Event Representation in Temporal and Geographic Context

Ryan Shaw and Ray R. Larson

School of Information
University of California, Berkeley
Berkeley, California, USA, 94720
ryanshaw,ray@ISchool.Berkeley.EDU

Abstract. Digital Libraries often fail to connect their contents to the wider context of information resources available that are about the same persons, related persons, places, or time periods and the events that happen to those persons, at those places and in a given time period. In this paper we critically examine a number of standards for organizing collections of archival, historical, genealogical, news, and personal information to see what kinds of resources they offer for modeling events. We present our ongoing work using event models to enable effective use of digitized biographical texts for a digital library environment.

1 Introduction: Modeling Events

Events are important organizing principles. Linked sequences of events constitute the stories we tell. Scheduled events crowd our calendars and bulletin boards. Recounted events fill news reports. Events and the people and things who participated in them are the raw material of our histories and mythologies. Arguably, events—things that happened, are happening, or will happen—are a primary way we plan, understand, and remember our lives. Yet our information systems often treat events as second-class entities, ephemeral bits of metadata for decorating more persistent entities like documents, objects, and resources. Or worse, life events and event metadata are ignored completely in Digital Libraries under the assumption that if the text of a document discusses such events, that is sufficient for access and linking to related events, persons, places, and time periods. This paper seeks to challenge such assumptions and asks, what would an event-oriented approach to organizing information look like?

Our concern in the "Bringing Lives to Light: Biography in Context" project (http://ecai.org/imls2006/) is to design, demonstrate, and evaluate standards and best practices for encoded markup, embedded search links, and editing tools that could be used to create more powerful digital biographical texts by connecting such texts with their geographical, temporal, and cultural contexts. Standardized ways to represent events would have a number of benefits for our project and for the wider world of Digital Libraries, including the creation of reusable infrastructure for managing documents in an event-centric manner, and

the opportunity to create generic tools for visualizing, exploring, and mining collections of media generated during, by, and about different kinds of events.

In this paper we examine and compare some different approaches for modeling events. Using a set of requirements for event models proposed by multimedia computing researchers [11], we survey a number of standards for organizing collections of archival, historical, genealogical, news, and personal information to see what kinds of resources they offer for modeling events. Finally we discuss ongoing work using event models to enable effective use of digitized biographical texts for a digital library environment.

2 Requirements for an event model

Westermann and Jain propose a number of requirements for a flexible, reusable approach to modeling events [11]. They suggest that an event model should provide resources for uniquely identifying events, distinguishing different types of events, supporting both discrete and continuous events, expressing various temporal aspects of events, linking events to places, describing the structure of participation (who and what are involved) in events, and associating events with documents. Moreover, they contend that such models ought to support various ways of relating events, and that these relations should themselves be first-class entities, so that different perspectives can be expressed via different sets of relations among events. Finally, they argue that it should be possible to qualify any of the associations or attributes of events as uncertain. In the following sections we examine a number of event-related standards to see how they measure up to these requirements.

Archivists, historians, and genealogists have developed a number of standards for representing events of interest, usually key events in people's lives [10]. News organizations and creators of calendaring software have also developed event-related standards. Here we will focus on standards that attempt to provide some non-trivial structure for event representations, skipping over those that simply set aside fields for free-text descriptions of events. The specific standards we consider here are the Encoded Archival Context tag library [1], the CIDOC Conceptual Reference Model [3], the Historical Mark-up and Linking Project XML schema [9], the GENTECH Genealogical Data Model [6], the Genealogical Data Communication XML schema [5], the Internet Calendaring and Scheduling Core Object Specification [4], and the International Press Telecommunications Council's Events Markup Language G2 [7].

2.1 Uniquely identifying events

The ability to uniquely identify events is critical for linking them to other kinds of metadata (people, places, time periods, things) and to documents. This could provide valuable linkages within a single Digital Library context. Ideally, we could uniqely identify events in a global context, allowing us to create references in an open, distributed manner without danger of collisions. In practice, few of

these standards provide clear mechanisms for assigning globally unique event identifiers.

The `event` tag of the Encoded Archival Context (EAC) tag library has an `id` attribute that ensures that events can be uniquely identified within a single valid XML document. In theory, if one published an XML event directory at a stable URI, this URI plus the unique `event` element ids would result in globally unique identifiers for the listed events. However, the standard does not discuss this, nor do ids seem to be required on `event` element. The Genealogical Data Communication (GEDCOM) XML schema, on the other hand, explicitly requires `id` attributes on its `EventRec` elements, as well as allowing `ExternalID` sub-elements for extending the schema to include various system-specific event identifiers. The Historical Mark-up and Linking Project (HEML) XML schema requires unique event URIs, specified via `uri` attributes on its `Event` elements.

The CIDOC Conceptual Reference Model (CRM) is designed to be an abstract data model that is realized by technical implementations within specific organizations. Thus the task of uniquely identifying entities is left up to implementors. Though the CIDOC CRM describes a mechanism for assigning names (`Appellation`s) to events, these names are not necessarily intended to be unique identifiers. The GENTECH data model, while also intended as an abstract specification, specifies that an event must have a unique `Event-ID` attribute.

The Internet Calendaring (iCal) Specification allows globally unique identifiers for events to be specified via a `UID` property on events, but this property is not required. Finally, the Events Markup Language (EventsML) specification states that persistent information about an event must have both a globally unique identifier (`guid`) and an unambiguous (but not necessarily unique) identifier (`conceptId`) assigned by a naming authority.

2.2 Distinguishing different types of events

The EAC does not provide any structured way to distinguish among different types of events. In HEML documents, one may classify an event using a single keyword per language, which rules out multiple classifications of an event from different perspectives. iCal also allows language-specific keywords, but is less restrictive in that it allows multiple keywords per language for an event. iCal proposes a core set of controlled terms for use as English-language category values, all of which are specific to the personal calendaring domain. In EventsML one can use any number of `facet` properties to distinguish among different kinds of events, using either free text or values from a controlled vocabulary.

GEDCOM events have a required free-text keyword `Type`, and an optional `VitalType` drawn from a small controlled vocabulary of terms. (`VitalType`s are primarily concerned with events such as birth or death that result in the creation or disappearance of entities.) GENTECH also requires events to have types. Rather than simple keywords, a GENTECH event type is a structure that specifies the roles that may be filled by participants (`Persona`s) in events of that type, or the (temporal or causal) relationships events of that type may have with other events. Finally, the CIDOC CRM defines three subclasses of `Event`

(`Beginning of Existence`, `End of Existence`, and `Activity`), and events can be more finely distinguished by linking them to taxonomies of event types. An event may be linked to more than one type, providing a flexible and powerful way to classify events.

2.3 Supporting discrete and continuous events

Since GEDCOM allows only a single date per event, it appears to be the only standard that does not explicitly support both discrete (instantaneous) events and continuous events (events with duration). In iCal and EventsML event times are specified as date ranges, but events that have a start time and no end time are considered to be instantaneous. EAC, HEML, and GENTECH all support specifying either point dates or date ranges for events. In the CIDOC CRM events have time-spans with durations, and these durations may be zero for instantaneous events.

2.4 Expressing temporal relations of events

Though all of the standards except GEDCOM support both discrete (instantaneous) and continuous (durative) events, they offer varying degrees of flexibility in how these points in time or time ranges may be specified. All of the standards support specifying times as calendar dates and times of day. But in some cases it may be necessary to be able to express relative times, so that one can assert that an event happened before, during, or after some time period or other event. This is especially useful if temporal ordering or containment of events is known but specific dates are not. It is also necessary for expressing domain-specific logical times, such as "the last debate before the Democratic primaries." EAC, iCal, GEDCOM XML, and EventsML have no explicit support for expressing relative temporal relationships. However, EAC does allow nesting of events (which might be interpreted as containment), and EventsML enables event containment via `broader` and `narrower` properties linking events.

HEML currently offers very limited support for modeling relative times, in that the date of an event can be specified to be after another event's date. However, the HEML `ChronologicalRelationship` element appears to have been designed to support other kinds of relative temporal relationships in the future. The GENTECH data model allows the statement of such relationships as `Assertion`s linking two events. Furthermore, since an `Assertion` is identified with the specific person making that assertion, the GENTECH model supports the creation of conflicting temporal relationships expressing different points of view, a significant advantage.

Unlike the GENTECH model, which supports the creation of relative temporal relationships but does not provide any guidance as to what these relationships should be, the CIDOC CRM defines a number of temporal relationship types based on Allen's temporal logic [2], including containment, overlapping, simultaneity, and disjuncture as well as basic ordering. Like the GENTECH model, the CIDOC CRM also supports the expression of alternative opinions via the

creation of multiple, potentially conflicting relationships. However, it is not clear how to link individual relationship assertions with specific opinion-holders.

2.5 Linking events to places

In the all of these standards, one can associate events with location terms, whether geographical place names (such as the Red Sea) or names of political jurisdictions (such as Macon, Georgia). EAC also provides a mechanism for linking these terms to specific controlled vocabularies or gazetteers via `valueauth` and `valuekey` attributes on the place element, which means that these terms could theoretically be resolved to physical coordinates. In HEML, GEDCOM XML, iCal, and EventsML, events can be linked with geographical locations specified as named latitude-longitude pairs or GPS coordinates (but not bounded geographic regions). Using the CIDOC CRM one can specify physical coordinates for places where events occurred, and these coordinates are not limited to simple latitude-longitude pairs, which theoretically allows bounded regions to be specified. The GENTECH model does not appear to have a standardized way to link place terms to physical coordinates.

2.6 Describing structures of participation

Perhaps the most interesting aspect of events is their function as conceptual nodes that link various actors and entities to particular places and times. The spectrum of different structures of participation in events range from very simple (a single scientist's important epiphany) to very complex (the Second Iraq War). Given their different areas of focus, the standards discussed here cover different parts of that spectrum.

Events in the EAC are understood to be elements in a chronological list of noteworthy events in the life of a person, family, or organization. Thus there is no support for expressing participation in an event as such: the sole participant in an EAC event is implied by the context in which it appears.

Most of the other standards use variations on a role-based approach to expressing structures of participation. EventsML event details can include any number of `participant` properties qualified by `role` attributes. Participants in iCal events are limited to scheduling-related roles, though there is a provision for "experimental" roles that could presumably be used to circumvent this restriction. HEML and GEDCOM XML simply allow participants in an event to be distinguished by a keyword indicating the role they played. It is not clear whether these keywords are to be drawn from a controlled vocabulary or chosen freely depending on the event.

The GENTECH model adds a bit more structure by allowing the type of an event to specify roles to be filled by various participants. For example, a "dissertation" event type might have roles for "student" and "advisor." Assertions expressing participation in specific dissertation events would then link specific people to these roles. Since the GEDCOM and GENTECH models are intended for genealogical use, their notion of participation in events is limited to people:

there is no way to describe the roles that non-human entities or objects may have played in an event, except as evidence for it.

Only the CIDOC CRM enables rich modeling of both the actors *and* the entities involved in an event. Actors can participate in events, and entities can be present at events. Certain kinds of entities such as ships and buildings may be sites at which events occurred. Events modeled as activities (rather than more abstract changes of state) can furthermore specify roles for the actors performing those activities, and relationships among various roles can be related in a taxonomy. Activities can also specify "modes of use" for objects used by these actors, and these too can be organized taxonomically. Finally, arbitrary entities can be described as having motivated or influenced some activity.

2.7 Associating events with documents

If we wish to use events as conceptual structures for navigating through collections of documents, it is important to be able to link them to specific documents or parts of documents. Yet this is where most of standards considered here are weakest: though most allow describing locations of event-related documents, there is little to no support for network links to digital documents.

EAC only allows very coarse linking of events with documents, in that entire EAC instances describing the events in the life of a person, family, or organizations can be associated with entire resources such as archival records, bibliographic items, or museum objects. It is not clear how single events might be linked to such resources, other than creating EAC instances that only contain a single event. In GEDCOM XML and the CIDOC CRM individual events can be linked directly to documents or other sources of evidence, while in the GENTECH model events are linked to evidence via assertions involving those events. None of these four standards has explicit provisions for resolving links to documents in a networked environment.

HEML, and iCal are more amenable to networked digital usage. In HEML individual events can be linked to multiple documents, either via bibliographic records or via hyperlinks to networked digital documents. iCal allows multiple documents to be attached to events via URIs. Absent a standard way to specify URIs linking to pieces of documents, however, neither of these standards can support linking events to (for example) specific frames of a video.

EventsML is also designed for networked environments, but it inverts the approach taken by HEML and iCal. Rather than linking event descriptions to related documents, users of EventsML link documents to event descriptions. For example, one would link a photograph of a wedding to information about that wedding by specifying the wedding's identifier as the value of the `subject` property on a `News Item` with the photograph as its content.

2.8 Relating events to one another

We discussed above temporal relations between events, including containment (sub-events within events). But there are additional kinds of relations among

events that would be useful to model, such as causality. Or one might want to relate alternative descriptions of an event, such as the differing accounts of the death of a samurai in Akutagawa's famous tale Rashōmon.

Only the GENTECH model supports the creation of arbitrary kinds of relations between events. Using GENTECH researchers can express event relations in the form of assertions that they support with documentary evidence or other assertions. Assertions may conflict with one another; in fact one researcher may make an assertion that negates or refutes another.

The CIDOC CRM also supports the creation of event relations independent of the events being related, but defines a taxonomy of types to which these relations must belong. Fortunately these types are fairly generic; in additional to the various kinds of temporal relations mentioned above, one event can influence or motivate another (variations on causality), or one event can be characterized as preparing for another (reflecting an interpretation of intentionality). Since the designers of the CIDOC CRM intended for users to extend it for particular domains, these relations could be sub-classed to create more specific kinds of relationships. The CIDOC CRM does not specify how relations might be associated with particular perspectives.

2.9 Expressing uncertainty

Knowledge about what has happened, is happening, or will happen is never perfect. Ideally one could capture this imperfection in an event model and make statements about the uncertainty of particular aspects of or relationships between events. Only the GENTECH model provides a robust way to make such statements. GENTECH users can associate assertions with a level of "surety" or certainty. The model supports the use of different approaches to characterizing certainty by allowing an assertion to point to the surety scheme being used.

HEML and EAC have some support for expressing uncertainty surrounding precise dates. HEML handles uncertainty of dates via an UncertainDateGroup element. The EAC mentions a certainty attribute on dates, but this attribute doesn't actually appear in the DTD. iCal does not support uncertainty beyond allowing events to be marked as "tentative." EventsML, GEDCOM XML and the CIDOC CRM do not have any specific provisions for recording uncertain information about events.

2.10 Summary of event-related standards

Table 1 summarizes some of the key differences discussed in this survey of event standards. As they were designed for specific applications rather than comprehensive event modeling, none of these standards fulfill every requirement Westermann and Jain propose. The CIDOC CRM and the GENTECH data model provide the richest and most flexible set of modeling tools, but even they lack some of the features identified above. Fortunately, there do not seem to be any critical conceptual conflicts among these standards, suggesting that it may be possible to define mappings among them.

Table 1. Key differences among standards that support event modeling

Standard	EAC	CIDOC	HEML	GENTECH	GEDCOM	iCal	EventsML
Unique event IDs?	weak		yes	yes	yes	yes	yes
Discrete and continuous events?	yes	yes	yes	yes		yes	yes
Relative temporal relationships?		yes	weak	yes			yes
Resolution to physical coordinates?	weak	yes	yes		yes	yes	yes
Participation of actors *and* objects?		yes					
Networked links to documents?			yes			yes	yes
First-class event associations?		yes		yes			
Uncertainty support?	weak		weak	yes			

3 Representing Life Events

We undertook this survey of event-related standards as part of an ongoing effort to enable more effective use of biographical texts in a digital environment. Currently most biographical reference sources provide prose narrations of people's lives. These texts are useful, but in a digital library environment it is desirable to also have reference services that can provide structured representations of life events. Clients of such services could send queries about people (specifying either names or unambiguous identifiers) and in return receive identifiers of important dates, places, artifacts, and other persons related to those people. Interfaces for browsing online resources about people could then use these identifiers to dynamically construct links that search for related resources, or to retrieve and directly display related resources, or to assemble maps and timelines that contextualize those people's life events.

We can view such services as an evolution of traditional authority records that identify people and provide links to alternative names and relevant evidence. Web services like OCLC's WorldCat Identities (`http://orlabs.oclc.org/Identities/`) have demonstrated the first steps in this evolution, showing how authority records can be enhanced with additional information about people and organizations. Advancing to the kind of biographical reference services we envision requires standards for representing and exchanging information about life events. This was our motivation for assessing the state of event-related standards.

Each of the standards discussed above was designed for specific purposes, none of which include our "authority records on steroids." Though we might have selected one and twisted it toward our ends, ultimately we decided that it

would be better to look elsewhere for suitable standards, while keeping in mind the desirability of being able to map to the standards examined here.

We settled upon using the BIO (`http://vocab.org/bio/0.1/`), an RDF vocabulary meant to be used in conjuction with FOAF (`http://xmlns.com/foaf/spec/`) to represent simple biographical information. A key advantage of using an RDF vocabulary is modularity: we can use classes, predicates, and properties from other RDF vocabularies as needed to fulfill our requirements, as will become clear in the following discussion of how BIO stacks up against Westermann and Jain's requirements.

BIO meets a few of the requirements simply by virtue of the fact that it is an RDF vocabulary. We can create instances of the `Event` class and assign them globally unique URIs. Since RDF was designed to describe web resources, we can naturally use it to link and annotate digital documents. Furthermore, the open world assumption on which RDF semantics are based allows us to express multiple conflicting statements about events. And because RDF supports reification of statements, we can treat those statements as first-class objects, allowing us to describe their provenance.

While the question of how to represent uncertainty in RDF is still a matter of debate, we expect that reification will also be useful for making statements about the surety of statements.

Other requirements can be met by appropriating classes, properties, and syntax from other RDF vocabularies. The BIO `Event` class has a `date` property that is defined to be a sub-property of the Dublin Core `date` term. This means that we can use the DCMI Period Encoding Scheme (`http://dublincore.org/documents/dcmi-period/`) to specify dates for our events. The Period Encoding Scheme provides a very flexible way to specify the temporal range of both discrete and and continuous events at various levels of precision. Using BIO's `event` property we can link events to instances of the FOAF `Person` class. Though BIO does not have any native properties for linking events to artifacts, we could extend it to support this by adopting vocabulary from the ABC Ontology [8], declaring the BIO `Event` class and the ABC `Event` class to be equivalent, and using the various ABC properties for linking events and artifacts. In this way we could model life events that involve the creation of artifacts like works of art or manuscripts. Likewise, we can use the BIO `place` property to link events to instances of the Geonames `Feature` class. This allows us to specify the location where an event occurred at a number of different levels of precision, from countries to specific buildings, and to link these locations to geographical coordinates and boundaries. Finally, we borrow properties from the OWL-Time ontology (`http://www.w3.org/TR/owl-time/`) to make statements about the relative temporal ordering of events in cases where specific dates are unknown or unstated in sources.

Figure 1 demonstrates how we can combine multiple RDF vocabularies to model events in the life of American frontiersman and politician Davy Crockett. Here we have modeled five events based on assertions mined from a biographical sketch of Crockett that appears in an 1859 edition of the Dictionary of the

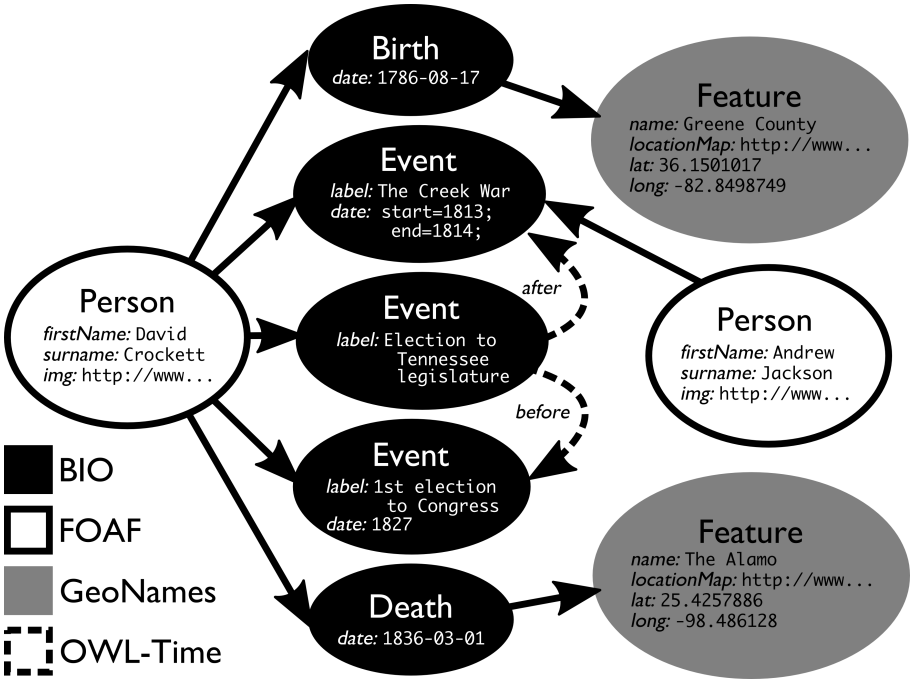

Fig. 1. RDF Combination of Schemas for Biographical Representation

United States Congress. They include Crockett's birth, his participation as a soldier in the Creek War, his election to the Tennesee state legislature, his first election to the U.S. Congress, and his death. The sketch gives exact dates and locations for Crockett's birth and death, which we have represented by using full date values and links to the Geonames feature records representing Greene County, Tennessee and The Alamo. However the date of Crockett's election to the Tennessee legislature is not specified: the sketch merely indicates that it occurred after he fought in the Creek War and before he was elected to Congress. We have faithfully modeled this using the OWL-Time properties `after` and `before`. Finally, the sketch mentions that he served under General Andrew Jackson in the Creek War, so we have linked the FOAF person record representing Jackson to the Creek War event as well. Thus we have created a network of records that would enable users to move from resources on Davy Crockett to resources on Greene County, The Creek War, Andrew Jackson, or The Alamo, or to see these events on a timeline, or to see these places on a map.

4 Conclusions

In the "Bringing Lives to Light" project we are focused on designing and developing support for the discovery and dynamic linking of contextual information to events in the lives of historic persons. Part of this effort is the development of

ways to represent events in the lives of individuals that will permit us to connect those lives with others (as with Crockett and Jackson above) as well as with particular places and periods. The use of RDF as a mechanism for integrating these event-oriented descriptions also permits us to easily take advantage of the rapidly developing set of tools being developed in the Semantic Web community. For example, we can use freely available tools from the Simile project at MIT (`http://simile.mit.edu`) to directly display on timelines and maps the RDF markup described in Section 3.

Another aspect of the project involves extraction of biographical events from digital texts. We are at work on developing NLP tools to identify life events in biographical texts and to then link those via search tools to related persons, places and events. In addition we have been converting structured and semi-structured biographical data (such as information extracted from the U.S. Congressional biography service at `http://bioguide.congress.gov`) to our RDF structure.

By extracting assertions linking people, places and events from prose passages, representing them in a structured form, and linking them to authoritative records with unique identifiers, we can go a long way toward addressing the problems of ambiguity that plague approaches relying solely on full-text search of humantities resources. Furthermore, rich enough representations can become valuable secondary resources in their own right, providing schematic overviews of the temporal and geographical context of people, places and events. Thus event-centric metadata is not only a powerful way to provide access to resources but is also a step towards new ways of using Digital Libraries.

5 Acknowledgments

The work presented draws on two projects supported by Institute of Museum and Library Services National Leadership Grants: Support for the Learner: What, Where, When, and Who (`http://ecai.org/imls2004/`) and Bringing Lives to Light: Biography in Context (`http://ecai.org/imls2006/`). Thanks also go to Vivian Petras, Jeanette Zerneke, and Kim Carl, and the efforts of Co-PIs Michael Buckland and Fredric C. Gey.

References

1. Ad Hoc Encoded Archival Context Working Group. Encoded archival context tag library (beta), 2004. http://www.iath.virginia.edu/saxon/servlet/SaxonServlet?source=/eac/documents/tl_beta.xml&style=/eac/shared/styles/tl.xsl.
2. J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.
3. N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff. Definition of the CIDOC object-oriented conceptual reference model and crossreference manual, 2003. http://cidoc.ics.forth.gr/docs/cidoc_crm_v_3.4.9_cross_referenced.zip.

4. F. Dawson and D. Stenerson. RFC 2445: Internet Calendaring and Scheduling Core Object Specification (iCalendar), 1998.
5. Family History Department. Gedcom xml specification release 6.0, 2001. http://www.familysearch.org/GEDCOM/GedXML60.pdf.
6. GENTECH. Genealogical data model phase 1: A comprehensive data model for genealogical research and analysis, 2000. http://xml.coverpages.org/GENTECH-DataModelV11.pdf.
7. International Press Telecommunications Council. EventsML G2 specification version 1.0, 2008. http://www.iptc.org/EventsML/.
8. C. Lagoze and J. Hunter. The ABC Ontology and Model. *Journal of Digital Information*, 2(2), 2001. http://jodi.tamu.edu/Articles/v02/i02/Lagoze/.
9. B. Robertson. HEML XML Schema Documentation, 2003. http://heml.mta.ca/heml-cocoon/schemadoc/2003-09-17/heml.
10. Text Encoding Initiative Consortium. Report on XML mark-up of biographical and prosopographical data, 2006. http://www.tei-c.org/Activities/PERS/persw02.xml.
11. U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE Multimedia*, 14(1):19–29, 2007.