Taking modeling seriously [in digital curation]

Foundational modeling challenges to foster future reuse of digital resources

Simone Sacchi, Karen M. Wickett
Graduate School of Library and Information Science
University of Illinois at Urbana–Champaign
501 E. Daniel Street, MC-493
Champaign, IL 61820-6211 USA
{sacchi1,wickett2}@illinois.edu

ABSTRACT

Interestingly (and unfortunately), conceptual data models developed within the digital preservation community are misused, and conceptual modeling techniques are underestimated in terms of how they can shape digital preservation, because the current modeling approach fails to answer the needs of the broader digital stewardship agenda. In general, we argue, we lack a unified digital curation model able to support curatorial activities that not only will fulfill the goal of "validating the communication from the past" but also support future reuse, or "communication with the future" [9]. Here we invite to take modeling seriously [in digital curation] and discuss the potential benefits in the long term.

Keywords

Modeling, digital preservation, digital curation, conceptual foundations

1. INTRODUCTION

Digital preservation takes place in the service of a broader stewardship agenda, one where digital resources must be curated not only to make their content accessible over time but also to enable their participation in ongoing research and scholarship, and proactively enable future, and possibly unanticipated, processing and analysis. The digital curation agenda —broadly interpreted as "maintaining and adding value to a trusted body of digital information for current and future use"— stress that "digital curation builds upon the underlying concepts of digital preservation whilst emphasising opportunities for added value and knowledge through annotation and continuing resource management"[1].

Interestingly (and unfortunately), conceptual data models developed within the digital preservation community [1, 14, 5], seem not to properly address some of these curatorial activities and goals. We believe that current models are misused, and conceptual modeling techniques are underestimated in terms of how they can shape digital preservation, because the current modeling approach fails to answer the needs of the broader digital stewardship agenda: it lacks the vision to enable and foster future, and possibly unanticipated, uses of digital resources. Here we focus on the OAIS Reference Model [1] as an example. Some foundational modeling aspects seem to fail:

1. *Information* —commonly understood as the ultimate target of preservation— and its possible *representations* (documents,data, records, etc.) are not modeled for what they *really* are and for their informative function. Instead, these models attend to the roles that certain entities play in a repository, confusing information-bearing entities with those roles. This limits the range of supported curatorial activities to those functionally consistent with those roles;

2. most conceptual entities are treated as black boxes, their logical structure left underspecified. This inhibits the shift from metadata description to semantic annotation, leaving most digital resources opaque to semantic–aware applications like those envisioned by the Semantic Web and Linked Data communities;

3. the explicit representation of some essential aspects are "outsourced" to other research agendas (e.g. provenance information). This inhibits the representation within the model of the entire lifecycle of digital resources, therefore limiting its characterization power.

In what follows we discuss these aspects in more detail.

2. WHAT IS MISUSED, MISSING OR CONFUSED

The OAIS Reference Model is routinely cited as an influential reference models in digital curation. While its functional and organizational components have influenced shaping preservation plans, its underlying Information Model —a conceptual data model— seem to only vaguely inform the development of preservation systems.

Most of the time the OAIS Information Model —along with other derived or compliant models [14]— seems to be used

[1] http://www.dcc.ac.uk/

just as a shared vocabulary useful for describing engineering solutions (e.g. format migration) to current preservation problems (e.g. technology obsolescence), in a myopic approach that does not (i) consider digital resources for what they *really* are and throughout their entire lifecycle, (ii) take advantage of the potential expressive power of conceptual modeling to support advanced digital curation activities.

2.1 Modeling information, and doing it consistently

The OAIS Information Model focuses on entities like Information Object while the digital curation agenda stresses the notion of digital information as its target. The notion of *Information Object*, for example, is extensively adopted in the the scientific literature and technical documentation. The OAIS Reference Model identifies an entire set of such objects, *Content Information* being one of them. Content Information is defined as "The set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Information"[1]. Similar definitions are applied to the other Information Objects in the model. To a certain extent, this entity is a useful modeling device: (i) it provides boundaries to otherwise loosely defined entities like "content" (ii) it allows for the assignment — and characterization — of attributes of digital resources (e.g. descriptive metadata) that seem to fit a general content–like or information–like entity. However, two main problems seem to emerge:

1. while Information Objects are claimed to denote *information*, most likely they denote *representations* of that information in a particular symbolic form (e.g. records, documents, data, etc.). Information itself is not modeled as a first class entity in this model.

2. It is a dramatic ontological mistake to confuse *information* —or any representation of information— with the set of things (e.g. a Data Object and its Representation Information in OAIS) that participate in granting access to it. If we take modeling seriously [11] information or symbolic representations of information —notwithstanding the variation in their definitions— can be hardly conceptualized as something "composed" of its Data Object and its Representation Information, *pace* the OAIS Reference Model.

The OAIS entities only model functions within preservation repositories —namely, in this case, aggregation of the requirements for interpreting bitstreams — and not what they are claimed to denote —namely, *information*.

Modeling digital resources for what they really are[2] —and within an ontologically consistent model— has the advantage of:

1. Avoiding conflation and confusion of entities that generate category mistakes in the assignment of properties to things. This is an essential requirement to support a correct identification of the preservation target. Are we aiming at preserving information regardless of any specific representation? Or is the goal to preserve particular features (significant property) that reside at the level of a specific representation?

2. Abstracting from current implementations, and allowing support for future curatorial activities. Precisely identifying the *things* that participate in the representation of information in digital form supports modeling multiple roles and functions they acquire in specific contexts, environments or [curatorial] activities.

2.2 Unpacking black boxes and modeling representations

When curators use the term "Information Object" they seem to refer to *representations* of information (documents, data, records, etc). These are symbol structures that express specific information in virtue of an assignment of meanings to specific tokens[3]. These *logical components*, if identified and modeled, provide "anchors" for finer–grained descriptions. However, entities like Information Object are modeled as atomic units: they are black boxes masking the internal structure of their logical components. Two main reasons inhibits further refinements:

1. The compositionality of this entity, is *actually* expressed: an Information Object is claimed to be *composed* of a Data Object and its Representation Information.However this is a fictive aggregation, not reflecting a logical structure ontologically consistent with what representaions of information realy are.

2. While it seems possible to "nest" Information Objects within other Information Objects creating a sort of hierarchical structure —and in fact Representation Information *is* itself an Information Object in OAIS — the logic of such compositionality is not specified.

This is a major limitation in their expressive power if these models must support certain semantic–oriented curatorial activities (e.g. semantic annotation [6]) and add "value and knowledge" to digital resources. This goal is usually achieved by breaking down symbol structures into logical units of descriptions in order to annotate the marked–up components accordingly to a domain ontology. The entire Semantic Web and Linked Data research agendas rely on such a modeling approach, one able to make explicit and computationally available the semantics documents and data. "Significance is in the eye of [machine] stakeholders"[2].

Of course, this is an active component of other —still related— research agendas, in particular the ones proposed by the descriptive markup, ontology development, data semantics, and annotation communities[4]. However, their modeling strategies are not directly aligned or merged with those developed in digital preservation, slowing the process of curating digital resources for consumption by both human beings and semantic–aware applications.

[2]Some preliminary suggestions can be found in [15]

[3]This generally applies to text–based digital resources, however for images, audios, and videos, the identification of tokens is more challenging.

[4]Here some examples of their achievements [8, 12, 7]

2.3 Provenance and the lifecycle of digital resources

The explicit representation of some essential aspects digital resources' lifecycle are "outsourced" to other research agendas, an example being their provenance information. Modeling provenance involves representing *things* and *events* participating in the lifecycle of digital resources that transcend preservation transactions within repositories. Among these events we enlist the indication events [13, 3] involved in the creation of a digital resource and all the transaction (history of creation, ownership, accesses and changes) bringing a resource to its current state. The Open Provenance Model[10] and the W3C PROV Ontology[5] model such information. The assessment of authenticity and trust are essential for digital stewardship, and a correct modeling and representation of provenance information is a core component for such assessment.

Modeling provenance explicitly within preservation models[6] —or at least make preservation models provenance–aware through extension points that support model alignment while preserving the modularity paradigm— foster their expressive power and provides a more consistent and unified model for digital resources lifecycle.

3. CONCLUSION

We discussed aspects of what we call conceptual and modeling foundations of digital curation, focusing on modeling strategies to foster future reuse of digital resources. In general, we argue, we lack a unified digital curation model — one that is ontologically robust, extensible and is informed by precisely–defined conceptual and modeling foundations— able to support curatorial activities that not only will fulfill the goal of "validating the communication from the past" but also support future reuse, or "communication with the future" [9], and, in general, the broader vision of digital stewardship.

We are aware that taking modeling seriously is a time-consuming activity. We are hoping to start a conversation about what the potential benefits are in the long term, not to dictate any particular approach.

4. ACKNOWLEDGMENTS

The authors wish to thank Allen Renear and David Dubin for their continuous help and invaluable suggestions. The research reported here has being carried out at the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois at Urbana-Champaign and funded by the National Science Foundation as part of the Data Conservancy (OCI/ITR-DataNet 0830976).

5. REFERENCES

[1] J. CCSDS. Reference model for an open archival information system (OAIS). Technical report, CCSDS 650.0-B-1, Blue Book, 2002.

[2] A. Dappert and A. Farquhar. Significance is in the eye of the stakeholder. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries*, ECDL'09, pages 297–308, Berlin, Heidelberg, 2009. Springer-Verlag. ACM ID: 1812838.

[3] D. Dubin. Encoded descriptions at face value. In A. Grove, editor, *Proceedings of the American Society for Information Science and Techonology*, volume 47 of *ASIS&T Annual Meeting Proceedings*, Pittsburgh, PA, Oct. 2010.

[4] M. Factor, E. Henis, D. Naor, S. Rabinovici-Cohen, P. Reshef, S. Ronen, G. Michetti, and M. Guercio. Authenticity and provenance in long term digital preservation: modeling and implementation in preservation aware storage. In *First workshop on on Theory and practice of provenance*, page 6, 2009.

[5] H. Heslop, S. Davis, A. Wilson, and N. A. o. Australia. *An approach to the preservation of digital records*. National Archives of Australia Canberra, 2002.

[6] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[7] L. Lefort, C. Henson, K. Taylor, P. Barnaghi, M. Compton, O. Corcho, R. Garcia-Castro, J. Graybeal, A. Herzog, K. Janowicz, et al. Semantic sensor network XG final report, W3C incubator group report (2011). Technical report, W3C, June 2011.

[8] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279 – 296, 2007.

[9] R. Moore. Towards a theory of digital preservation. *International Journal of Digital Curation*, 3(1), 2008.

[10] L. Moreau, B. Plale, S. Miles, C. Goble, P. Missier, R. Barga, Y. Simmhan, J. Futrelle, R. E. McGrath, J. Myers, P. Paulson, S. Bowers, B. Ludaescher, N. Kwasnikowska, J. V. d. Bussche, Ellkvist, Tommy, J. Freire, and P. Groth. The open provenance model (v1.01). Technical report, University of Southampton, July 2008.

[11] A. H. Renear. Taking modeling seriously. In *Knowledge Organization and Data Modeling in the Humanities*, volume 47, Providence, RI, Mar. 2012.

[12] R. Sanderson and H. V. d. Sompel. *Open Annotation Collaboration: Alpha Data Model Summary*. 2009. Published: Published by the Open Annotation Collaboration at http://www.openannotation.org/documents/OAC-Model_UseCases-alpha.pdf.

[13] B. Sandore and J. Unsworth. ECHO DEPository Ñ phase 2: 2008-2010 final report of project activities. pages 30–37. University of Illinois at Urbana-Champaign, June 2010.

[14] R. Sharpe. PLANETS data model overview. Technical Report IF8-D1, 2009. please request from info@planets-project.eu.

[15] K. M. Wickett, S. Sacchi, D. S. Dubin, and A. H. Renear. Identifying content and levels of representation in scientific data. In *To be published in: Proceedings of ASIS&T 2012: the 75th Annual Meeting of the American Society for Information Science and Technology*, volume 48, Baltimore, MD, 2012.

[5] http://www.w3.org/TR/prov-o/

[6] An example of such an approach can be found here [4]