This paper was presented at the 2008 TEI Annual Members Meeting, held November 6–8 2008, in London, England, United Kingdom.

FRBR Group 1 Entities and the TEI Guidelines[1]

Kevin S. Hawkins

Introduction

When speaking about literature or about text encoding, we sometimes use terms like *work*, *text*, and *document* quite loosely—so loosely, in fact, that you could hold an entire graduate-level seminar[2] or publish a whole book[3] to discuss the meanings of these three words. While lexical ambiguity is a common feature of human language, loose usage may point to a deeper ontological ambiguity—or simply a lack of clarity—over what is being discussed. The various members of a *bibliographic family* are confused not only by novice text encoders but also, I believe, by the many contributors to the TEI guidelines. Clarifying this confusion over what is the object of encoding may help us to get around some of our persistent problems in applying TEI markup and lead to texts that are more machine-readable than at present.

While there are many ontologies of bibliographic families, for this analysis I will apply the FRBR model to TEI text encoding. I first want to acknowledge my debt to Klemens Bobenhausen and Hans Walter Gabler, whose paper at last year's meeting inspired me to explore this topic.[4]

The FRBR model as a data model

Functional Requirements for Bibliographic Records: Final Report (FRBR) was first published by the International Federation of Library Associations and Institutions (IFLA) in 1998. It has as one of its aims "to provide a clearly defined, structured framework for relating the data that are recorded in bibliographic records to the needs of the users of those records."[5] While this goal was originally aimed at increasing interoperability of catalog records, it has come to be thought of as part of an effort to rethink library catalogs entirely, abandoning the last vestiges of catalog cards.

Most interest in and discussion of FRBR focuses on the entities in "group 1": *work*, *expression*, *manifestation*, and *item*. (Henceforth, these terms will be used only in the FRBR sense.) Definitions and examples are given in the table below:

Entity	Definition	Examples
work	"a distinct intellectual or artistic creation"[6]	Bulgakov's *Master and Margarita*

[1] I greatly appreciate the help Thomas M. Dousa, Dave Dubin, Allen Renear, Richard Urban, all of the University of Illinois at Urbana-Champaign, for providing references and invaluable guidance in the development of this paper.

[2] For example, a course entitled "Document, Text, Work" was offered in the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign in the fall of 2002.

[3] See, for example, Richard P. Smiraglia, *The Nature of "A Work": Implications for the Organization of Knowledge* (Lanham, Md.: Scarecrow Press, 2001).

[4] Klemens Bobenhausen and Hans Walter Gabler, "Markup as Theory of Text," presented at *TEI@20: 20 Years of Supporting the Digital Humanities*.

[5] IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report. As Amended and Corrected through February 2008* (http://www.ifla.org/VII/s13/frbr/frbr_2008.pdf) 7.

[6] IFLA, p. 17.

expression	"the intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms"[7]	the text of the first version, which Bulgakov burned in a stove; the censored version published in *Moskva* magazine in 1966–67; the English translation by Michael Glenny
manifestation	"the physical embodiment of an expression of a work"[8]	the Glenny translation published in paperback by Harper & Row; the Glenny translation published as an audiotape
item	"a single exemplar of a manifestation"[9]	my copy of the Harper & Row paperback edition of the Glenny translation

Note that FRBR explicitly includes non-print media.

The FRBR group-1 entities are often thought of as constituting a hierarchy, with work at the top and item at the bottom,[10] but the FRBR report is ambiguous on this point[11] and, furthermore, it has been noted that the idea of a hierarchy for FRBR is problematic.[12] Instead of *hierarchy*, I will use a term that is slightly less ontologically dishonest—*levels of text*—to refer to the these four entities, which vary from the most abstract (work) to the most concrete (item).

The FRBR report admits to using entity-relationship analysis[13] but gives a clear disclaimer that its conceptual model "does not carry the analysis to the level that would be required for a fully developed data model."[14] Indeed, deficiencies both in the functional requirements[15] and in the provisional data model[16] have been identified. A

[7] IFLA, p. 19.
[8] IFLA, p. 21.
[9] IFLA, p. 24.
[10] See, for example, David Mimno, Gregory Crane, and Alison Jones , "Hierarchical Catalog Records: Implementing a FRBR Catalog," *D-Lib Magazine* Oct. 2005 11(10): http://www.dlib.org/dlib/october05/crane/10crane.html.
[11] There is only one use of the term "hierarchical" in the FRBR report, in the introduction: "Further study could be done on the practical implications of restructuring MARC record formats to reflect more directly the hierarchical and reciprocal relationships outlined in the model" (see IFLA, p. 6).
[12] See Allen H. Renear and Yunseon Choi, "Modeling Our Understanding, Understanding Our Models: The Case of Inheritance in FRBR" *Proceedings of the 69th ASIS&T Annual Meeting* (2006) http://eprints.rclis.org/archive/00008158/01/Renear_Modeling.pdf (preprint) and Karen Coyle, "Hierarchy v. relationships," *Coyle's inFormation: Comments on the Digital Age, Which, As We All Know, Is 42* (Nov. 7, 2007) http://kcoyle.blogspot.com/2007/11/use-of-hierarchy-as-organizing.html.
[13] See IFLA, p. 6.
[14] See IFLA, p. 3.
[15] See, for example, Alexander C. Thurman, "FRBR and Archival Materials: Collections and Context, not Works and Content," in *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools* (Westport, Conn.: Libraries Unlimited, 2007) 97–102.
[16] See, for example, Allen Renear and David Dubin, "Three of the Four FRBR Group 1 Entity Types are Roles, not Types," *Proceedings of the 70th Annaul Meeting of the American Society for Information Scienec and Technology (ASIST)* (2007) http://hdl.handle.net/2142/9094 (preprint).

working group convened by the Library of Congress even cautioned against adopting the FRBR model until it can be studied further.[17]

There is promising work to formulate an object-oriented definition of FRBR called $FRBR_{OO}$,[18] which attempts to harmonize FRBR with an existing ontology for the museum community called the CIDOC Conceptual Reference Model (CRM);[19] however, I will use the original FRBR model in this discussion for its simplicity.

Markup as a data model

While FRBR attempts to create a data model of the bibliographic universe, what about text encoding? Does markup correspond to a data model of text?

The advantage to using TEI and other similar markup is its ability to describe the content of the text and not just its appearance, as a typesetting language like TeX or a word processor used without styles would do. Markup (or word processor styles) allows us to manipulate the content in useful ways: for example, if you encoded foreign words separately from other italicized words, you could easily produce a list of all foreign words in a text in order to produce a glossary, or you could produce a scholarly edition in which foreign words were in bold instead of italics.

However, this prioritization of content objects is easily taken to its extreme. A TEI novice may at first want to encode every underlying content object that can be found, but you soon realize the impossibility of the task because not everyone agrees on the underlying structure of the text. Marking up text is about describing its structure as you see it, but it's also about describing those content objects that are relevant to you. While you may be able to mark definitively all sorts of structures in the text, you only bother with certain ones because you have limited resources—time, staff, and computing power—to handle those structures of interest.

However, we are often cautioned not to make encoding decisions based on currently available software because someday new software will be developed that doesn't have the same constraints as your current software. What, then, is the basis for our decisions? There is a vague notion of trying to represent the underlying features and structure of the text. While we accept that markup is used to solve specific problems, many of us think of it as corresponding to the Platonic reality of the structure of text.

This can get messy, as we all know. For example, should a footnote number (a <ref> or <ptr/> tag) at the end of the sentence (surrounded by <s> tags) occur before or after the "</s>"? While it's helpful for software development if we all reach the same decision about the footnote and the <s> element—and the TEI strives to reach consensus on this—I think we need to accept that scholars have sufficiently different views of text to prevent a perfect unified ontology from ever being codified. While the TEI allows customization to add elements, for example, only certain types of changes to the fundamental model of text in TEI are allowed. I believe this striving to model Platonic

---

[17] Library of Congress Working Group on the Future of Bibliographic Control, *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control* (http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf): 32–33.

[18] International Working Group on FRBR and CIDOC CRM Harmonisation, *FRBR Object-Oriented Definition and Mapping to $FRBR_{ER}$: Version 0.9 Draft* (2008) http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR_oo_V0.9.pdf.

[19] ICOM/CIDOC Documentation Standards Group, *Definition of the CIDOC Conceptual Reference Model: Version 4.2.5a* (2008) http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.5a.pdf.

reality in the face of more modest functional requirements is symptomatic of our confusion over the role of TEI markup.

Since both the FRBR model and the TEI guidelines attempt to meet certain functional requirements through loose, imperfect data models, it may prove helpful to borrow from the FRBR model to improve the TEI guidelines. In particular, we should see which of the FRBR group-1 entities might correspond to the object of encoding when using TEI markup.

What does markup describe?

The TEI was designed for marking up digital copies of physical objects. More recently, it has also been used for marking up born-digital documents and even for composing them, with P4 Lite and now P5 taking these cases into account. However, creating a TEI document as a transcriber and creating one as an author are very different actions, as Allen Renear[20] and Wendell Piez[21] showed us in their explorations of the descriptive/procedural distinction, previously assumed to be straightforward. To take an example, if I transcribe a phrase in an early printed book and declare that it *is* a chapter heading, that is very different from me composing a text today and noting that I would like a phrase to *be* a chapter heading.

However, if we put aside encoding one's own text, we find there are also variations in usage of markup when encoding someone else's document. For example, if I tag an instance of "Wall Street" with <name type="noun"> or with <name type="synecdoche">, I describe as a linguistic object the noun phrase in the text. But if I tag an instance of "Wall Street" with <name type="place" ref="#place_wallstreet">, I describe the referent of the noun phrase—the thing referred to.

At a higher level, however, it's not clear what the sum of tags in a TEI document represents. Are we representing only the item that was the source document, or are we also representing the manifestation, expression, or possibly even the work? It seems likely to me that we often represent more than one of these in the same TEI document. The TEI guidelines discuss various structures in the text which often fail to form a hierarchy;[22] what if these structures actually come from different levels of text according to the FRBR model?

Separate TEI documents for each level of text

Since the FRBR model posits that there are four different levels of text, what if we create a separate TEI document to represent each level? Consider this abridged list of data elements that might be used in the body of these TEI documents:

FRBR group-1 entity	TEI elements within <body>
work	?

[20] Allen Renear, "The Descriptive/Procedural Distinction is Flawed," *Markup Languages: Theory and Practice* 2(4): 411-420.
[21] Wendell Piez, "Beyond the 'Descriptive vs. Procedural' Distinction," *Proceedings of Extreme Markup Languages* (2001) http://www.idealliance.org/papers/extreme/proceedings/html/2001/Piez01/EML2001Piez01.html.
[22] TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (Oxford, Providence, Charlottesville, Nancy: TEI Consortium, 2008, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf): xxxiii-xxxiv (section v.ii).

expression	<choice> <corr> <floatingText> <l> <lg> <orig> <name> <p> <quote> <reg> <s> <sic> <w>
manifestation	<catchwords> <fw> <lb/> <pb/>
item	<add> <damage> <del> <handShift/> <unclear>

You'll notice that the expression level corresponds to various content hierarchies (prosodic, poetic, syntactic, etc.), the manifestation to the physical hierarchy, and the item to those pesky arbitrary spans (in TEI sometimes better treated as pairs of milestones) that interfere with hierarchies at the expression or manifestation level. It seems that we are already thinking in terms of levels of text; in fact, if these levels were always encoded separately, we could avoid many cases of needing overlapping hierarchies in our XML! Unfortunately, overlapping hierarchies could still occur, especially at the expression level: for example, the structure of drama overlaps with the syntactic hierarchy when one character completes the sentence of another.

Upon closer examination, we see that elements do not all map so cleanly to only one of the four levels. Two examples come to mind. First, a particular edition (manifestation) of a text (expression) might have a typo in it; in this case, <orig> or <corr> might be more appropriate at the manifestation level rather than the expression level, unless we are purists and claim that this one typo warrants the creation of a new expression. Second, a printing defect might make every copy of that edition (manifestation) turn out to have a smudge; in this case, <unclear> would be better suited to the manifestation level.

How about the TEI header? This part of a TEI document is actually much more complicated to reassess since it already describes both the TEI document and the source document, so with this extra dimension we would potentially need eight TEI documents! But even if we did this, there is considerable ambiguity in the TEI guidelines which I am

not sure how to resolve: for example, should <sourceDesc> describe the source document as a work, expression, manifestation, or item? Should the <fileDesc> describe the TEI document as a work, expression, manifestation, or item? Many elements could apply at more than one level: the <title> of a work would be its uniform title (in the world of cataloging), whereas the <title> of the expression would be an identifier for a particular version of a text, and the <title> of the manifestation would be that text which appears on the title page. (While the FRBR model does not prescribe a title attribute for the item entity, it might make sense to have one for recording the title in cases where it differs from other items of that manifestation, such as when the item is defaced.[23]) Many of these ambiguities mirror ambiguities in cataloging practice (which has not yet been FRBRized): for example, are subject headings assigned for the work or the expression, and do notes in a catalog record apply to the work, expression, manifestation, or item?

My goal is not to eliminate overlapping hierarchies or avoid any of the ingenious solutions proposed for expressing overlapping hierarchies in markup;[24] rather, it seems that the need for multiple hierarchies in a single XML document is symptomatic of broader ontological confusion—of trying to encode different levels of text in a single XML document and not distinguishing between them. Clarifying the level of text that is the object of encoding—and the level of text which the creator of the TEI document hopes to convey—may help us, and maybe even our machines, to think more clearly.

Making XML truly machine-readable

It is often said that XML is designed to be machine-readable, though its verbosity as a data-storage format makes it perhaps the most human readable of such data formats. Nevertheless, text encoders creating TEI documents in XML—or for that matter catalogers creating records in MARC format—underspecify that which is obvious to a human reader. This deficiency has been deeply explored by those working on the BECHAMEL Markup Semantics project,[25] and I think we can agree that we are actually far from having TEI documents which can be processed by machine. In short, machine-readable is not the same thing as machine-processable.

Machine processing requires a sound data model, which FRBR is not. $FRBR_{OO}$ is much better suited to this role. While I remain pessimistic of our ability to formulate a single data model of text, if the TEI ever did so, it would need to incorporate a data model of bibliographic entities such as $FRBR_{OO}$. In the meantime, keep the original FRBR model, or your favorite model of bibliographic families, in mind as you create TEI documents: it just might help you think more clearly as a scholar of literature.

---

[23] See Kevin S. Hawkins, "Entailment of Entities and Implicature of Attributes in the FRBR Model," presented at *Modern Information Technologies and Written Heritage: From Ancient Texts to Electronic Libraries* (2008) http://www.ultraslavonic.info/preprints/20080618.en.pdf.

[24] CONCUR, Goddag, LMNL, MECS, TexMECS, Trojan Horse markup, and XCONCUR are discussed in C. M. Sperberg-McQueen and Claus Huitfeldt, "Markup Discontinued: Discontinuity in TexMecs, Goddag structures, and rabbit/duck grammars," *Proceedings of Balisage: The Markup Confererence* (2008) http://www.balisage.net/Proceedings/html/2008/Sperberg-McQueen01/Balisage2008-Sperberg-McQueen01.html.

[25] See David Dubin and David J. Birnbaum, "Interpretation Beyond Markup," *Proceedings of Extreme Markup Languages* (2004) http://www.idealliance.org/papers/extreme/proceedings/html/2004/Dubin01/EML2004Dubin01.html and citations therein to previous work.