DRAFT: The Basic Representation Model for Digital Preservation and Information Organization
This paper draft is being shared as part of the SIG-CM ASIST 2020 Workshop. Please DO NOT DISTRIBUTE THIS DRAFT

Karen M. Wickett

Contents

Abstract

Information models to handle preservation and information organization, such as the Functional Requirements for Bibliographic Records, the OAIS Reference Model, and the CIDOC Conceptual Reference Model have faced major stumbling blocks in their adoption, due in part to the difficulty in applying them consistently to digital objects. This paper describes and analyzes those challenges, and presents The Basic Representation Model, a general framework to account for the multiple and varied levels of representation that occur in the creation, management, and preservation of digital objects. This model was developed in the context of a project focused on the preservation of scientific data, and has the additional potential to support fine-grained information modeling in bibliographic, scientific, and cultural heritage domains. In particular, the model can support clear and explicit connections between information organization, information management, and digital preservation.

1 Introduction

[Section goals: The potentials for connecting information organization, information management, and digital preservation. My goal with the paper is to provide a general model of the objects and relationships that are fundamental to the representation and transmission of information. My goal with this section is to motivate the need for such models. I think what is missing in many models is a flexibility to handle the multi-level realization and embodiment relationships that typically occur in digital systems.]

2 Background

My goal in this section is to sketch the intellectual history of the Basic Representation Model and review and critique the related models that were the primary inspiration for BRM.

2.1 General Theories of information

The development of conceptual models to explain the entities and relationships involved in the recording and transmission of information is an important activity for the field of library and information science. These models inform system design, guide research, and provide insight into the nature of information and information-bearing objects. The Basic Representation Model bridges the gap between general models of the what information is, and the encoding and transmission of information into documents and digital objects.

At their most general, theories of information are oriented around communication and the transmission of messages that carry content. For example, the Mathematical Theory of Communication (Shannon and Weaver, 1964) views information content as quantifiable in terms of the reduction in uncertainty given by the reception of a signal. While this model encapsulates an engineering approach to building robust channels for signal transmission and does not accord well with the realities of human communication (Ma, 2012), it highlights the fact that information consists in some sense of content.

The centrality of information to human communication has driven many philosophical explorations of the topic. The theories developed by Barwise and Perry (1983), Dretske (1981), and Devlin (1995) characterize meaning and information as highly contextualized phenomena. Information-bearing messages have meaningful content only and because of human action - in expression of content and in the interpretation of signals. Floridi brings together the philosophical and engineering approaches in his General Definition of Information, which treats information as consisting of some quantifiable amount of data, which is well-formed by some syntax and meaningful under some interpretation (Floridi, 2013). Floridi's and other philosophical accounts are explanatory for understanding the basic nature of information and its fundamental role in human communication, but do not connect to the management of information as approached in library and information studies.

Bates's "Fundamental Forms of Information" takes a similarly generalist approach and presents a wide-ranging analysis of information in its many forms. Bates advocates for a suite of general definitions for information, with her most general account defining information as a "pattern of organization of matter and energy" (Bates, 2006). The general account is divided into several classes of information, with the most relevant for the work discussed here being her definition of Recorded Information as "Communicatory or memorial information preserved in a durable medium." The management of records and the information they contain is a core activity for library and information science, and The Basic Representation Model described below is concerned with information in this sense.

In "Information as Thing", Michael Buckland presents a view of the information field that analyzes the work, outputs and perspectives on information in terms of "information as thing", "information as process" and "information as knowledge" (Buckland, 1991). Buckland points out that "Information-as-thing is of special interest in the study of information systems. It is with information in this sense that information systems deal directly." The conceptual models of information-bearing objects (discussed below) that have been developed to promote interoperability and system-building since them all confirm Buckland's stance on information (as thing) by identifying entities and relationships pertaining to the expression, encoding and storage of information in some concrete form.

One of the primary entities identified by Buckland as a focal point for the analysis of information systems the document. Although the precise nature of documents has also been argued to be contextual (citep Buckland 1997), Elaine Svenonius grounds her analysis of information organization systems in the notion of the document, defined as "an information-bearing message in recorded form" (Svenonius, 2000). Svenonius then goes on to define and analyze information organization systems in terms of set operations on documents as determined by various features. Svenonius' account is explanatory of the purposes and activities that are fundamental to information organization, but does not analyze documents as bearing information in recorded form.

[TRANSITION]

2.2 Related models

2.2.1 FRBR and related models

The Functional Requirements for Bibliographic Records (FRBR) was produced by IFLA in 1998 and has been a focal point in the evolution of bibliographic cataloging since then (IFLA, 2009). Many folks have thought about how to apply it to digital objects in general (Renear and Dubin, 2007), video games and virtual worlds (McDonough et al., 2010), scientific data (Hourclé, 2008) (Sacchi et al., 2011b) (Wickett et al., 2012).

FRBR Group One entities and relationships. Since FRBR presents these in pretty agnostic terms, it tempting to apply the relationships outside the scope of monographic publishing. But it's not clear that FRBR even works that well for tracking the entities and relationships involved in, for example, the publication of XML documents, or e-books. Certainly problems arise when you try to apply it to scientific data, see (Wickett et al., 2012).

From Wickett et al. (2012) :

> Another problem with FRBR is that its entity types appear to not represent fundamental types of things, but rather roles that fundamental things enter into in particular circumstances ((Renear and Dubin, 2007); (Guarino and Welty, 2000)). This makes it hard to identify what features are contingent properties and what features apply to fundamental types, as well as making extension and refinement of the model convoluted.

2.2.2 OASIS Reference Model

So, OAIS is really important and influential for digital preservation systems (CCSDS, 2002). Everybody be making their SIPs and their DIPs. And it seems to work really well for the one level of representation it works at, which is files. But then it also gives a few models of information objects and content that are not easy to understand and don't work as general models. As Simone put it, One Thing is Missing or Two things are confused (Sacchi et al., 2011a).

An incomplete list of important concepts to explain and critique from OAIS: information object, data object, representation information, designated community.

2.2.3 CIDOC-CRM

CIDOC-CRM is an ontology for descriptions of cultural heritage objects. I'm including it here because it aims to give a high level ontology that museum descriptions can be mapped to to enable interoperability. I see a similar potential role for BRM but with scientific data. CIDOC is very oriented around cultural heritage and museum object description practices.

This makes me wonder though, about how BRM maps to CRM. Also can BRM serve as a sort intermediate layer that would enable mapping of scientific data to various scientific ontologies, like SWEET, etc. This is one way were I see potential for BRM to serve as a bridge between digital preservation and information organization. BRM accounts for both semantic content and the encoding of digital information, so it is possible to describe and link those pieces of information. The encoding layers and mappings between symbol structures and PME are information that is essential to digital preservation. The semantic content is the level where information organization activities such as indexing and knowledge organization take place.

2.3 Data Conservancy Data Concepts Group

The Data Conservancy Data Concepts (DCDC) group was a research group hosted at the Center for Informatics Research in Science and Scholarship at the University of Illinois at Urbana-Champaign, as part of the NSF Data Conservancy project. The aims of the research group were to clarify key concepts in the encoding and representation of scientific research data. As part of this work, the group worked to identify appropriate models enable description and preservation of digital datasets that were the result of scientific data collection, observation, analysis, and aggregation processes.

These mapping efforts focused on popular models from library and information science and resulted in a series of papers (Sacchi et al. (2011b), Dubin et al. (2011), Wickett et al. (2012)). The group proposed the Systematic Assertion Model (SAM), a conceptual model for scientific data that positioned the content of scientific data as propositional content warranted by an observational or computational event, and data as the symbolic representation of that data content.

The Basic Representation Model provides the representational backbone for the Systematic Assertion Model. SAM is an event-based model that accounts for the agents and events that essential to the origination of scientific data. BRM provides the entities and relationships that those agents and events act on or produce. The popular models that the group reviewed in Sacchi et al. (2011b) all had big problems for us, which I'll tell you about in the next subsection. They didn't have enough fine-grained divisions, or they seemed to conflate important objects, relationships, or activities.

The entities and relationships in BRM were first named and published on by Dave Dubin as part of Preservation Model 1, which was a result of the EchoDep project (Sandore and Unsworth, 2010).

2.4 A motivating example

I think a worked example would be useful for readers. Wickett et al. (2012) uses an example of a species occurrence record. I would like to develop an original example for this paper. I would appreciate feedback on what features might make for a useful example in this context.

This is where I would do the initial presentation of a motivating example. Then I would discuss the example through the presentation of the model.

Example ideas for workshop feedback:

- Enron email archive and associated datasets: https://www.cs.cmu.edu/ ./enron/ . This would let me discuss the re-representation of the archive in ways that are targeted at particular kinds of analysis - e.g. social network analysis versus linguistic analysis. There have been many versions produced. This example may be TOO FRUITFUL.

- Another email archive - Don Knuth's email collection. Note that "The full text of the emails is only accessible on a workstation in the Field Reading Room, which is open to all members of the general public.“

- a digital repository example -

- a scientific data example - part of the point of this paper is to argue for this as a general model, not specific to scientific data. but this would probably be fastest for me to write.

3 The Basic Representation Model

3.1 BRM Entities

One of the achievements of the Basic Representation Model is that it identifies the genuine entity *types* (following Guarino and Welty (2000)) involved in the recording and expression of semantic content, and distinguishes those from the contingent *roles* those entities play in particular contexts. As a criterion for distinguishing types from roles one can adapt Guarino and Welty and apply this rule: If it is possible that something that is an F might not have been an F, then being an F is a role that things have; otherwise F

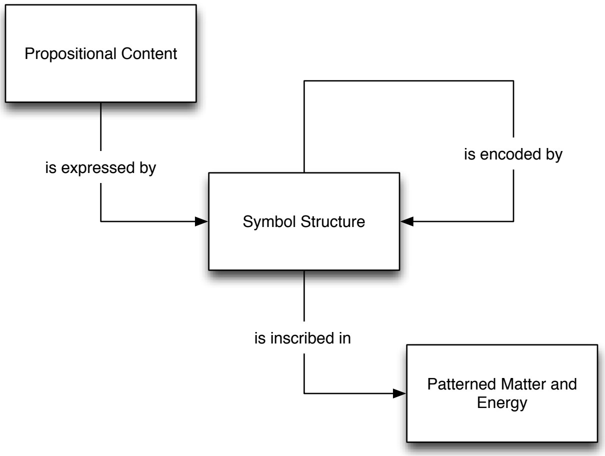

Figure 1: The Basic Representation Model

is a type of thing. So, using their example, since it is possible that someone who is student might not have been a student (i.e., might not have enrolled this year), *student* is role. But since it is not possible that something that is a person might not have been a person (and still exist), *person* is a type of thing.

Taking this strategy, the Basic Representation Model proposes three fundamental types of things that participate in the representation of semantic content as information, be it in the form of digital objects or in more traditional forms such as printed text. These entity types are *Propositional Content*, *Symbol Structure*, and *Patterned Matter and Energy*.

Propositional content. Allen Renear likes to sometimes define propositions as the bearers of truth values. Propositional content is the kind of thing (and the only kind of thing) that can be true or false. But the concern here is not with whether those propositions are true or false, or whether they pertain to our actual world. Just that they are purely content. This maps closely to Floridi's concept of the semantic content in his definition of information as semantic content.

Text from Wickett et al. (2012):

> In our model propositions appear as the language–independent content expressed by symbol structures. In the sense intended propositions may be defined as all and only those things that are either possibly true or possibly false. That is, they are the proper subjects of truth values. The symbol structure that expresses a proposition may also be considered true or false, but only in a derivative sense: derivatively "true" if the proposition it expresses is true, and derivatively "false" if the proposition it expresses is false. A common alternative account of propositions defines them

> as the proper objects of epistemic attitudes, such as belief or doubt. For our purposes these two accounts of proposition may be considered co-extensive: the class of things that can be true or false is identical with the class of things that can be the object of epistemic attitudes. Although the significant role of propositions in our model is as the expressed content of symbol structures, the definitions just given allow propositions to exist independently of symbol structures.

Symbol Structure. These are arrangements of elements into discernable patterns.

Text from Wickett et al. (2012):

> In our model symbol structures are abstract arrangements of symbols that, in a given context, express propositions. Individual symbols themselves are the atomic components of symbol structures. Although the symbol structures in our examples are in some language with a determinate semantics, our model allows symbols and symbol structures to express different propositions in different languages or different contexts. Examples of abstract objects that can serve as symbol structures include graphs, relations, and sequences, along with more familiar kinds of symbol structures like strings of characters.

Patterned matter and energy. We are living in a material world, and to encounter information we must encounter some material objects that record it in a way we can interpret. The name for this entity type takes inspiration from Marcia Bates' Fundamental Forms of Information, but separates the matter or energy that is patterned from the patterning, since the patterning is an abstract arrangement of some physical material.

Text from Wickett et al. (2012):

> Whereas both propositions and symbol structures are abstract objects, patterned matter and energy is a concrete quantity of matter and energy that manifests a physical arrangement that is the physical inscription of an (abstract) symbol structure. In order for a digital object to effectively communicate information, there must be some instantiation of the symbol structures in a physical medium that an agent can interact with.

3.2 BRM Relationships

Is Expressed By. This seems to be lifted from FRBR. This relationship can stand between propositional content and symbol structures. The symbol structure that stands in this relationship is a *primary symbol structure.* The distinction between primary and non-primary symbol structures was important for identifying data content.

Text from Wickett et al. (2012):

> Every meaningful digital object will use symbol structures to express propositions. For instance, a digital object may use RDF triples to express propositions about species occurrence. We use the *is Expressed By* relationship type for this technical sense of “express”. The *is Expressed By* relationship type represents the fact that the propositional content of a digital object is understood as being expressed by a symbol structure that is the primary expression — the *Primary Symbol Structure* — for that content in a particular context. *is Expressed By* represents a general relationship that is instantiated between specific propositional content and a specific symbol structure. An event–based account of how this relationship is actually instantiated for scientific data is provided by the Systematic Assertion Model.

Is Encoded By. this is a relationship between symbol structures. It can describe encoding relationships in digital systems. It is the possibility of recording information about relationships between symbol structures in the expression and encoding of information that is particularly useful for modeling information in digital computing systems. These layers of symbol structures in encoding relationships are key to the recording and storage of information in computational systems.

The layers of encoding and representation in a digital system are modeled by a series of *is Encoded By* relationships between *Symbol Structures*.

Text from Wickett et al. (2012):

> A digital object will typically map the symbol structures that express propositions into other symbol structures. We call this mapping from symbol structure to symbol structure an encoding of one symbol structure by (or into) another. For instance, a digital object may map RDF triples into the XML/RDF serialization language. Or those same triples might be encoded in the N3 serialization language. In each case we have the same Primary Symbol Structure – the RDF triples that express propositional content – but a different encoding of that primary symbol structure. Symbol structures that are encodings of other symbol structures may in turn be encoded by still other symbol structures. For instance the N3 symbol structure may itself be encoded in a UTF-8 byte sequence. Unpacking the encoding levels provides a more complete and consistent way to represent what changes when digital objects undergo transformations, like format migrations.

Is Inscribed In. This is the relationship between a symbol structure that expresses content and a concrete arrangement of matter or energy.

Text from Wickett et al. (2012):

> The Is Inscribed In relationship type represents the fact that a particular symbol structure is represented in a physical medium through a mapping between the symbol structure and a particular concrete arrangement of matter and energy.

3.3 Interpretive Frames

Text from Dubin et al. (2011):

> But the digital data resources that concern us are encoded symbol structures that express data content. Our problem is the contingent nature of this connection: data express their conceptual content not simply in virtue of their arrangement and structure, but always with reference to what we call *interpretive frames.* These are abstractions that frame the interpretive context for symbolic expressions.
>
> At the risk of understating their complexity, one can think of interpretive frames as functions or mappings between structural propositions at different expressive levels, or from structural propositions to conceptual propositions. Examples of interpretive frames include the grammatical rules expressed by an XML Schema, coded character sets such as ACSCII, the convention of writing numbers as strings of Arabic numerals with ten as the implied numerical base, the Hierarchical Data Format standard, and all dialects of the English language as they are spoken today. Interpretive frames alos include any systematic expressive choices that may be local to a particular digital resource, such as a correspondence between successive rows of a spreadsheet and the order of transactions in a scientific experiment.

4 Discussion

4.1 Levels of representation

Information models that assume a single level (or a fixed set of levels) of representation in digital objects are always going to suck to apply. I'm not sure this needs to be a separate section, really, or whether it might just be most effective to critique the representation of digital objects by the various models in the background section.

Note: discuss YT project?

4.2 Application Areas

[Here's a big question mark for me. There are not any real deployments of this model to discuss here. So is my goal here to argue that people should try to do that? Or is the model really better at explanation than for operationalization? Is the lack of deployment purely because we are not developers and haven't had the opportunity to build a system that works this way? Or is the model too high level and works better as a teaching tool? I would appreciate feedback on this point.]

4.2.1 Scientific data management and preservation

Ideas for application scenarios to discuss:

- instrument data
- astronomical data
- biodiversity records

4.2.2 Bibliographic data and Text encoding management

- Jacob and Dave's case study from Balisage 2018.
- MARC records into XML into linked open data

5 Future work

- Integration with provenance models

6 Conclusion

References

J. Barwise and J. Perry. *Situations and Attitudes.* MIT Press, 1983.

Marcia J. Bates. Fundamental forms of information. *Journal of the American Society for Information Science and Technology*, 57(8):1033–1045, 2006. doi: 10.1002/asi.20369. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20369`.

Michael K Buckland. Information as thing. *Journal of the American Society for information science*, 42(5): 351–360, 1991.

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

Keith Devlin. *Logic and information.* Cambridge University Press, 1995.

F.I. Dretske. *Knowledge and the Flow of Information.* Blackwell, 1981.

David Dubin, Karen M. Wickett, and Simone Sacchi. Content, format, and interpretation. *Proceedings of Balisage: The Markup Conference 2011*, 2011. doi: 10.4242/balisagevol7.dubin01. URL `http://dx.doi.org/10.4242/BalisageVol7.Dubin01`.

Luciano Floridi. *The philosophy of information.* OUP Oxford, 2013.

Nicola Guarino and Christopher A. Welty. A formal ontology of properties. In *EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, pages 97–112, London, UK, 2000. Springer-Verlag. ISBN 3-540-41119-4.

Joseph A. Hourclé. FRBR applied to scientific data. In *Proceedings of the American Society for Information Science and Techonology*, volume 45, pages 1–4, 2008. doi: 10.1002/meet.2008.14504503102. URL `http://dx.doi.org/10.1002/meet.2008.14504503102`.

IFLA. Functional requirements for bibliographic records: Final report. Technical report, International Federation of Library Associations and Institutions, 2009. URL `http://www.ifla.org/files/cataloguing/frbr/frbr`$_2$008.*pdf*.

Lai Ma. Meanings of information: The assumptions and research consequences of three foundational lis theories. *Journal of the American Society for Information Science and Technology*, 63(4):716–723, 2012.

Jerome McDonough, Matthew Kirschenbaum, Doug Reside, Neil Fraistat, and Dennis Jerz. Twisty little passages almost all alike: Applying the frbr model to a classic computer game. *Digital Humanities Quarterly*, 4(2):1869–1883, 2010.

A. H Renear and D. Dubin. Three of the four FRBR group 1 entity types are roles, not types. *Proceedings of the American Society for Information Science and Technology*, 44(1):1–19, 2007.

Simone Sacchi, Karen M. Wickett, Allen H. Renear, and David Dubin. One thing is missing or two things are confused: An analysis of OAIS Representation Information. *Poster presented at the Seventh Internation Digital Curation Conference*, 2011a.

Simone Sacchi, Karen M. Wickett, Allen H. Renear, and David S. Dubin. A framework for applying the concept of significant properties to datasets. In *Proceedings of the American Society for Information Science and Techonology*, New Orleans, LA, 2011b.

Beth Sandore and John Unsworth. *ECHO DEPository — Phase 2: 2008–2010 Final Report of Project Activities*, section 4.2, pages 30–37. University of Illinois at Urbana-Champaign, June 2010.

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* The University of Illinois Press, 1964.

Elaine Svenonius. *The intellectual foundation of information organization.* MIT press, 2000.

Karen M Wickett, Simone Sacchi, David Dubin, and Allen H Renear. Identifying content and levels of representation in scientific data. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012.