the next paragraph). From the perspective of many communities, *affiliation* was a clear refinement of agent semantics. However, there are serious problems with dumbing-down such a qualifier as indicated in Figure 5. The first panel of the figure shows the use of such a qualifier in a simple HTML syntax[14] for a record associated with a book by *Allison Lurie*, who is affiliated with *Cornell University*. The second panel of Figure 5 shows a simple unqualified record for a book by the author *Gary Cornell*. The third panel shows the result of "promiscuous" dumbing-down[15] of the record; stripping off the qualifiers and accepting the tokens *–Alison, Lurie, Cornell* – as values for the element creator. A simple query on the creator field would degrade the quality of the search through false hits. While the problem may seem trivial and easily fixable for this simple example, the problem becomes intractable with a huge number of records and unlimited qualification by distributed communities. It is unacceptable to either promiscuously dumb-down – making false hits the rule rather than the exception – or, as an alternative, throw out qualified values – creating the balkanization alluded to earlier.

---

[12] CIMI, the Consortium for the Interchange of Museum Information, has been one of the leading experimenters with DCMES. CIMI established an XML DTD for DCMES and worked with its members to create a large number of unqualified records. Its experiments indicated that DCMES, without qualification, served as a useful basis for coarse level resource discovery. However, later attempts to extend those experiments to the use of qualification indicated that these semantic extensions interfered with the original core interoperability requirement. The comments reported here are personal communication with John Perkins in May, 2000.

[13] This paraphrasing captures the essence of the qualification guidelines in heretofore unpublished documents of the DCMI.

[14] This syntax is for illustrative purposes and is not the syntax being considered within the DCMI.

[15] Credit goes to Ron Daniel Jr. (now of Metacode Technologies Inc.) for this phrase.

Tom Baker and others associated with the DCMI propose a solution that may prove to be a workable compromise. The solution builds on the notion that qualification of the DCMES should proceed on well-defined *qualification principles* that constrain qualifiers to basic semantic refinement (e.g., defining sub-types of elements such as *illustrator* as a qualifier for *creator*) and value encoding (e.g., defining that a *date* value is encoded according to ISO8601). These interoperability principles will be publicly disseminated, accompanied by a set of exemplary qualifiers that demonstrate the principles, in a DCMI document due third quarter 2000. Baker then proposes that qualification proceeds within distributed communities but that a *usage board*, similar to that which exists for natural language dictionaries, periodically vets qualifiers in use and maintains a registry of qualifiers with annotation indicating their conformance to the interoperability principles. Such a registry would, over time, be implemented using mechanisms such as RDF schema that would permit implementations that consume DC metadata descriptions to automatically check conformance to published interoperability standards. This solution does not prevent communities from developing qualifiers that confound interoperability (there is no solution to that problem) but is beneficial both because it is grounded in the principle of DCMES as a vehicle for simple resource description and provides a mechanism for monitoring conformance to that principle.

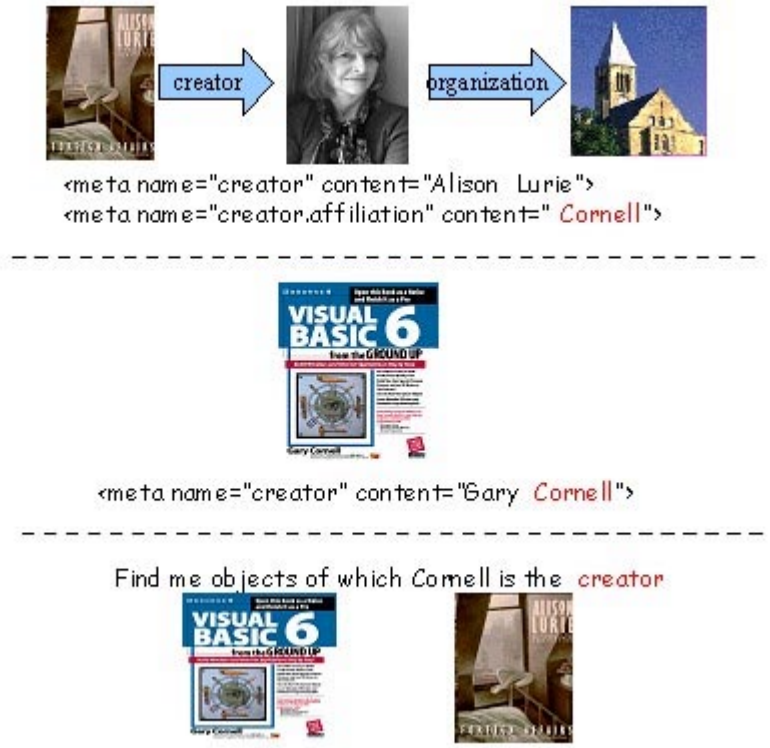

Figure 5 - Uncontrolled qualification vs. interoperability

At the time of writing of this paper (June 2000) this compromise has more or less been adopted as the official policy of the DCMI. We are optimistic that it will serve as a useful vehicle for maintaining the scope of the DCMES and thereby facilitating interoperability among DCMES records. The experiences of the past, however, where scope was lost in the context of creeping functionality, suggest that a certain level of vigilance vis-à-vis these principles will be necessary.

4 Agents of change

We began this paper by mentioning reality and its complexity. In this section we look at this complexity with a finer lens and attempt to understand how this complexity impacts descriptive schema.

The example illustrated in Figure 5, in which attributes of the agent (affiliation) are intermixed with attributes of the resource, provides a glimpse into the core of our argument, which is as follows. Key to the simplicity of descriptive schema such as the DCMES, as it was originally conceived, is their resource-centric data model. This model, as described earlier, has a simple flat structure whereby attributes (e.g. *title)* and their values (e.g., *Mona Lisa in Curlers)* are associated with reasonably mundane (real-world) objects – e.g., documents, books, pictures, and the like. Richer more complex descriptions confound this model since they inevitably include the attributes of multiple entities. This was shown in Figure 5 where attributes of the creator were intermixed with attributes of the document. Such intermixing, as shown earlier, compromises the effectiveness of the schema as a mechanism for descriptive interoperability.

We suggest that that a framework for building richer descriptions must address two needs:

1. *Expanded and Refined Vocabulary* – Qualification of the DCMES effectively expands the available vocabulary that can be used in descriptions. However, as noted earlier, this expansion needs to build on a more refined foundation that accounts for the fact that the core vocabulary will serve as the root of more complex terms.
2. *Expressive Structure* – The data model, the rules for assembling the metadata vocabulary into statements, should be able to unambiguously express the boundaries between different entities.

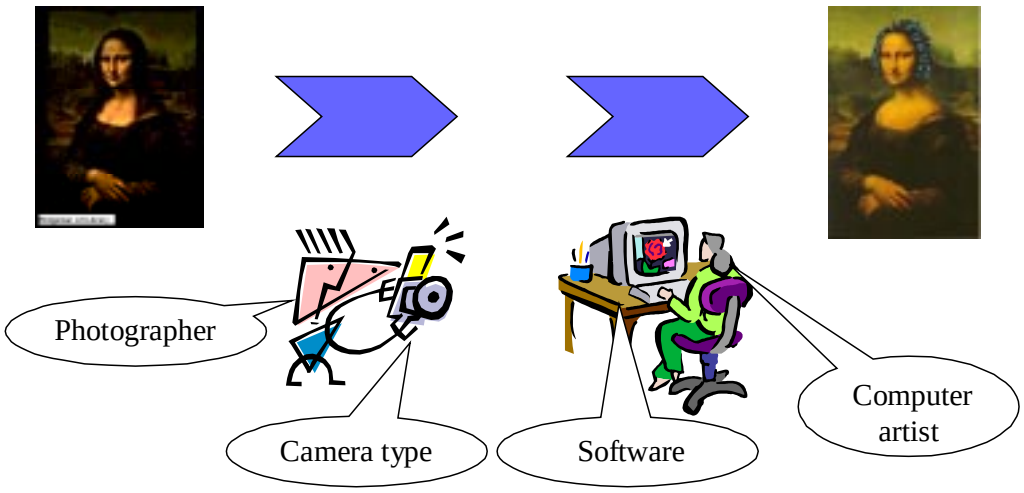

Figure 6 – A closer look at resources, entities, and their relationships

We do not suggest that these more complex needs be met by ignoring the goals that originally motivated the DCMES. Instead we advocate establishing frameworks for the creation of more

complex descriptions that can co-exist with simpler ones as separate packages in a Warwick Framework or RDF like container framework[16].

A closer look at the "Mona Lisa in Curlers" illustrates the nature of more expressive structure. Figure 6 illustrates some of the complexity underlying the "Mona Lisa in Curlers" resource. As indicated, the Castaldo work is a derivation of the original work by Leonardo da Vinci. The derivation process consisted of a number of events and agents and tools related to those events. For example, the image of the original da Vinci painting was digitized perhaps by a photographer using a specific type of digital camera. This digital image was then altered using some image processing software (e.g., Photoshop) by an artist on some type of computer system. There are numerous other details and processes not shown in the illustration. The essential point is that complex descriptions that meet the needs of specific descriptive communities will involve descriptions of these other entities. For example, the digital imaging community would certainly be interested in descriptions of the camera type and imaging software. Yet, providing those descriptive components within the flat data model, as attributes of the Mona image, presents the problem described in the earlier section.

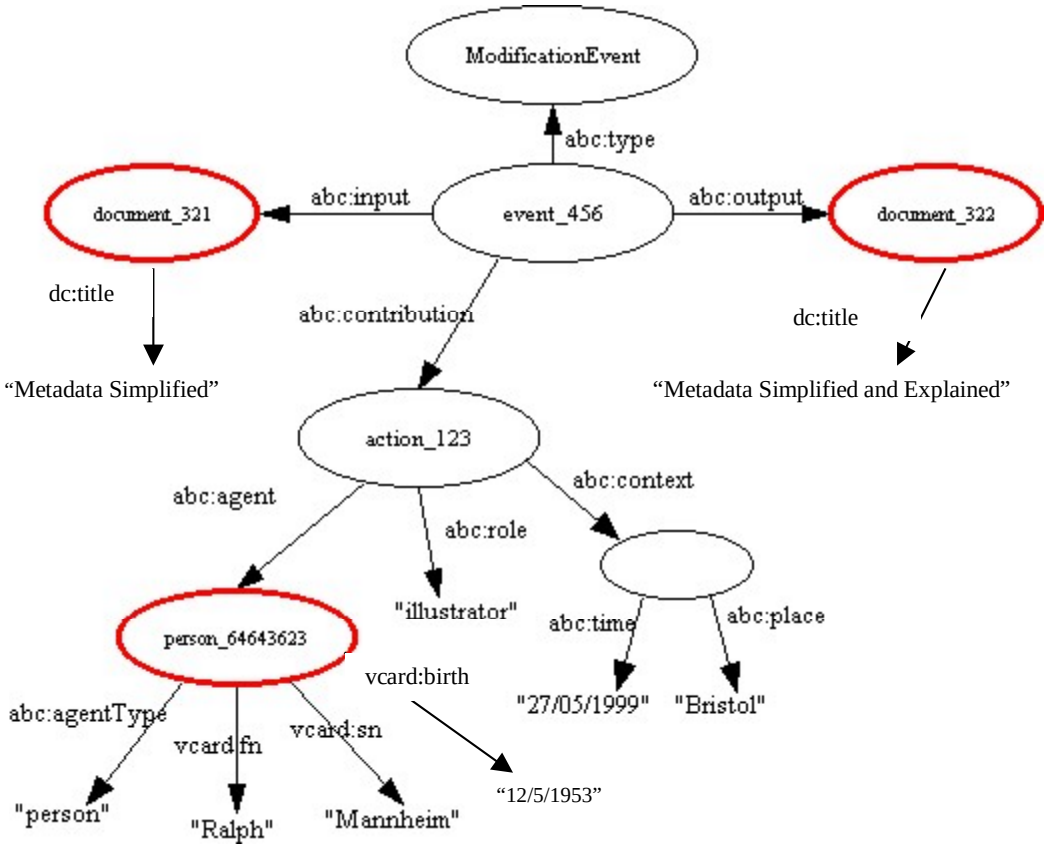

Figure 7 - Event-aware descriptive data model

[16] As we describe in [16], the notion of separate packages can be a logical, rather than a strictly physical concept. Given a well-defined and expressive underlying representation, such as that expressed in the ABC model [11], it should be possible to project automatically both simple DCMES views and more complex views.

If a flat resource-centric data model is not sufficient for richer descriptions, then what is the better alternative? This is an issue that is being examined by a number of descriptive communities. For example, the bibliographic community (i.e., Libraries) has become increasingly aware of the shortcomings of the generally flat AACR [15] model for describing resource inter-relationships. The IFLA FRBR framework [1] recognizes the lifecycle aspects of intellectual content and distinguishes between abstract works, their manifestations, and the items that are produced from those manifestations. Similarly, the rights management community [3, 21] has noted that representing transactions and the information related to those transactions is essential for metadata concerned with managing intellectual property. Finally, the archival and record-keeping communities have stated the importance of process orientation for descriptions [6, 7].

Our own work in the Harmony Project [2] builds on these earlier efforts and argues that *event-awareness* is vital for the understanding and expression of more detailed descriptions of resources. The details of this are beyond the scope of this paper and are described more fully in other documents [11, 17]. A brief summary of the event-aware concept is as follows.

Resources, as shown simply by the Mona Lisa in Curlers example and as noted in the IFLA FRBR, are the tangible result of an evolution of transformations and derivations. An important aid towards understanding this evolution of an individual resource and the derivative relationships between resources is to characterize the events that are implicit in the evolution or derivation. For example, the evolution from *work* to *expression* may contain an implicit *composing event*. The process of making implicit events explicit – making them *first-class objects* – provides well-defined attachment points for common descriptive concepts such as agency, dates, times, and roles. The clear and uniform definition of such attachment points then makes it possible for automated processes, computer programs, to unambiguously distinguish between entities and their attributes.

We have been experimenting in the Harmony Project with a highly expressive form of an event-aware model. This is illustrated in Figure 7 using an RDF-like graph representation. The figure shows a transition between two resources, labeled “document_321” and “document_322”. As shown, separate entities such as agents, roles, documents, and contexts are cleanly separated in the model. Such separation makes it possible for programs to cleanly differentiate between dates related to the agent “person_64643623” born on “12/5/1953” from the date related to the modification of the resources (the date “27/05/1999” that is the context of the modification event).

5 Is it all worth it?

Undoubtedly the model presented in **Figure 7** is complex. Furthermore, the representation, manipulation, and querying of such a model will require tools far more powerful than simple HTML META tags or existing relational databases. (Such tools are currently the subject of the extensive work by both the RDF and XML communities in the W3C).

We do not suggest that this level of complexity is the only possible alternative to the simple DC schema. Simplicity and complexity are two endpoints along a spectrum and, correspondingly, we need metadata sets that are well-suited to the varying points along this spectrum. At the same time, we need to seriously consider functionality vs. cost trade-offs when formulating metadata standards and applying them. Bill Arms points out such a trade-off in [5] and raises the issue that reduced functionality (simplicity) and reduced costs may be the proper choice for a large class of resource discovery needs. Applying Pareto’s 80/20 rule, perhaps we would see an 80% improvement in overall resource discovery, at a relatively low cost, through the application of very simple descriptive schema. Perhaps the use of a very simple core set of descriptive elements (maybe not even fifteen) makes more sense from a cost/benefit standpoint than extensive work on

complex representations and the creation of descriptions that match those representations. Certainly serious investigation and evaluation in the digital library and metadata community on this cost/functionality trade-off would be a sensible path of research.

Whatever the conclusions of such a study, we argue that trying to hide complexity in a simple metadata set such as DCMES leads to unacceptable compromises. The resulting metadata is inadequate for simple discovery purposes – dumb-down principles don't work – nor is it sufficient for complex description, which requires clear delineation of entities and their properties. Our goals for discovery and description are better served by abiding to principles of modularization and seeking solutions that are bounded by well-defined scope and purpose.

How these observations fit in with plans for the Dublin Core Metadata Set and the DCMI is obviously an open question that will only be decided by that group as a whole. These observations are only those of the author, although shared in discussions with a number of other DCMI participants. One conclusion that seems to be shared among a number of participants is that viewing the DCMES as a *record* format may be off-target. The attempt by communities to create original descriptive records using the DCMES seems to inevitably result in an effort to shoehorn richer descriptions into the DCMES elements. Rather, it appears to be much more sensible to view the DCMES as a simple projection of inevitably richer and idiosyncratic original descriptive surrogates that conform to the needs of individual communities. This projection should serve the role originally framed for the DCMES; as a means of facilitating initial discovery of these richer descriptions. The richer descriptions should not be based on qualification of the DCMES elements, but should be built on more expressive data models. Moving the focus of the DCMI away from the 15 elements to a forum for discussing these richer data models would be a sensible transition.

Acknowledgements

This paper benefits from discussions with various colleagues in the metadata and, in particular, the Dublin Core community including John Perkins, Bill Arms, Clifford Lynch, Stuart Weibel, Eric Miller, John Kunze, Jane Hunter, Dan Brickley, Sandy Payette, and Ron Daniel Jr. Special thanks go to Tom Baker for his careful reading, well-deserved criticisms, and valuable suggestions. Without his help, I'd still be confusing vocabularies, statements, and schemas. Support for the work in this document came from funding through NSF Grant 9905955.

References

[1] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions, http://www.ifla.org/VII/s13/frbr/frbr.pdf, March 1998.

[2] *The Harmony Project*, http://www.ilrt.bris.ac.uk/discovery/harmony/.

[3] *INDECS Home Page: Interoperability of Data in E-Commerce Systems,* http://www.indecs.org/.

[4] *Platform for Internet Content Selection,* http://www.w3.org/PICS/.

[5] W. Y. Arms, *Digital libraries*. Cambridge, Ma.: MIT Press, 2000.

[6] D. Bearman and K. Sochats, "Metadata Requirements for Evidence.," Archives & Museum Informatics, University of Pittsburgh, School of Information Science, Pittsburgh, PA` http://www.lis.pitt.edu/~nhprc/BACartic.html, 1996.

[7] D. Bearman and J. Trant, " Electronic Records Research Working Meeting May 28-30, 1997, A Report from the Archives Community," *D-Lib Magazine* (July/August 1997), http://www.dlib.org/dlib/july97/07bearman.html, 1997.

[8] J. L. Borges, *Other inquisitions, 1937-1952. Translated by Ruth L.C. Simms. Introd. by James E. Irby*. Austin: University of Texas Press, 1964.

[9] G. C. Bowker and S. L. Star, *Sorting things out : classification and its consequences*. Cambridge, Mass.: MIT Press, 1999.

[10] D. Brickley and R. V. Guha, "Resource Description Framework (RDF) Schema Specification," World Wide Web Consortium, W3C Candidate Recommendation CR-rdf-schema-20000327, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/, March 27 2000.

[11] D. Brickley, J. Hunter, and C. Lagoze, "ABC: A Logical Model for Metadata Interoperability," Harmony Project, Working Paper, http://www.ilrt.bris.ac.uk/discovery/harmony/docs/abc/abc_draft.html, 1999.

[12] L. M. Chan, J. P. Comaromi, J. S. Mitchell, and M. P. Satija, *Dewey Decimal Classification: A Practical Guide*, Second ed. Albany: Forest Press, 1996.

[13] L. Dempsey and S. Weibel, "The Warwick Metadata Workshop," *D-Lib Magazine*, July/August , http://www.dlib.org/dlib/july96/07weibel.html,, 1996.

[14] B. Furie, *Understanding MARC Bibliographic: Machine-Readable Cataloging*. Washington DC: Cataloing Distribution Office, Library of Congress, 1998.

[15] M. Gorman, *The concise AACR2, 1988 revision*. Chicago: American Library Association, 1989.

[16] C. Lagoze, "From Static to Dynamic Surrogates: Resource Discovery in the Digital Age," in *D-Lib Magazine*, 1997.

[17] C. Lagoze, J. Hunter, and D. Brickley, "An Event-Aware Model for Metadata Interoperability," submitted to ECDL 2000, Lisbon, 2000.

[18] C. Lagoze, C. A. Lynch, and R. D. Jr., "The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata," Cornell University Computer Science, Technical Report TR96-1593, http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593?abstract=, June 1996.

[19] O. Lassila and R. R. Swick, "Resource Description Framework: (RDF) Model and Syntax Specification," World Wide Web Consortium, W3C Proposed Recommendation PR-rdf-syntax-19990105, http://www.w3.org/TR/PR-rdf-syntax/, January 1999.

[20] D. Levy, "Cataloging in the Digital Order," presented at The Second Annual Conference on the Theory and Practice of Digital Libraries, 1995.

[21] G. Rust and M. Bide, "The INDECS Metadata Model," http://www.indecs.org/pdf/model3.pdf, July 1999 1999.

[22] S. Weibel, "Metadata: The Foundations of Resource Description," *D-Lib Magazine*, July , http://www.dlib.org/dlib/July95/07weibel.html,, 1995.

[23] S. Weibel, R. Iannella, and W. Cathro, "The 4th Dublin Core Metadata Workshop Report: DC-4, March 3-5, 1997,National Library of Australia, Canberra," *D-Lib Magazine*, June, 1997.

[24] S. Weibel and E. Miller, "Image Description on the Internet: A Summary of the CNI/OCLC Image Metadata Workshop," *D-Lib Magazine*, January, 1997.

[25] S. L. Weibel and C. Lagoze, "An Element Set to Support Resource Discovery: The State of the Dublin Core," *International Journal of Digital Libraries*, 1 (1), 1997.