DOCUMENT RESUME

ED 454 867 IR 058 158

AUTHOR Lagoze, Carl
TITLE Business Unusual: How "Event-Awareness" May Breathe Life into the Catalog?
SPONS AGENCY National Science Foundation, Arlington, VA.
PUB DATE 2000-11-00
NOTE 21p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144.
CONTRACT 9905955
AVAILABLE FROM For full text: http://lcweb.loc.gov/catdir/bibcontrol/lagoze_paper.html.
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cataloging; Change; Information Technology; *Library Catalogs; Library Role; Metadata; Models
IDENTIFIERS *Data Models; Technological Change

ABSTRACT

This paper proposes changes in the use of the catalog and the model upon which it rests. The first section describes why these changes are necessary if the library is to transition effectively into the digital age, including: the disruptive context caused by technological change; the costs associated with the catalog; the changing nature of information, how it is delivered, and who takes responsibility for organizing and describing it; and dimensions in which metadata varies from the catalog record, e.g., specialization, decentralization, and democratization. The second section describes one dimension of a new data model--event-awareness--and why it must be an important component of a new cataloging model. This section focuses on the following issues: the move away from relatively fixed physical artifacts to generally fluid digital objects; the difficulty of establishing integrity, trust, and authenticity in the networked environment; and the decentralization and specialization of resource description and problems of mapping among these descriptive vocabularies. The third section provides the outline of an event model and how it might be used. (Contains 38 references.) (MES)

Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?

Carl Lagoze
lagoze@cs.cornell.edu
Department of Computer Science
Cornell University

Prepared for Bicentennial Conference on Bibliographic Control for the New Millennium, Library of Congress, November 15-17 2000

Final version

Business Unusual

Since the nineteenth century, the modern library has been the preeminent institution of responsibility and trust in the information landscape. The Catalog has done much to make this possible by providing a uniform vehicle for access and management of a variety of information resources. Rapid growth of the Internet and the revolutionary transition from physical to digital artifacts jeopardize the role of the catalog and the library institution itself. The conservative "business as usual" perspective of libraries must shift to "business unusual" – radical changes in the catalog, its role and its composition, are needed for the library to endure in the digital age.

The increasing perception of information as a commodity suggests that there are lessons to be learned from the business world. In his popular management book Clayton M. Christensen [13] describes the threats and opportunities for businesses in the face of a *disruptive* technology[1]. Whereas a *sustaining* technology improves the performance of an established product, and therefore appeals to an existing customer base, a disruptive technology brings 'to a market a very different value proposition than had been available previously'. In his book, Christensen demonstrates how disruptive technologies establish a failure framework that historically has led to the exit of established companies from a market and, in many cases, their eventual demise.

Research libraries are unquestionably confronted with a suite of disruptive technologies, so numerous that they can be described as a *disruptive context*. The elements of this disruptive context include well-

known technical advances such as low-cost computers, the availability of broadband networking in the home and office, and advances in protocols and delivery systems on the Web. In addition, there are non-technical factors such as changes in the publishing framework (e.g,, the movement to 'author self-archiving' as described by Stephan Harnad [20]), and the increasing rate of change in many fields and corresponding increasing demand for immediate availability of research results. In combination, these factors seriously undermine the practices, and in fact the *raison d'être*, on which the research library has relied for over a century.

The Catalog stands exposed to the full force of this disruption. Over the past century research libraries have expended considerable effort evolving the catalog as a sustaining technology, adapting it to new genre of materials - audio, video, and software – and new delivery systems – from cards to MARC formatted electronic records and the integrated library systems that store and provide access to them. There is no question about the high functionality of the 'cataloging product' and its success in uniformly imposing order [26] on a variety of resources to facilitate their discovery, access, and management.

Yet, the nature of disruption, as described by Christensen, is that apparent success of a product often belies fundamental threats to its viability. In the case of the catalog these threats are both intrinsic and extrinsic.

The most substantive intrinsic threat to the viability of the catalog as we know it rests largely in the costs associated with it, which by and large is a result of its complexity. While automated sharing of cataloging records has produced substantial economies of scale, the cost of an original cataloging record, for which estimates range from 50 to 110 $US [15], makes it among the most expensive tasks in the library. The increasing burden of this expense led Bill Arms [3] to question whether the current cataloging practice can continue to exist amidst relatively static library budgets and the increasing number of resources to catalog. Arms suggests that a wiser resource allocation might be the use of cheaper automated descriptive methods even though the results would be admittedly less functional.

The extrinsic threat to the survival of the catalog comes from the changing nature of information, how it is delivered, and who takes responsibility for organizing and describing it. These changes can be characterized as follows:

- *Scale* – The sheer volume of information available on the Web and the rate of growth severely stresses the economics of traditional cataloging (described above).
- *Permanence* – The impermanence of digital information defies attempts to establish fixity, which is essential to traditional cataloging.
- *Authenticity* – The breakdown of traditional publishing models on the Web disrupts conventional mechanisms for establishing the authenticity of an information resource.
- *Variety* – The rapid introduction of new genres of digital information and the demand for specialized descriptive methods for these resources undermines the notion of *uniform access* upon which the traditional cataloging model rests.

Within this changed context, various types of metadata distinct from the catalog[2] are evolving as truly disruptive technologies – often cheaper, simpler, and admittedly less functional than the traditional catalog. In contrast to the catalog record, which is a self-contained complex organizational scheme developed and maintained by a closed community of professionals, metadata in general varies across a number of dimensions:

- *specialization* – formats and schemas often reflect the needs of specific communities
- *decentralization* – production and maintenance of metadata occurs in distinct communities of expertise that rarely share common practices or standards.
- *democratization* – some metadata initiatives, notably the Dublin Core Metadata initiative, are targeted for creation and maintenance by non-professionals.

As such, metadata offers the possibility of substantially lowering the cost of describing resources and making those descriptions more appropriate for the communities that use the resources. Furthermore, a number of metadata initiatives are focusing on descriptive domains largely unexplored by traditional cataloging records; for example rights management [35].

How can the research library maintain its enduring order-making role in the face of these disruptive challenges and technologies? How must cataloging and cataloging practices change so that libraries can continue to add value to the information infrastructure? There are no simple answers to these questions. The answers, as such, must address the institutional, technical, and theoretical foundations of cataloging practice. Hopefully, conferences such as this one provide the opportunity for evaluating the challenge and developing an inventory of ideas from which the community can move forward.

My view, as presented in this paper, is that adaptation to the networked information context will require rather radical changes to the role of the catalog and the cataloging model. This view and the material presented in this paper builds on some ideas that were put forward in the recently published National Research Council study of the Library of Congress [15], in which I participated[3]. As stated in this study:

The committee understands that it will be a tremendous challenge to change the base model for metadata (e.g., from resource-centric to relationship-centric) in a world of widespread data exchanges (the MARC records that are the basis of cooperative cataloging) and reliance on turnkey software (commercial integrated library systems that are based on MARC). However, it is certain that library-type metadata practices will at some point need to be reexamined in the light of a changed world. It is certainly valid to ask when the time will come where there is sufficient understanding of this changed world to undertake such a process. It is not productive to ignore the fact that changes are inevitable and dramatic.

The premise underlying this statement is that the resource-centric descriptive model upon which current cataloging practices are built, whereby discrete descriptive records are associated with fixed information artifacts, is incompatible with networked digital information. This new context has radically different

information entities, decentralized information production and management, and troublesome questions about authenticity and trust. It requires a model that can flexibly express the relationships between resources, abstract concepts, and multiple descriptions of those resources and concepts[4]. Complex relationships are not unique to the digital world – examples such as translations, editions, transcriptions, and the like – are well-established in physical genres and have bedeviled catalogers for years. The nature of networked digital information, however, greatly increases the complexity of resource relationships and demands a descriptive model that fully represents those relationships.

The goal of this paper is to examine one dimension of such a new data model – *event-awareness* – and why it must be an important component of a new cataloging model. Summarized briefly, an event-aware model raises events or state-transitions to first-class status, thereby allowing descriptive properties to be associated with these transitions, as well to the information entities that are inputs, outputs, and tools for these events. Using "translations" as an example, an event-aware model defines the translation act as a "first-class object" and associates properties such as the date of translation and the translator to that translation object.

The beginning of the paper describes why event-awareness is necessary for a new cataloging model. This necessity comes from both the nature of the digital objects that the catalog must describe and the role that libraries and the catalog need to play in the digital context. The latter portion of the paper provides the outline of an event model and how it might be used. It is not my intention in this paper to provide a complete solution to the problems facing the catalog. However, I hope that some of the ideas provided here may hint at the directions such a solution may need to take.

Why event-awareness?

What has changed in the digital milieu that makes an event-aware model relevant? This section focuses on the following issues:

- The move away from relatively fixed physical artifacts to generally fluid digital objects.
- The difficulty of establishing integrity, trust, and authenticity in the networked environment.
- The decentralization and specialization of resource description and problems of mapping amongst these descriptive vocabularies.

Fixity and Fluidity

Fixity is an underlying assumption of the traditional cataloging model. Fixity is realized in the two most significant first-class entities in the traditional model – the *work* and the *document*. The "first-classness" of these entities lies in the fact that they are the locus for association of attributes created by the cataloging process [37]. Fixing the work provides the locus for the association of time and space independent attributes such as author, title, edition, and subject. Fixing the document, as a particular

space-time manifestation of a work, provides the locus for associating attributes related to publication (e.g, date) and location (e.g.. library shelf). The instantiation of a cataloging record in a library's catalog establishes another layer of fixity; the linkage between a bibliographic description, a work, and document recognized as a manifestation of that work.

What is a "document" in the digital context, how does it differ from other information objects, and what is the nature of its fixity? Michael Buckland asks many of these questions in "What is a "document"' [10]. Buckland notes how the digital world, where everything exists "as a string of bits", calls into question traditional information science distinctions between documents and other information objects (e.g., processes, images, digital artwork). If "digital documents" bear a striking resemblance to "digital museum objects" or to "digital archival objects", then certainly the descriptive distinctions between these communities need to be reconsidered[5]. David Levy writes about issues of fixity and fluidity in physical and digital documents [27]. While Levy states that both physical and digital documents have degrees of fixity and fluidity, he calls attention to the significance of "technologies of fixity". Whereas both digital and physical documents move between states of fixity, there is a marked acceleration of these state transitions in digital documents; Levy calls it "the rhythm of fixity and fluidity".

The quickening of this rhythm is sufficiently problematic to call into question the integrity of the relationship of a catalog record to a digital document, thereby weakening the base integrity of the record. Examine how such relationship between record and digital object is established in the catalog. The recommended method [33] for fixing the relationship of a MARC catalog record with a networked document is through the 856 field: "Electronic Location and Access". The predominant content of this field, given the dominance of the Web for the delivery of digital content, is a URL. The fragility of URLs, or any pointer across the network[6], is well known. This fragility may take the form of catastrophic disappearance of the referenced object (known in HTTP as a 404 error), or, even more insidious, modification of the object and resulting changes in its information content (see [30]).

A solution to this conundrum – fixing the network reference – is non-trivial. One brute force "solution" is to give up on networked references, copy the objects to a local repository, and assume responsibility for their stability. As suggested in the NRC report [15], however, an attempt to indiscriminately move the "library as container" notion from the physical to the digital world is simply not realistic. Crespo and Garcia-Molina [18] suggest another solution, using techniques such as hashing for establishing bit-wise fixity. While this may appear to be a workable solution, it fails to account for the fact that exact bit-wise correspondence is not really the issue when it comes to the integrity of the cataloging record[7]. Generally, the more important issue vis-à-vis the integrity of a descriptive record is fixity of the *meaning* of the document [30] that the record purports to describe. Furthermore, bit-wise fixity is essentially meaningless when the fundamental nature of some digital objects rests in their dynamic nature (e.g., what exactly are the fixed bits the online of the New York Times at http://www.nytimes.com).

The inherent fluidity of many digital objects suggests that a "fixation with fixity" may in fact be a red herring. My suggestion is that a more realistic approach towards cataloging digital object is to incorporate fluidity into the cataloging model itself. The record should model a digital document as a

series of transition events, and should describe the nature of the events, the agents responsible for the events, and the times and places of those change events.

No doubt, this "answer" to the cataloging model opens up a number of questions that will need to be examined by the cataloging and research community:

- What is the granularity of the event record that should be recorded for digital objects? Abstractly any event can be deconstructed recursively to infinitely granular levels. The challenge in any such event model is to understand how finely granular a change history should be; the answers will inevitably be community and situation specific.
- If existing resource-centric cataloging is expensive, what are the costs of incorporating events in a new model? Like many metadata problems, there will need to be solutions that combine automated and human effort. In our Project Prism at Cornell, we are examining the use of monitoring surrogates [34] as one means of flexibly tracking status of digital objects and perhaps assisting in the maintenance of event records.

Although these and other open questions remain for an event-aware model, it does address the pervasive need to address the fluidity of a large class of digital objects. The failure of the traditional catalog to do this is a serious impediment to the transition of the library to the digital context.

A Foundation for Trust

Mechanisms for trust (and component issues of integrity, authenticity, security, and privacy), which are well-established in the bricks and mortar information context, have proven to be among the most difficult to transfer to the digital milieu. Two major national studies [16, 36] and a variety of research projects have examined issues related to how to establish trust between parties, how to be certain about the authenticity of information, how to protect privacy, how to securely protect information, and how to disseminate information in a controlled fashion in the digital realm.

What is the role of information professionals, libraries, and, in particular, the catalog (and metadata in general) in resolving such trust and integrity issues? I suggest that these organizations and tools have an essential role. Furthermore, the catalog can facilitate this role only if it has the ability to record events in the lifecycle of digital objects.

The perspectives of information professionals and researchers from a variety of communities – archival, computer science, and preservation – provide some valuable background on this issue. Picking up the theme of the previous section, the issue of fixity, or lack thereof, is a large part of the problem. As noted by David Levy [28]:

Assessments of authenticity in the world of paper and other stable, physical media rely heavily on the existence of enduring physical objects. ... What happens in the digital case if there are no stable,

enduring digital objects?

Peter Hirtle [21] describes how archivists, preservationists, and librarians share the same problem of authentication of digital objects and how this demonstrates the need for a common approach. The similarity between the issues face by archivists, preservationists, librarians, and others including the museum community is a concrete example of the questions raised by Michael Buckland in "What is a Document" [10]. The issues prevalent in each community merge as their individual media are commonly represented as bits on disk or over a network.

One approach from the archival perspective, advocated by David Bearman[8] [4], is for trusted custodial agencies to maintain metadata that records the provenance of the digital object. Bearman and his partners stress the importance of custodial control over provenance metadata, in contrast to control of the objects themselves. He reaches a conclusion about centralized storage of digital objects that sounds very similar to that of the NRC Library of Congress study [15]:

Archivists cannot afford – politically, professionally, economically or culturally – to acquire [electronic] records except as a last resort.

In later work [5] Bearman describes a metadata model for such a task – one that has a strong event orientation. Paul Conway [17] reaches a similar conclusion for the preservation community, stating that the solution to establishing integrity of digital objects lies in "documenting successive modifications to a given digital record".

We see in all of these statements the common argument that unlike physical objects, where authenticity is sometimes derivable from the object itself[9], authenticity of digital objects can generally only be established by endowing the objects with metadata, which is then maintained by trusted institutions. Clifford Lynch [29] addresses this *trust* issue directly and describes how all assertions of authenticity for digital objects are grounded in levels of trust:

...there is no question of authenticity through comparison with other copies; there is only trust or lack of trust in the location and delivery processes and, perhaps, in the archival custodial chain.

Lynch points out that there are a number of existing developing technologies that assist in establishing trust, but that all of these technologies recursively reduce to institutional trust (e.g., the institution or combination of institutions from which a provenance chain was derived); in other words, trusting the institutions that hold custody over the metadata establishing provenance.

How does this all translate to the role of the library and the catalog? The rapidly growing dependence on (born-again and born) digital information through society – in schools, business, education, and the like – presents a large-scale authenticity crisis. There is a compelling need for trusted organizations to step forward with tools to alleviate this crisis. I believe an essential value-added role that the library can add to the networked information environment is to act as a leader, or at least a primus inter pares, is

establishing trust. I believe that the catalog should be the mechanism that facilitates this role. To accomplish this, the catalog must be able to act as a record keeping tool; one that is useful for documenting the events that take place in the origination of and modifications to digital content.

Metadata as a cross-community activity

In the Warwick Framework [24] we advocated a modular model of metadata – individual descriptive *packages*, contributed by distinct communities of expertise, that are aggregated and associated with networked resources within a metadata *container*. This modular model is realized in the RDF (Resource Description Framework) [25], which the Web Consortium is advocating as the basis for Web metadata.

The decentralization of this descriptive model is dramatically different from that presumed by the catalog, which is generally framed as a "one-stop shopping" descriptive context under the control of a well-defined professional community. Undeniably, the centralization and well-defined control regimes of the traditional catalog generally leads to high-quality descriptive records, where quality is measured as adherence to well-defined standards and rules.

It is not productive, however, to argue platonic notions of quality in the face of two countervailing factors. First, the benefits of specialization in distributed, community-specific metadata are considerable. Although AACR2 and MARC encoding has proven adaptable for a variety of resources, it simply not capable of expressing descriptive semantics in specialized areas[10]. Any attempt to incorporate such specialized semantics in a general cataloging model would only lead to greater complexity and resulting greater cost. Second, the economics of cataloging, described earlier in this paper, make it impossible for libraries to ignore the cost savings possible by leveraging descriptive information supplied by metadata from external organizations.

What then is the distinct role of the library and the catalog in this decentralized descriptive environment? I suggest that a useful approach is to enthusiastically accept descriptive diversity and adopt a role as *mediator*. Rather than absorbing semantics (and descriptions) from distributed communities, libraries should promote the catalog as a mapping[11], or interoperability mechanism, amongst distributed descriptions. Technologies such as RDF and its schema language [8] make it possible to undertake such a mapping role amongst individual descriptions that are distributed across the Web[12].

I have no doubt that this suggestion might meet some resistance from my library colleagues whom have already been asked to accept the notion of providing access and some responsibility for content not entirely in their control. This suggestion takes the idea one step further by conceiving of the catalog as not only an access point for distributed resources, but as a distributed resource in its own right.

The existing resource-centric catalog is not an adequate basis for such semantic mediation. Scalable and extensible mapping among different metadata vocabularies will require a model that recognizes distinct entities that are common across virtually all descriptive schemas – people, places, creations, dates, and

the like – and that includes events as first-class objects.

The importance of this event-awareness in the model can be explained as follows. Understanding the relationship among multiple metadata descriptions (and ultimately the vocabularies on which they are based) begins by understanding the entities (resources) they purport to describe. Understanding these entities entails a comprehension of their lifecycle and the events (and corresponding transitions and transformations) that make up this lifecycle.

This argument builds upon the following observations. Descriptive communities can be distinguished by the events that are of significance to them. For example, a community that focuses on the history of production of a film may consider the "event" associated with the insertion of a certain scene into a film significant. As a result that event may be explicit in their descriptive vocabulary – for example, that community may have a metadata attribute that describes the date of the scene insertion. Another community, say one concerned with the presentation of that film on a screen, may consider that event irrelevant and may not be concerned with the "is part of" relationship of the scene to the movie.

A particular metadata description, a record from some community in some schema, actually refers to a *snapshot* of some entity taken in a particular state - a perceived fixity of the entity in a particular time and place that perforce elides events or lifecycle changes that are outside the domain of interest by the particular descriptive community. The granularity of that snapshot (and the number of elided or revealed events) varies across metadata vocabularies. For example, a Dublin Core description, intended for relatively basic resource discovery, is a particularly coarse snapshot. A Dublin Core description of a postcard of the Mona Lisa might list Leonardo Da Vinci as the creator even though numerous events took place in between Da Vinci's creation and the representation of the Mona Lisa on a postcard. On the other hand, an INDECS description, for which the events associated with transfers of rights are extremely important, might describe more fine-grained event snapshots. Establishing the identity of the events implied in the respective snapshots makes it possible to associate descriptive properties in each metadata description with these events, which then facilitates mapping among properties in the metadata descriptions.

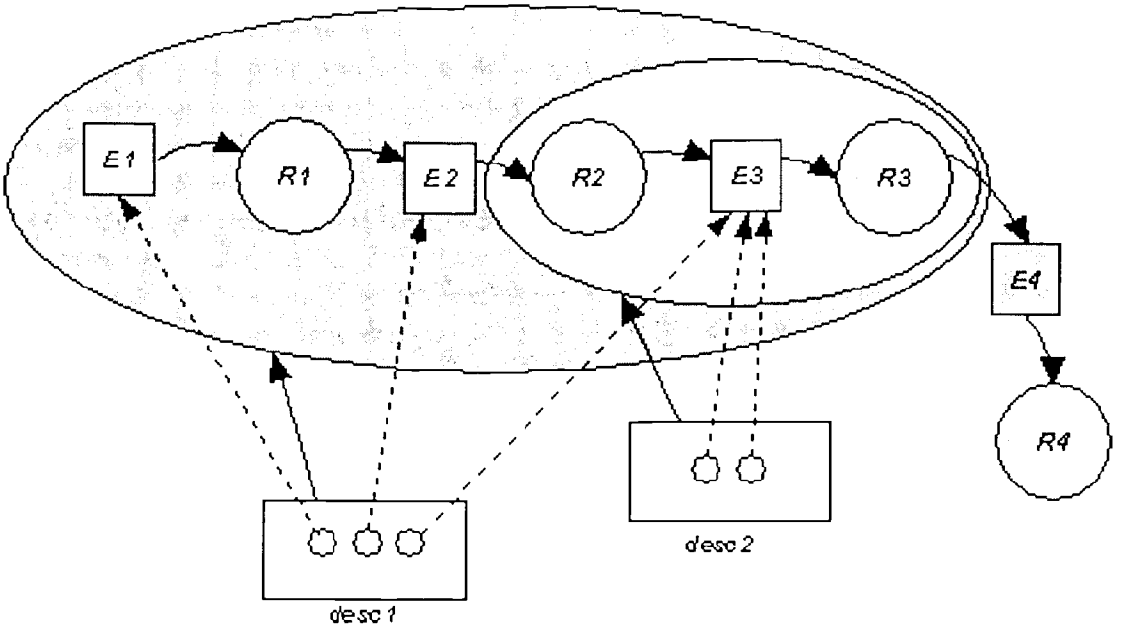


Figure 1 - Metadata and events

This basic concept of using events in mapping amongst metadata schema is illustrated in Figure 1. The larger circles represent manifestations of a resource as it moves through a set of event transitions; the events are represented by the squares interspersed between the circles. For example, event *E1* may be a creation event that produces resource *R1*. This resource may then be acted on by a translation event - event *E2* - producing resource *R2* and so on. The rectangles at the bottom of the figure represent metadata descriptions (instances of particular metadata vocabularies), and the ellipses that enclose part of the resource/event lifecycle represent the snapshot of the lifecycle addressed by that particular metadata description. For example, the larger dark-shaded ellipse represents the snapshot described by *desc1*, and the smaller light-shaded ellipse the snapshot described by *desc2*. The smaller circles within each descriptive record are the actual elements, or attributes, of the description. The dotted lines (and the color of each circle) indicate the linkage of the metadata element to an event - as shown the elements in d*esc1* are actually associated with three different events that are implicit in the snapshot. For example, the attributes (moving from left to right) may describe *creator, translator,* and *publisher,* which are actually "agents" of the events. As shown, the three rose colored elements are all associated with a single event *E3*, implying a relationship between them that can be exploited in mapping between the two descriptive vocabularies that form the basis for the different descriptions.

The Nature of an Event Model

This paper has up to this point presented a number of justifications for the incorporation of event-awareness into the cataloging model. This section illustrates event-awareness by summarizing the modeling work in the Harmony Project. The full details of the Harmony work are out-of-scope for this

paper. The interested reader should consult the research papers and reports [8, 9, 22] that provide greater details.

Over the last year, the Harmony Project has been examining a number of metadata vocabularies in an attempt to understand the entities and relationships that are common across them. The result is the so-called ABC model, which declares these entities as a set of base classes to which properties relevant to information content and its lifecycle can be attached. These entities are Resource (the primitive entity as it is defined in RDF), Event, Input, Output, Act (with Agent and Rols), and Context (with Date and Time). A UML representation [7] of the ABC model is shown in Figure 2.

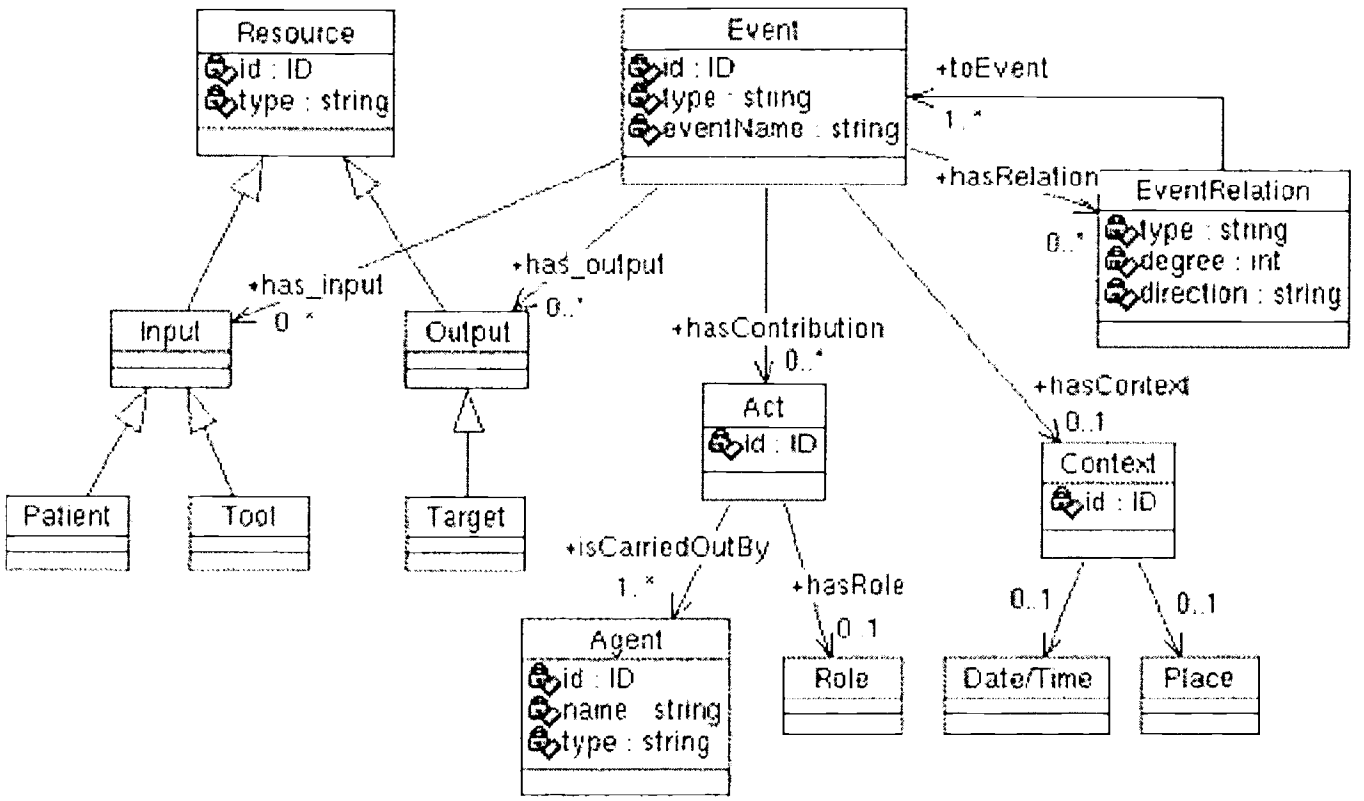


Figure 2 - UML representation of ABC model

We have tested and continue to refine this model in a number of experiments. For example, consider the following simple example of a digital audio:

The recorded performance was part of the "Live at Lincoln Center" series, made at The Lincoln Center for the Performing Arts on April 7, 1998 at 8PM Eastern time. The orchestra is the New York Philharmonic, and the musical score is "Concerto for Violin". The actual audio is a 130 minute MP3 encoding.

This example is represented in the ABC model in Figure 3 using UML-like symbols.

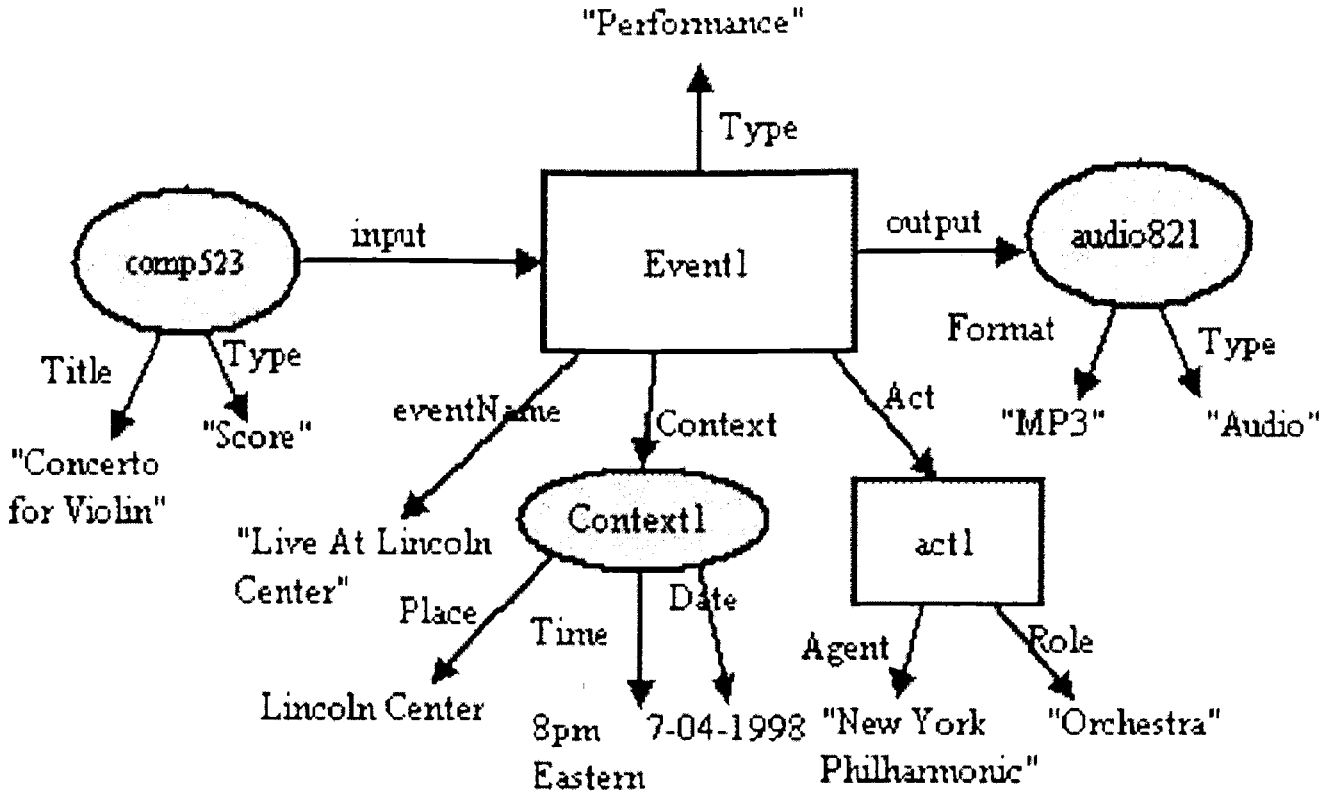


Figure 3 - Example of ABC event-aware model

As illustrated in both Figure 2 and Figure 3, the model provides well-defined attachment points for various properties, by explicitly representing entities. Thus, the date of performance is defined as a property of the "performance" event, rather than as a property of the audio. The usefulness becomes clearer if we expand the example and include a "composition" event that feeds into "comp523" in Figure 3, with a "Date" property of 3-01-1804. This stands in contrast to a resource-centric model in which both dates (and perhaps) several others would be listed as cataloging properties of the single audio resource.

At this point we have experimented with the ABC model for mapping between a number of metadata schemas including Dublin Core, ID3 tags embedded in MP3, MPEG-7 descriptions in DDL, and the CIDOC CRM model. We have demonstrated that it is possible to do simple mappings using XML schema [6, 38] and XSLT [14]. The limitations of these tools has constrained the expressiveness of these mappings and in Harmony we are beginning to experiment with more powerful tools such as a metadata term ontology and the use of a general mapping rule language.

Conclusion

This paper has proposed radical changes in use of the catalog and the model upon which it rests. It has described why these changes are necessary if the library is to transition effectively into the digital age.

Changes of such magnitude obviously require careful consideration and strategic planning on the part of libraries and associated information professionals. They will require libraries to take a prominent role in research initiatives and, correspondingly, allocate resources to develop and hire the professionals capable of participating and leading such research. Being too conservative will only widen the disconnect between the rapidly changing information environment and the manner in which libraries profess to manage it. I end with an appropriate admonition from the NRS report [15] (taking the liberty to replace explicit references to "the Library of Congress" with "libraries"):

The alternative to progress along these lines is simple: [libraries] could become a book museum....But a library is not a book museum. A library's value lies in its vitality, in the way its collections grow, and in the way that growth is rewarded by the diverse and innovative uses to which its collections are put. [Libraries] will, by the choices [they] make now and in the next months and years, determine how much of that vitality will survive into the new millennium and how well [they] can avoid subsiding into diminished relevance.

Acknowledgements

This paper benefits from discussions and joint research with a number of people. The approximately 14 months spent on the NRC Library of Congress study were among the most valuable in my life and I thank all my colleagues there for their inspiring thinking. I owe special thanks to my colleagues in the Harmony project, especially Jane Hunter who has done wonderful work on metadata mapping using the Harmony ABC model. Thanks also to Clifford Lynch for referring me to the valuable papers form the CLIR authentication workshop. Finally, I express gratitude to the organizers of the workshop for inviting me and giving me the chance to think about these issues. Naomi Dushay also supplied invaluable editing advice. Support for work on this paper came from NSF Grant 9905955.

References

[1] *Dublin Core/MARC/GILS Crosswalk*, http://lcweb.loc.gov/marc/dccross.html.

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

[3] W. Y. Arms, "Automated Digital Library: How Effectively Can Computers Be Used for the Skilld Tasks of Professional Librarianship?," *D-Lib Magazine*, 6 (7/9), http://www.dlib.org/dlib/july00/arms/07arms.html,, 2000.

[4] D. Bearman, "An Indefensible Bastion: Archives Repositories in the Electronic Age," Archives and Museum Informatics, Pittsburgh, Technical Report 13, 1991.

[5] 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.

[6] P. V. Biron and A. Malhotra, "XML Schema Part 2: Datatypes," World Wide Consortium, W3C Working Draft WD-xmlschema-2-2000025, http://www.w3.org/TR/xmlschema-2/, April 7 2000.

[7] G. Booch, J. Rumbaugh, and I. Jacobson, *The unified modeling language user guide*. Reading Mass.: Addison-Wesley, 1999.

[8] 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.

[9] 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.

[10] M. K. Buckland, "What is a "document"?," *Journal of the American Society of Information Science*, 48 (9), 1997.

[11] P. P. S. Chen, *Entity-relationship approach : the use of ER concept in knowledge representation*. Washington, D.C.: IEEE CS Press, 1985.

[12] P. P. S. Chen, *The entity-relationship approach to logical database design*. Wellesley, Mass.: QED Information Sciences, 1991.

[13] C. M. Christensen, *The innovator's dilemma : when new technologies cause great firms to fail*. Boston, Mass.: Harvard Business School Press, 1997.

[14] J. Clark, "XSL Transformations (XSLT)," World Wide Web Consortium, W3C Recommendation REC-xslt-19991116, http://www.w3.org/TR/xslt, November 16 1999.

[15] Committee on Information Strategy for the Library of Congress, *LC21: A Digital Strategy for the Library of Congress (2000)*: National Academy Press, Washington, DC, 2000.

[16] Committee on Intellectual Property Rights in the Emerging Information Infrastructure, *The Digital Dilemma: Intellectual Property in the Information Age*. Washington, D.C.: National Academy Press, 2000.

[17] P. Conway, "The Relevance of Preservation in a Digital World," Northeast Document Conservation Center, Andover, MA http://www.nedcc.org/plam3/tleaf55.htm, February 1999.

[18] A. Crespo and H. Garcia-Molina, "Archival Storage for Digital Libraries," presented at Third ACM International Conference on Digital Libraries, Pittsburgh, PA, 1998.

[19] L. Duranti, *Diplomatics: New Uses for an Old Science*. Lanham, MD: Scarecrow Press, 1998.

[20] S. Harnad, "Free at Last: The Future of Peer-Reviewed Journals," *D-Lib Magazine*, 5 (12), http://www.dlib.org/dlib/december99/12harnad.html,, 1999.

[21] P. B. Hirtle, "Archival Authenticity in a Digital Age," presented at Authenticity in a Digital Environment, Washington, D.C., 2000.

[22] J. Hunter and D. James, "Application of an Event-Aware Metadata Model to an Online Oral History Archive," presented at ECDL 2000, Lisbon, 2000.

[23] ICOM/CIDOC Documentation Standards Group, *CIDOC Conceptual Reference Model*, http://www.ville-ge.ch/musinfo/cidoc/oomodel/.

[24] C. Lagoze, "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata," *D-Lib Magazine*, 2 (7/8),http://www.dlib.org/dlib/july96/lagoze/07lagoze.html,, July/August, 1996.

[25] 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.

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

[27] D. M. Levy, "Fixed or Fluid? Document Stability and New Media," presented at 1994 European Conference on Hypermedia Technology, 1994.

[28] D. M. Levy, "Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment," presented at Authenticity in a Digital Environment, Washington, D.C., 2000.

[29] C. Lynch, "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis on the Central Role of Trust," presented at Authenticity in a Digital Environment, Washington, D.C., 2000.

[30] C. Lynch, "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information," *D-Lib Magazine*, 5 (9), http://www.dlib.org/dlib/september99/09lynch.html,, 1999, September.

[31] S. McKemmish, G. Acland, N. Ward, and B. Reed, "Describing Records in Context in the Continuum: the Australian Recordkeeping Metadata Schema," Monash University, Records Continuum Research Group http://www.sims.monash.edu.au/rcrg/publications/archiv01.htm, 1998.

[32] Metadata Ad Hoc Working Group, "Content Standard for Digital Geospatial Metadata," Federal Geographic Data Committee, Washington DC FGDC-STD-001-1998, http://www.fgdc.gov/standards/documents/standards/metadata/v2_0698.pdf, 1998.

[33] N. B. Olson, *Cataloging Internet Resources*. Dublin, OH: OCLC Online Computer Library Center, Inc., 1997.

[34] S. Payette and C. Lagoze, "Value-Added Surrogates for Distributed Content: Establishing a Virtual Control Zone," *D-Lib Magazine*, June , http://www.dlib.org/dlib/june00/payette/06payette.html,, 2000.

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

[36] F. B. Schneider and National Research Council (U.S.). Committee on Information Systems Trustworthiness, *Trust in cyberspace*. Washington, D.C.: National Academy Press, 1999.

[37] E. Svenonius, *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press, 2000.

[38] H. S. Thompson, D. Beech, M. Maloney, and N. Mendelsohn, "XML Schema Part 1: Structures," World Wide Web Consortium, W3C Working Draft WD-xmlschema-1-

2000225, http://www.w3.org/TR/xmlschema-1/, April 7 2000.

---

[1] Thanks to Stuart Weibel of OCLC for introducing me to the notion of metadata as a disruptive technology.

[2] Traditional cataloging is one form of metadata – a form of description or "data about data".

[3] This paper is not intended as a summarization of that report. Although this paper benefits from conversations during the NRC study, the thoughts and opinions expressed here are of the author alone.

[4] The interested reader may wish look at the many good sources of information on data modeling including the classic materials on E-R (entity relationship modeling) [11, 12], and the excellent work in various descriptive communities [2, 23].

[5] The museum metadata community [23] and archival metadata community [31] have recognized the importance of event-oriented models.

[6] Actually a URL is one of but several types of "locators" in the 856 field. For example, the contents may be a URN; a so-called permanent and location-independent identifier. While the permanence of a URN is an attractive concept, from the implementation point of view a URN is simply one or more levels of indirection to a URL, where permanence rests on the stability of the agency maintaining the indirection mechanisms. Moral: URNs really provide no real technical solution to the problems of fixity discussed here.

[7] For example, hashing techniques are generally insensitive to the difference between a trivial font change and a change in the wording of a paragraph.

[8] The community of people advocating this approach with David Bearman is collectively known as the "Pittsburgh Project".

[9] See explanations of the science of diplomatics in [19].

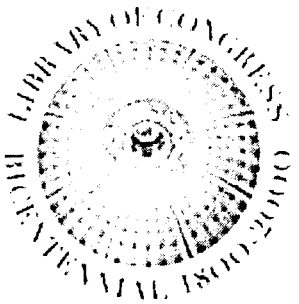
[10] Consider the highly descriptive FGDC standard for geospatial resources [32].

[11] Mapping among descriptive formats is not entirely new to the cataloging community. There have been numerous experiments with *crosswalks* between MARC-based cataloging records and metadata in its various forms [1]. These crosswalks generally presume a role where the catalog is the superior form

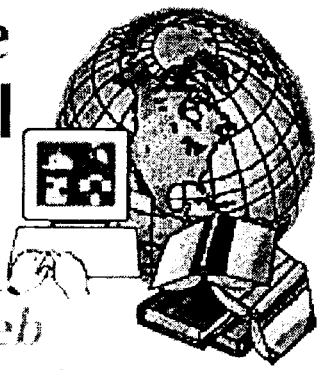
and other metadata forms have reduced functionality and, therefore, importance. This is different from the catalog acting as a mapping mechanism among distributed metadata packages that in their composite equally contribute to the "cataloging record" of a digital object.

[12] The result is a Warwick Framework-like container whose packages are distributed across multiple servers.

Library of Congress
Comments: lcweb@loc.gov (October 19, 2000)



Bicentennial Conference on Bibliographic Control for the New Millennium



Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Conference Hom Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Carl Lagoze

Digital Library Scientist
Dept. Of Computer Science
Cornell University
Ithaca, NY 14853

Business Unusual: How "Event-Awareness" May Breathe Life Into the Catalog?

About the presenter: Carl Lagoze is Digital Library Scientist in the Computer Science Department at Cornell University. In this capacity he leads a number of digital library research efforts in the Department and across the university, collaborating with the University Library and Office of Information Technology. Mr. Lagoze's research is funded through a number of NSF, DARPA, and industry grants, most notably a major grant from the multi-agency Digital Libraries Initiative Phase 2. In general, this research can be characterized as investigations into the technical and organizational issues in the development and administration of distributed digital libraries. The recent focus of this research is on policy: What are the policies that need to be asserted to ensure the reliability, security, and preservation of content and services in distributed digital libraries and what are the mechanisms for enforcing those policies? Mr. Lagoze is the co-inventor of Dienst, a widely deployed protocol and architecture for distributed document libraries. He is also the co-author of the Warwick Framework, a modular metadata model for digital content, which is a conceptual basis for the Resource Description Framework (RDF), now a WWW metadata standard. Mr. Lagoze's professional activities include serving on the advisory committee of the Dublin Core Metadata Initiative, serving on the program committee of U.S. and international digital library conferences, and numerous talks both in the U.S. and internationally on his research on metadata and digital library architecture.

Cataloging Directorate Home Page

Library of Congress Home Page

Full text of paper is available

Summary: (revised 11/1/00)

The speaker proposes that the digital context presents a dramatically new context than that which was addressed by the traditional cataloging model. Whereas the catalog has depended on relatively fixed resources delivered by a relatively stable set of role players (publishers, authors, information intermediaries), the digital context is characterized by fluidity in both content and those who provide it. The speaker proposes new roles for the catalog based on this new reality and a new data model that meets the needs of these roles. An "event-aware" model of cataloging, one that recognizes digital resources as inherently dynamic, will allow the research library to adapt to the realities of the digital millenium.

Library of Congress
November 1, 2000
Comments: lcweb@loc.gov

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").