Balisage: The Markup Conference 2011
Proceedings

Content, Format, and Interpretation

David Dubin
Research Associate Professor
University of Illinois
<ddubin@illinois.edu>

Karen Wickett
University of Illinois
<wickett2@illinois.edu>

Simone Sacchi
University of Illinois
<sacchi1@illinois.edu>

Balisage: The Markup Conference 2011
August 2 - 5, 2011

How to cite this paper

Dubin, David, Karen M. Wickett and Simone Sacchi. "Content, Format, and Interpretation." Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2 - 5, 2011. In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, vol. 7 (2011).
doi:10.4242/BalisageVol7.Dubin01.

Abstract

The connection between notation and the content it expresses is always contingent, and mediated through complex layers of interpretation. Some content bears directly on the encoder's intention to convey a particular meaning, while other content concerns the structures in and through which that meaning is expressed and organized. *Interpretive frames* are abstractions that serve as context for symbolic expressions. They form a backdrop of dependencies for data management and preservation strategies. Situation semantics offers a theoretical grounding for interpretive frames that integrates them into a general theory of communication through markup and other notational structures.

Table of Contents

Acknowledgments

Introduction

The distinction between a digital resource's content and its expressive format is usually described in different terms than the content/presentation distinction familiar to markup researchers and practitioners. In both cases one understands that the same content can be formatted or presented in different ways. But the word “format” typically connotes a discrete symbolic notation—one that might encode conceptual content, structural information, presentational instructions, or all three. “Presentation” is usually understood as patterns of energy or matter that visually or audibly communicate (via shared graphical or auditory interpretive conventions) resource structure and content to human minds. Standardized and proprietary digital file formats are the most familiar of these notations.

Proposals for semantic enrichment or digital preservation often focus on methods for transforming resources from one format into another. Colloquial XML can be transformed into RDF via XSLT Sperberg-McQueen and Miller, 2004, or into horn clause assertions through a Prolog application Dubin, 2003. But although notations like RDF and first order logic may admit more expressive distinctions than colloquial XML, such transformations at best merely re-express resource semantics in a more convenient form for drawing inferences or some other purpose“those semantics aren't inherent in the notation. The connection between resource and content (i.e., a symbol structure and the content it expresses) is always contingent: the same symbols might just as easily express different content, or no content at all Renear and Dubin, 2007. In the context of some particular assertion event, correct interpretation of encoded content is typically mediated through many expressive layers. In the following sections, we discuss the relationships among content, structure, and presentation, and situate them with respect to our ongoing research in scientific data management.

Background

This work is part of the Data Conservancy, an ongoing scientific data management project funded by the National Science Foundation's Office of Cyberinfrastructure Choudhury and Hanisch, 2009. Our aims are to develop formal terminology and identity conditions for concepts of general importance to the management and use of scientific datasets (e.g., observation, data content, version, format, etc.). Our proposed formalizations are expressed as terminological axioms in the Description Logic ALC Schmidt-Schauß and Smolka, 1991 Baader et al., 2003. Although these may later base ontologies that can direct automated reasoning over data set descriptions, our current aims are merely analytic: we propose, challenge, and revise the models in the context of reviewing and informing data curation practices and system design decisions. For example, we suggest that a model separating abstract propositional content of a scientific assertion from the observation event justifying that assertion may ease data integration across a series of related studies (e.g., replication of findings):

Figure 1

- **Equation (a)**
  $\text{Proposition} \sqsubseteq \text{AbstractThing}$

- **Equation (b)**
  SimpleProposition ⊑ Proposition
- **Equation (c)**
  ComplexProposition ⊑ Proposition
- **Equation (d)**
  Conjunction ⊑ ComplexProposition
- **Equation (e)**
  Observation ⊑ Event
- **Equation (f)**
  Computation ⊑ Event
- **Equation (g)**
  Assertion ⊑ Event
- **Equation (h)**
  SystematicAssertion ≡ Assertion ⊓ ∃warrantedBy.(Observation ⊔ Computation)
- **Equation (i)**
  (Proposition ⊓ ∃substanceOf.SystematicAssertion) ⊑ DataContent
- **Equation (j)**
  (Proposition ⊓ ∃conjunctOf.DataContent) ⊑ DataContent

The reader is invited to imagine simple propositions as standing (as reified RDF statements do) in subject, predicate, and object relations to entities and properties in a scientific domain like chemistry or ecology. But unlike reified RDF, our simple propositions are completely abstract, requiring no concrete expression. Hayes, 2004. Propositions standing in the same subject, predicate, and object relations are strictly identical. On this understanding, different data sets might have exactly the same propositional content, but differ in the observations or computations that justify their assertions. Similarly, two scientists might appeal to exactly the same observation events as justification for very different (or even contradictory) assertions.

On the content of digital resources

In the context of our research on scientific data, we view resource "content" as propositional in nature. A *proposition* is an abstract thing which can be the object of propositional attitudes (such as belief or doubt) and the bearer of truth values. We consider propositions to be the language independent entities that are the meanings of those sentences (or other symbol structures) that express them. Artistic and literary resources may have forms of non-propositional content that are inseparable from the expressive choices of their creators, but artistic and literary content are not our focus in this study.

Specifically, we are concerned with two kinds of propositional content:

Conceptual Content	Conceptual content is the distinct intellectual contribution supplied by the digital resource, which in our study concerns entities, properties, and relations in a scientific domain. This type of content corresponds, roughly, to the "work" entity type in the FRBR model IFLA, 1998, or, with a slightly different connotation, the "Deliverable Unit" in the PLANETS model Sharpe, 2009. Conceptual content is typically considered the main preservation target, though on our account such content, being abstract, is not subject to corruption and so isn't literally preserved.
Structural Content	The second kind of propositional content concerns abstract structures in and through which conceptual content is expressed and organized. The paragraphs, chapters, and footnotes of conventional documentation are among these structures, as well as database relations, spreadsheet rows, and lines and arcs of vector graphics. Examples of structural content would include the fact that a particular text string is a paragraph, or that an arc has particular coordinates in an abstract display plane.

The digital data resources that concern us are encoded symbol structures that express scientists' claims, with our analysis aimed at supporting format migration, digital preservation and data integration. Abstract symbol structures and propositions do not undergo changes of state Renear and Wickett, 2009, and so the problem is one of maintaining a connection between conceptual content and the structures that express it. This is easier when structural content is directly encoded within a digital resource as, for example, with XML declarations, PostScript prologues, and other forms of metadata. In the following sections we consider the connections between the propositions expressed through these technologies, and the chain that links the bit level to the conceptual level.

Data Expression and Interpretive Frames

By the account in the earlier section, data content are a subset of abstract propositions, obtaining their status in virtue of their systematic assertion by a researcher. 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:

Figure 2

- **Equation (k)**
  $\text{SymbolStructure} \sqsubseteq \text{AbstractThing}$
- **Equation (l)**
  $\text{InterpretiveFrame} \equiv (\text{AbstractThing} \sqcap \exists \text{interpretiveContextFor.SystematicAssertion})$
- **Equation (m)**
  $\text{Data} \equiv \text{SymbolStructure} \sqcap \exists \text{primaryExpressionFor.SystematicAssertion})$

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 ASCII, 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 also 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.

In pointing to contingent interpretations as “our problem,” we don't mean that to suggest encoding standards, markup technologies, or even common data management practices are seriously flawed. While we're motivated by practical problems, such as under-documented spreadsheets, in highlighting the complexities of interpretation we don't mean to suggest that effective tools and solutions are lacking. But discussions of these methods tend to foreground regularity in a resource's primary expressive structure, and neglect the interrelationships among interpretive frames at different levels of abstraction.

Working Example

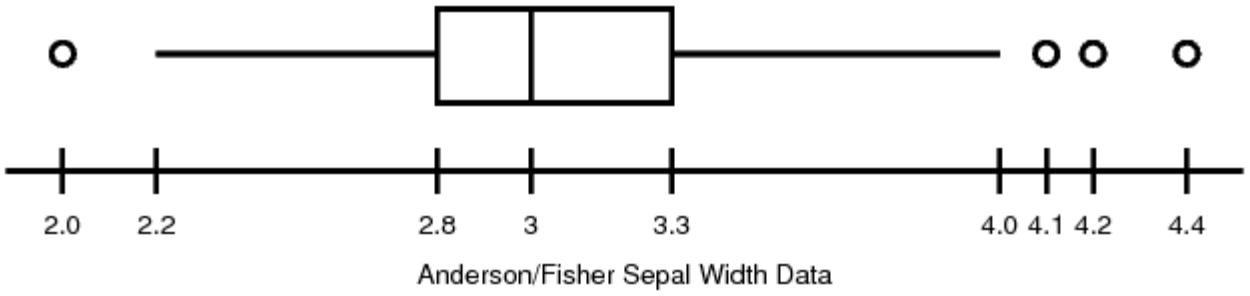
The following digital image can serve as an example of the distinctions we wish to draw. The resource Fisher5 is an Encapsulated PostScript file Adobe Systems, 1990 Its prologue consists of reusable functions, written to draw box-and-whisker plots from frequency distribution parameters. The final lines of the file lay out the parameters for the single plot:

```
/outliers 1 2.0 1 4.2 1 4.4 1 4.1 4
/left 2.2 /loq 2.8 /med 3 /upq 3.3 /right 4.0 /min 2.0 /max 4.4
/label (Anderson/Fisher Sepal Width Data) box

showpage
```

Displayed in an appropriate document viewing application, the file's presentation looks like this:

Figure 3



The following propositions comprise Fisher5's conceptual content:

- A certain frequency distribution is called “Anderson/Fisher Sepal Width Data.”
- The minimum value of that distribution is 2.0.
- The maximum value is 4.4.
- The median of that distribution is 3.
- The upper and lower hinges are 3.3 and 2.8, respectively.
- The distribution has four outliers, one each at values 2.0, 4.2, 4.1 and 4.4.
- 2.2 and 4.0 are (respectively) the lowest and highest values that lie within 1.5 midspreads of the

hinges.

Structural content would include (among other things):

- Fisher5 is an Encapsulated PostScript File
- The bounding box coordinates for this resource are 175,655 and 487,745.
- the octet 0x6d at offset 0x622 is a Latin lower case letter m.
- "/med" is a PostScript label
  "/med" names a parameter to the function "box."
  "/med" identifies the median of a distribution.

ASCII, PostScript, John Tukey's graphical convention for distribution summaries, and a special-purpose language for encoding box plots are among the interpretive frames that connect the listing above to the conceptual propositions it expresses.

Among the format migration options to be considered for Fisher5 in a preservation scenario are keeping the resource in its original PostScript expression, transformation into vector PDF, or conversion into a raster PNG file. Strictly speaking, all three options preserve the conceptual content for human beings able to display the file using viewing software, provided that those viewers have an understanding of Tukey's box plot conventions. The current PostScript file encodes conceptual content in a declarative notation: median, range, hinges, and outliers are expressed in the scale of the original data, not the PS/PDF display plane coordinates. Those declarations would disappear in a translation from PS to PDF (usually understood as a lossless transformation). On the other hand, syntactically correct PostScript offers no guarantee of page independence (or, for that matter, halting). This PostScript file uses a non-embedded font that may not be as commonly available in the future as it is today. And the undocumented Postscript-based box plot markup language will be unfamiliar to people who might have an interest in extracting the data.

It would be relatively easy to transform box plot markup language into RDF, preserving all of the conceptual propositions listed above, and avoiding the shortcomings of PostScript, PDF, and PNG. Such an RDF re-expression could also include structural information, such as that Fisher5 is a box-and-whisker plot. But unlike a PDF or PNG translation, the resulting RDF would not express a box plot, and the advantages Tukey's notation offers for rapid visual assessment and comparison would not be available. We don't mean to suggest that this is a dilemma, or that no better migration options than these four are available (SVG might offer the best of all of them, for example). But interpretive frames would form a backdrop of dependencies for any such solution.

Situation Semantics and Interpretive Frames

The usefulness of frameworks based situation semantics Barwise and Perry, 1983 for understanding the assignment of meaning to XML structures has been argued for by Wrightson Wrightson, 2001 Wrightson, 2005 and Wickett Wickett, 2010. Barwise and Perry use situation semantics to model the meaning of indicative sentences as a relation between a situation[1] in which the sentence was uttered (the *discourse situation*) and a situation that the sentence describes (the *described situation*). The framework proposed by Wickett focuses on treating metadata records encoded in XML as a kind of utterance and,

following Barwise and Perry, examining how specific elements of XML documents contribute to inform consumers of the resource situations that were used assign meaning to the document as a whole. Situation semantics can be used here to give a theoretical grounding for interpretive frames that integrates them into a general theory of communication through markup and other notational structures.

In the case of data encoded in XML documents, we can also consider the document to be a series of indicative statements. In general a discourse situation gives an assignment for a speaker, an addressee, a (space-time) discourse location, and an expression. In terms of the framework (axioms) for encoding presented above, the speaker is the agent that commits to an expression, the discourse location is partially given by the assertion event, and the expression is the symbol structure that is the primary expression for the systematic assertion indicated in an assertion event. The role of the addressee and the end-point of the discourse location are left open until the document is viewed by some consumer of the data, only at this point will we have a complete discourse situation.

The described situation for data is a situation in which the real-world entities referred to by the symbol structures have the properties indicated by the relevant set of claims. In other words, the described situation is one in which the propositions that are the substance of the assertions (and therefore are data content) are all true. Since the described situation may not come to pass, we allow for data that is in error, by referring to things that do not exist or assigning properties to things incorrectly.

In Situations and Attitudes, Barwise and Perry discuss *resource situations*, the situations that the actors participating in a discourse situation have access to and use to identify and assign referents for the expressions that make up an utterance. Interpretive frames, as presented above, are a particular kind of resource situation. One kind of interpretive frame is the resource situation that govern the mappings between symbol structures and the things they refer to. This mapping was discussed by Barwise and Perry (and Wickett) as the speaker's *connections*. This interpretive frame assigns things like identifiers to individual plants in laboratory study, or assigns one column of a spreadsheet to a particular property of those plants. The preservation of meaning (in translation or simply within a single discourse situation) requires that the connections established by the addressee of an utterance are the same as those intended by the speaker.

XML documents, and digital objects in general, operate as communicative artifacts in virtue of a chain of computational structures that provide a background in which bitstreams can be understood as encoding symbolic structures. These interpretive frames are pointed to by things like standards for character encoding and by the various standards and specifications for hardware and software that allow us to create files and share them across systems. Barwise and Perry discuss how in natural language utterances, expressions that occur at one point in a discourse situation can supply a *setting* that influences how expressions that occur at another point in the discourse situation are understood. We can understand the interpretive frames that govern things like character encodings as resource situations that supply the necessary settings under which bitstreams can be interpreted as characters.

Discussion and Implications

One of the goals of the Data Conservancy project is to support interoperability of scientific data products. An interoperable data product is one for which given any addressee (consumer of the data product), the

set of connections that link the symbol structures to referents (objects of study, properties, values, etc.) are the same as those intended by the agent that indicated those symbol structures in the original assertion event. Representing structural propositions directly, either by asserting them (as with metadata annotations) or expressing them via encoding technologies like XML is one part of our strategy for helping to achieve this goal. Documentation of interpretive frames that connect propositions at different abstraction levels is another part of that same strategy.

We can see an application of these ideas in the OAIS Reference Model, which requires the inclusion of "representation information" as part of an Archival Information Package. This representation information is intended to give "information necessary to render and understand the bit sequences constituting the Content Data Object" Lavoie, 2004. However, it is important to draw a distinction between an interpretive frame and documentation of the frame. While OAIS representation information is necessary and can provide documentation of important aspects of the interpretive frames against which some data object is created, it must itself be in the form of a symbolic structure. On our view, interpretive frames are abstract mappings that correspond roughly to a *situation* Barwise and Perry, 1983. Therefore documentation can express elements of an interpretive frame, but a document cannot, by itself, *be* an interpretive frame.

Conclusions

Document markup solutions already do a better job than other notations in explicating structural content, and connecting it to appropriate interpretive frames. XML documents begin by declaring what they are, which encoding governs the interpretation of bit patterns, and (typically) what schema provides a syntax for the document. XML metadata applications offer numerous other forms of documentation and linking to bridge interpretive gaps. Most of the observations we offer here can be found stated either directly or indirectly by proponents of semantic documentation and enrichment frameworks like Formal Tag Set Definition and Intertextual Semantics Marcoux et al., 2009. But professional and research literature on markup semantics tends to foreground the role of markup itself in licensing inferences Sperberg-McQueen et al., 2002 Sperberg-McQueen and Miller, 2004 Sperberg-McQueen, 2011. Archiving standards like OAIS give an impression that "representation information" can supply needed interpretations, rather than simply document encoding choices. We recommend a different emphasis.

In our work with scientific data, the author/researcher's assertion event—rather than the resulting expression structure—seems to us the locus at which key identities are established. According to our axioms, it is these assertions that make propositions into data content, and supply symbol structures with their contingent meanings. The encoder of a data set can be likened (as Wendell Piez has suggested) to the player in a nomic game Piez, 2009, accepting some responsibility for creating the constitutive rules that govern his or her choices.

Acknowledgments

This research was supported by NSF Grant OCI-0830976. The authors wish to thank Allen Renear, the GSLIS Research Writing Group, and the anonymous reviewers of a prior Balisage submission for suggestions that have improved this paper.

Bibliography

[Adobe Systems, 1990] Adobe Systems Incorporated. *PostScript Language Reference Manual, Second Edition*. Addison-Wesley, Reading, MA, 1990.

[Baader et al., 2003] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The description logic handbook: theory, implementation, and applications*. Cambridge Univ Press, New York, 2003.

[Barwise and Perry, 1983] J. Barwise and J. Perry. *Situations and Attitudes. A Bradford Book*. The MIT Press, Cambridge, MA, 1983.

[Carlyle, 2006] A. Carlyle. Understanding FRBR as a conceptual model FRBR and the bibliographic universe. *Library Resources and Technical Services*, 50(4):264–273, 2006.

[Choudhury and Hanisch, 2009] S. Choudhury and R. Hanisch. Data conservancy: Building a sustainable system for interdisciplinary scientific data curation and preservation. In *Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data (PV) Conference, Madrid, Spain*, 2009.

[Dubin, 2003] D. Dubin. Object mapping for markup semantics. In B. T Usdin, editor, *Proceedings of Extreme Markup Languages 2003*, Montréal, Canada, August 2003.

[Hayes, 2004] P. Hayes. RDF semantics. Published by the World Wide Web Consortium at url http://www.w3.org/TR/rdf-mt/, February 2004.

[Holdsworth and Sergeant, 2000] D. Holdsworth and D. M Sergeant. A blueprint for representation information in the OAIS model. In B. Kobler and P. C. Harihan, editors, *Eighth Goddard Conference on Mass Storage Systems and Technologies: In cooperation with the 17th IEEE Symposium on Mass Storage Systems*, pages 413–428, Greenbelt, MD, 2000. NASA.

[IFLA, 1998] International Federation of Library Associations (IFLA). *Functional Requirements for Bibliographic Records: Final Report*. UBCIM Publications-New Series. Vol. 19, München: K.G.Saur, 1998.

[Lavoie, 2004] B. F Lavoie. The open archival information system reference model: Introductory guide. *Microform and imaging review*, 33(2):68–81, 2004.

[Marcoux et al., 2009] Y. Marcoux, C. M. Sperberg-McQueen, and C. Huitfeldt. Formal and informal meaning from documents through skeleton sentences. In *Proceedings of Balisage: The Markup Conference 2009*, volume 4 of *Balisage Series on Markup Technologies*, Montréal, Canada, August 2009. doi:10.4242/BalisageVol3.Sperberg-McQueen01.

[Piez, 2009] W. Piez. How to play XML: Markup technologies as nomic game. In *Proceedings of Balisage: The Markup Conference 2009*, volume 4 of *Balisage Series on Markup Technologies*, Montréal, Canada, August 2009. doi:10.4242/BalisageVol3.Piez01.

[Renear and Dubin, 2007] A. H. Renear and David Dubin. Three of the four FRBR group 1 entity types are roles, not types. In Andrew Grove, editor, *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology*, Medford, NJ, 2007. Information Today, Inc.

[Renear and Wickett, 2009] A. H. Renear and K. M. Wickett. Documents Cannot Be Edited. In *Proceedings of Balisage: The Markup Conference 2009*, volume 3 of *Balisage Series on Markup Technologies*, Montréal, Canada, August 2009. doi:10.4242/BalisageVol3.Renear01.

[Schmidt-Schauß and Smolka, 1991] M. Schmidt-Schauß and G. Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48(1):1–26, 1991. doi:10.1016/0004-3702(91)90078-X

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

[Sperberg-McQueen and Miller, 2004] C. M. Sperberg-McQueen and E. Miller. On mapping from colloquial XML to RDF using XSLT. In *Proceedings of Extreme Markup Languages*, Montréal, Canada, August 2004.

[Sperberg-McQueen et al., 2002] C. M. Sperberg-McQueen, D. Dubin, C. Huitfeldt, and A. Renear. Drawing inferences on the basis of markup. In B. T Usdin and S. R. Newcomb, editors, *Proceedings of Extreme Markup Languages 2002*, Montréal, Canada, August 2002.

[Sperberg-McQueen, 2011] C. M. Sperberg-McQueen. What constitutes successful format conversion? towards a formalization of 'Intellectual content'. *International Journal of Digital Curation*, 6(1), 2011.

[Wickett, 2010] K. M. Wickett. Discourse situations and markup interoperability. In *Proceedings of Balisage: The Markup Conference 2010*, volume 5 of *Balisage Series on Markup Technologies*, Montréal, Canada, August 2010. doi:10.4242/BalisageVol5.Wickett01.

[Wrightson, 2001] A. Wrightson. Some semantics for structured documents, topic maps and topic map queries. In *Proceedings of Extreme Markup Languages*, Montréal, Canada, 2001.

[Wrightson, 2005] A. Wrightson. Semantics of well formed XML as a human and machine readable language: Why is some XML so difficult to read. In *Proceedings of Extreme Markup Languages*, pages 1–11, Montréal, Canada, 2005.

---

[1] The technical notion of a situation is close to our intuitive one: a situation occurs at a space-time location and involves individuals participating in certain roles and standing in relations. It also closely corresponds to the notion of a state of affairs, especially since situations are abstract objects that may or may not obtain.

Author's keywords for this paper: Semantics; Situation Theory; Digital Preservation

David Dubin

<ddubin@illinois.edu>
Research Associate Professor
University of Illinois

David Dubin is a Research Associate Professor at the University of Illinois Graduate School of Library and Information Science in Champaign, IL. David conducts research on foundational issues of information representation and description.

Karen Wickett

<wickett2@illinois.edu>
University of Illinois

Karen M. Wickett is a doctoral student at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Simone Sacchi

<sacchi1@illinois.edu>
University of Illinois

Simone Sacchi is a Doctoral Student at the Graduate School of Library and Information Science and

research assistant at the Center for Informatics Research in Science and Scholarship under the NSF granted Data Conservancy Project. His research interests are in conceptual foundation of digital curation and knowledge representation.