Proppian Structural Analysis and XML Modeling

Article · January 2001

CITATIONS
13

READS
127

1 author:

Scott Malec
University of Texas Health Science Center at Houston
7 PUBLICATIONS **30** CITATIONS

SEE PROFILE

PROPPIAN STRUCTURAL ANALYSIS AND XML MODELLING

Scott Alexander Malec

University of Pittsburgh

Abstract (Please control whether every title you refer to in your contribution appears in the bibliography)

In this paper, the development of an XML DTD called Proppian Fairy Tale Markup Language (PftML) is proposed. The DTD was developed by implementing a feedback loop with the markup of approximately 20 Russian magic tales. Propp's "functions", the fundamental units of a Russian magic tale narrative were found to be analogous to structural metadata, and as such renderable by hierarchically arranged "elements" in XML documents. PftML utilizes a DTD to create a formal model of the narrative structure of Russian magic tales. Additionally, tags throughout the corpus remain consistent and limits to the framework of the markup are established without remainder.

Keywords

XML, Markup, Narratology, Folklore

According to Gerald Prince Dictionary of Narratology 1987

1. Introduction

Narrative lies at the heart of what it means to participate in the life of man the "wise". Indeed, the word narrative itself stems etymologically from the Latin *gnarus*, meaning "knowing" or "expert" and which derives from the Proto-Indo-European /*gna*/, meaning "to know," *i.e.* "The narrator is one who knows"

(Prince Pp. 39; why not in bibliography). Surprising then, in the two thousand and three hundred some odd years after Aristotle's *Poetics*, that remarkably little ink has been shed upon so human an endeavor as story telling. At the dawn of the twentieth Century however, there was a windfall of new approaches to the study of narrative, such as the early efforts of scholars like Alfred Lord Bates and Milman Perry. Continuing to this day, the scope of intellectual interest in the narrative form has expanded beyond the province of literary studies and into such fields as anthropology (Colby, Dundes), semiotics, cognitive science (Lehnert, Rumelhart, van Dijk), linguistics and even computer science (Lang). What these approaches share is a belief in the need to analyze narratives through segmentation, for example into "plot units" (Lehnert), narremes and syntagms. In this paper, I will examine some of the practical and theoretical issues surrounding the analysis of a particular corpus of narratives, that of Afanas'ev's *Русские народные сказки.* (*Russian Folk Tales*). This is the very same work that V. I. Propp used to inform his methodology of decomposing narratives, his principle contribution to the field of narratology. In fact, the crux of this project is the use of a subset of the same Russian language corpus from which Propp drew, since it allows for an empirical test of the conclusions of Propp's initial analysis against the original data.

Over the course of three years, I developed an XML DTD called Proppian Fairy Tale Markup Language (PftML) and have completed the markup of 20 Russian magic tales conforming more or less to this document type definition. The materials that I used had to be born-again digital, and here I describe the process in terms of the technical details (DTDs, markup, formalisms, hermeneutics), the difficulties that I encountered with my content model and the course for the future of this ongoing project. I shall also address theoretical problems and make observations pertinent to text encoding in general.

2. Overview

2.1 Background

The beginnings of this project lay with an insight that Propp's "functions", the fundamental units of a Russian magic tale narrative, are analogous to structural metadata, and as such renderable by hierarchically arranged "elements" in XML documents. This idea was fomented in part by the debate in Humanities Computing (HC) concerning the Ordered Hierarchy of Content Objects (OHCO) hypothesis. This hypothesis argues that texts can be decomposed into simple hierarchically ordered data trees not subject to remainder variables. In principle, the idea is that a morphological analysis of narrative (a type of text) – implemented in XML – would be a promising candidate to provide an illuminating angle onto the OHCO debate. To this end, I have developed an XML application based on Propp's *Morphology of the Folktale* (1928), called Proppian fairy tale Markup Language (PftML). PftML utilizes a DTD to create a formal model of the structure of the Russian magic tale narrative and standardizes the tags throughout the corpus.

2.2 Key Definitions and Concepts

In this section, I will define and describe some of the crucial concepts for understanding my work as presented here.

Structural metadata is a term used to describe the metadata which "relates different objects and parts of objects to each other." (Arms 2001: 71). Structural metadata represented by the markup of a poem, for example, might consist of annotations for the compositional levels of the line, the stanza, and the canto number.

The decomposition of narratives into their respective elements, which Propp called "*functions*", and the representation of these by means of structural metadata is the crux of the PftML project. According to Propp, a "[F]unction is understood as an act of a character defined from the point of view of its significance for the course of the action." (Propp 1968: 21). In PftML markup, a

function manifests itself compositionally as a chunk of text, which is more often than not a sentence in a narrative, but which can also be a sentence fragment. One of the purposes of textual analysis using PftML is to be able, by adequately describing the compositional structure in terms of the elements and content model of the XML DTD, to decompose tales into Proppian functions without remainder. As discussed in the section entitled "Issues and Problems", however, this is not always the case.

XML (eXtensible Markup Language) is a meta-markup language. A meta-markup language is one that can be used to describe new markup languages. Unlike HTML, XML has no fixed vocabulary or syntax. By contrast, the vocabularies (XML "Elements") and syntax (in Extended Backus-Naur Formalism [EBNF] notation) utilized in XML documents may be created through the use of either a DTD (Document Type Definition) or an XML schema. XML was derived from SGML (Standard Generalized Markup Language) and preserves many useful and powerful features of its parent markup language, including the aforementioned DTDs, as well as descriptive markup and data that is independent of any hardware / software platform. It has become the preferred langauge due to its elegance and economy of use. Created to describe the hierarchical patterns of data, it is well-suited to the sophisticated forms of analysis applied in corpus linguistics, bioinformatics, aeronautic documentation systems and other domains within academia, medicine and industry. Because the formal nature of Propp's ideas lends itself so readily to the formal nature of XML modeling, scholarly activity such as PftML could provide powerful structural insights that are not apparent to the eye of an unaided human reader, however scrupulous. Although support for overlapping hierarchies, such as that which SGML supports is desired at times, XML's simplicity is often its own reward.

Vladimir Iakovlevich Propp (1895-1970) was a Russian folklorist loosely affiliated with Russian Formalism (cf. Ehrlich 1955 [why not in bibliography?]), a literary movement in the 1910's and 1920's of the last century that sought to uncover the "devices" (приёмы) that lend to the "literary" its unique substance, thereby differentiating itself from other linguistic utterances. Propp studied Germanic philology at St. Petersburg University and then taught lan-

guages in secondary schools for some time after the October Revolution. Extracurricular reading of East Slavic (Ukrainian, Great Russian, White Russian and Ruthenian) folklore, a field that Propp felt was neglected in his country, combined with his passions for literary theory (including that of Veselovsky, father of the field of comparative literature in Russia, and Formalism) and the morphological writings of J. W. v. Goethe, brought this scientist to the conclusions that he reached in his most highly acclaimed work. In 1928, Propp published his widely influential treatise on the devices of narrative, the Morphology of the Fairy Tale, which sparked intense interest in the structural features of narrative texts, among other modes of cultural production. In the Morphology, Propp treats a corpus of 100 tales in Afanas'ev's collection and discovers basic recurrent units of the magic tale plot ("functions") and the *ars combinatoria* employed implicitly to arrange them. Propp defines a folkloric morphology to be "a description of the tale according to its component parts and the relationship of these components to each other and to the whole." [bibliographical information missing] Propp's original goal with his early work was to derive a morphological method of magic tale classification, based on the arrangements of "functions", just as one would place a living organism within a particular taxon based on the rhythm of its parts.

3. Propp's Approach

3.1 The Gist

The gist of Propp's approach was to devise a list of elements, or vocabulary, and then to determine the syntax, or rules of arrangement, of those elements. Curiously, this is the self-same process used in writing a DTD to describe documents that are to be born-again digitally and encoded in XML/SGML.

When examining a corpus of texts, one first discovers which composition elements are present across the corpus and establishes a list of elements and then determines how these components are put together – the calculus of possible combinations. This second part is described in the content model in the DTD, which one constrains or "loosens" in correspondence with the character-

istics of the document set. After one determines the patterns of the form and content uniform across the corpus, one then decides which information it would be important to extract and which determines sets of attributes with which to qualify these elements.

Presciently, the mathematician Alonzo Church summed up this process in a 1951 article describing the semantics of formalized languages:

> As primitive basis of a logistic system it suffices to give, in familiar fashion: (1) the list of primitive symbols or vocabulary of the system (together usually with a classification of the primitive symbols into categories, rules and rules of inference). (2) The formation rules, determining which finite sequences of primitive symbols are to be well-formed expressions, determining certain categories of well-formed expressions, among which we shall assume that at least the category of sentence is included. (Church 1951: page?).

The set of elements in a DTD are this vocabulary, while the content model, which describes how the elements are put together, manifests the document syntax. In substance, these ideas were echoed by the French literary critic and philosopher Paul Riceour (as quoted in Culler 1975: 26) in his critique of structuralism, and then again by Maler and Andaloussi as a coherent methodology for developing DTDs (Maler / Andaloussi 1996: 30): list, categorize, model, and validate.

3.2 Propp's Conclusions

Propp posited the following general rules by which his corpus of tales was composed:

> Functions of characters serve as stable, constant elements in a tale, independent of how and by whom they are fulfilled. They constitute the fundamental components of a tale;
>
> The number of functions known to the fairy tale is limited;
>
> The number of functions known to the fairy tale is limited;

> All fairy tales are of one type in regard to their structure. (Propp 1968: 21)

These simple rules are a syntax amenable to formulation in terms of a content model of an XML DTD, with only minor qualifications (discussed below), and, along with the identification of Proppian functions (see definition above), form the basis of this research program into structural metadata of Russian magic tale narratives.

Propp seems to me to be so intriguing because the type of formalism of expression he describes seems to be precisely that of computing humanists involved in text encoding. On the surface, at least, Propp's idea lends itself so well to such an interdisciplinary project.

4. Selection of Material

I selected the texts from Afanas'ev's corpus from two groups:

1. Aarne-Thompson type 480: tales of the "Cinderella" or the *Золушка* type (see Smirnov for an exhaustive collection of tale variants)
2. tales already analyzed in Propp's appendix

I chose some tales because they were classified as Aarne-Thompson 480s, which meant that they have to do with step-daughters and evil-stepmothers, such as Cinderella. I found them to be of interest because in some of these tales the role of donor (the Proppian function performed by the "Fairy God Mother") was fulfilled alternatively by *Морозко*, the Russian Jack Frost, Baba Yaga, a hideous witch (see Johns 1998 for discussion), or even a Mare's Head, in a Ukrainian variant.

The other tales I selected because Propp had already laid out his analysis in Appendix III of the 1968 English translation. I could thus perform the scholarly and scientific task of comparing my results against the analysis of Propp himself. Rather as a 21st Century Miguel de Cervantes, I found myself in a position of difficulty not unlike Pierre Menard in Borges "Pierre Menard, Author of Quixote": It was considerably more intellectually strenuous a task to write

Propp's Morphology a second time, and I have yet to plumb the depths of this irony.

5. Markup Observations

Structural metadata is not always a clear cut issue, as the parsing of literary structures raise important issues endemic to the fields of hermeneutics and taxonomy.

Marking up texts, I noticed several distinct interpretive stages that a text would go through once it had been copy edited to a tolerably clean state. These three stages of markup are outlined here:

1. Getting a "feel" for the text – what type of story is it? Is the hero a victim-hero (with "villainy" for a cardinal function) or a seeker-hero (with "lack" as a cardinal function)?
2. Preliminary Analysis – which functions present themselves on the surface, draft a "sketch" of what the plot looks like; create a list of Proppian functions and characterize the significant dramatis personae who move the plot forwards;
3. Parsing – this stage may take several runs, and involves the arduous task of actually doing markup.

As Lou Burnard has pointed out elsewhere, a markup is a tool to make concrete a particular interpretation of a text (cf. Burnard 2001: page?). The final parsing of any story up above may or may not alter the DTD. Of considerable philosophical and hermeneutic interest, the DTD is here a type of formalism which makes manifest the encoder's understanding of the whole corpus of texts, the outer limits of constraints to which any particular story may or may not adhere. This feedback loop, or hermeneutic circle, provides an intellectual thrill which can lead to a kind of vertigo in the novice, as there are an overwhelming number of questions which remain unanswered: when should I alter the DTD to conform to an apparently aberrant document instance? What are my ultimate goals for this project? What constitutes a function in the final analysis? But there can be no final analysis, only increasingly rarefied under-

standing, and more questions. Gadamer, in speaking of exegesis of scripture, applies ? equally well to Russian magic tales here:

> The literal significance of Scripture, however, is not univocally intelligible in every place and at every moment. For the whole of Scripture guides the understanding of individual passages: and again this whole can be reached only through the cumulative understanding of individual passages. (Gadamer 1989: 175)

Likewise, the DTD guides the interpretation of individual Proppian functions, while this DTD, representing the sum of the understanding of the syntax of the entire corpus, is further refined through reference to Proppian functions as they instantiate themselves in particular texts. Stated succinctly, "the hermeneutic circle, an intrinsically mystical notion, derived from the observation that to understand a part, its function in the whole must be clear; yet the function of the whole can only be derived from an understanding of its parts." (Burnard 2001: 31).

Various violations of expectation naturally occurred when marking up the texts. These 'violations' include the following:

1. implicit functions;
2. violations of the "sequence" rule; and
3. ambiguity of functions.

In XML analysis, 'implicit functions' mean that the function is implied in the text from context, but not present on the surface structure of the text. When you're speaking of a 'cardinal' function such as 'villainy' or 'lack', you need something there, otherwise there would be neither a move nor a fairy tale, so you wind up with an empty function, with perhaps a comment.

Since XML employs extended Backus-Naur Formalisms, XML is unforgiving when you break the sequence of DTD's elements as adumbrated in the content model. Finally, part of the purpose of the whole XML / Propp project, as I see it, was to clarify what Propp meant, to resolve ambiguities in the morphology by reinterpreting his own textual evidence and perhaps reproduce his results.

Other interesting problems and question that come to mind which I have specifically encountered with attempting to codify Propp's schema in the form of a DTD have been:

1. "non-function" elements – what to do with the remainder, the filler useful in oral culture, but useless in written (eg. скоро сказка сказывается)
2. what qualifies as a "function"?
3. how loosely can Propp suffer to be deconstrained? or was Propp referring more to the "gravity" of cultural tendency?

Speaking of inter-subjectivity, one of the basic premises in the philosophy of science is that others should be able to duplicate one's results. The humanities under post-modernism, as currently configured, make no such assumption. Yet markup is one area in which interpretations of texts of different scholars from different backgrounds can occasionally seem to congeal with one another. Yet, when it comes to the interpretation of humanistic media, there will always be some grounds for dispute.

My own parsing of the texts would on occasion differ from those of my project advisor, David J. Birnbaum. When they would differ markedly, we would engage in dialogue, and argue our points pro and contra respective interpretations. Most of the time, one of us would concede with the other, since such dialogue engenders understanding, and it is through such heady discourse that one learns to come to grips with the limitations of one's own frame of reference and the biases inherent in one's relationships to the texts.

John Unsworth wrote the following in his "What is Humanities Computing?":

> Consensus-based ontologies (in history, music, archaeology, architecture, literature, etc.) will be necessary, in a computational medium, if we hope to be able to travel across the borders of particular collections, institutions, languages, nations, in order to exchange ideas. Those ontologies will in turn exist in a network of topics, a web of grading zones, to use a term that Willard McCarty has used to explain humanities computing, having borrowed that term from a book that itself borrows concepts of anthropology to explain the practice of physics. And as that genealogy of that metaphor suggests, come tomorrow, we will require the rigor of computational methods in the discipline of the humanities not in spite of, but because of, the way that human understanding and human creativity violate containment, exceed representation, and muddle distinctions. (Unsworth year: page).

Thus, it is often through acknowledging the limitations of our models and their inadequate to describe in full the products of the human intellectual that we notice and take delight in the exception to the rule, to the highly constrained syntax of a tightly woven XML DTD, for example.

There are, of course, yet other philosophical rat-holes to be explored, such as the precarious art of using metaphors in scientific thinking. Max Black (why not in bibliography?, year? page?) wrote: "a metaphor is the tip of a model." I shall avoid digressing here, but for an interesting discussion of this point, read Fernand Hallyn's *Metaphor and Analogy in the Sciences*.

6. Current Work

My current work involves the creation of a web based interface for querying my corpus of Russian magic tales in XML. This step represents a major step for me from digitally representing the structure of the plot of narrative texts via descriptive markup toward actually processing them. In my current research, I am using PftML to explore data structures (e.g. embedding, repetition, sequencing, etc.) latent in natural language narratives, such as Russian magic tales, in searching for unexpected correlations between functions, and in devel-

oping an alternative system of folktale classification (from the Aarne-Thompson) based on Propp's work.

In addition to making text more navigable and findable (a subject of enduring interest to the A & I community), the development of a querying device and an interface for markup can aid in making the markup more consistent across the corpus. As I marked up the text and came to grips with the corpus, my understanding of what constituted the particular functions altered and affected my markup. With a tool such as what I am currently in the process of developing, I will be able to regulate and render the markup consistent across the corpus, an hermeneutic feat inconceivable without such a tool. [e]Xist, an Open Source native-XML Database Management System for storage and retrieval of my corpus is currently being used to bring more consistency to the markup of my corpus.

In addition, compatibility with TEI-headers for administrative metadata is planned with PftML to reflect current best practices in text encoding. Since TEI is modular, I have only to incorporate the TEI-headers to help to make the administrative metadata of my DTD more granular.

7. The Future of PftML

My grand vision for this project is that it is preliminary to more extensive work: namely, to explore other open technologies (XSLT, SVG) which would ultimately enable me to visually represent comparisons of the internal structural metadata (through markup) of textual narratives.

These resultant graphs would address the principle concern of Propp's early work: the morphological classification of Russian magic tales, as opposed to the thematic approach of the Finnish School of Aarne-Thompson. Morphology signifies here a hypothetical classification scheme according to a segmental arrangement of functions. These segments have been represented in tales through the arrangement of elements specified by attributes of decomposed XML documents. Since according to Propp, the magic tales in Afanas'ev's corpus are instantiations of a single Ur-tale, one can analyze how closely the

schema of one tale follows the schemata of other tales. This type of analysis would reveal which tales are more closely related in an empirical way, and further Propp's work to devise a typology of magic tales based on composition textology. One might also make a study of the development of expertise with a markup system, rather as Simon and Chase (1973) analyzed the "chunking", or pattern recognition, in the cognitive behavior of chessmasters.

8. Conclusion

I hope that this work would help to emphasize the significance of Propp's contribution to narrative studies. If we are to design systems that reflect structures inherent in complex data, such as textual narratives, it will be necessary to face the limitation of both our understanding of them and of the capabilities of the tools utilized to represent this limited understanding. At the same time, we must accept these limitations in the interest of praxis, deriving what value we can from them, while never ceasing to explore ways in which this utility may be improved.

Bibliography (first names must be given in full)

Afanas'ev, A. N. (1957) *Russkie Narodnye Skazki*. Moscow: Gosodarstvennoe Izdatel'stvo Khudozhestvennoi literatury.

Afanas'ev, A. N. (1973) *Russian Fairy Tales*. Trans. Norbert Guterman. New York: Pantheon Books.

Aarne, Antti / Thompson, Stith (1964): *The Types of the Folktale: A Classification and Bibliography* (= Folklore Fellows Communications 184). Helsinki: Suomalainen Tiedeakatemia.

Arms, William Y. (2001): *Digital Libraries*. Cambridge: MIT.

Borges, Gorge Luis (1964) *Labyrinths: Selected Stories & Other Writings*. Ed. by Donald A. Yates and Jannes E. Irby. New York: New Directions Publishing Corporation.

Burnard, Lou (2001): "Title of contribution", in: Fiormonte, Domenico / Usher, Jonathan (eds.): *New Media and the Humanities: Research and Applications.* Proceedings of the first seminar Computers, literature, and philology. Oxford: Oxford Humanities Computing Unit pages?.

Chase, William G. / Simon, Herbert A. (1973): "The Mind's Eye in Chess", in: *Visual Information Processing*. New York: Academic Press.

Church, Alonzo (1951): "The Need for Abstract Entities", in: *American Academy of Arts and Sciences Proceedings* 80: 100-113.

Colby, Benjamin N. (1973): "A Partial Grammar for Eskimo Folktales", in: *American* Anthropologist 75: 645-662.

Culler, Jonathan (1975): *Structuralist Poetics*. Ithaca: Cornell University Press.

Dundes, Alan (1964): *The Morphology of North American Indian Folktales* (= Folklore Fellows Communications 195). Helsinkii: Suomalainen Tiedeakatemia.

eXist Open Source XML Database <http://exist-db.org> [date of last visit]

Fiormonte, Domenico / Usher, Jonathan (eds.) (2001): *New Media and the Humanities: Research and Applications*. Proceedings of the first seminar Computers, literature, and philology. Oxford: Oxford Humanities Computing Unit.

Gadamer, Hans-Georg (1989): *Truth and Method*. Trans. Joel Weinsheimer and Donald G. Marshal. New York: Continuum.

Hallyn, Fernand (ed.) (2000): *Metaphor and Analogy in the Sciences*. Dordrecht: Kluwer Academic Publishers.

Johns, Andreas (1998): "Baba Iaga and the Russian Mother", in: *Slavic and East European Journal* 42, 1: 21-36.

Lang, Raymond (1997): *A Formal Model for Simple Narratives*. Doctoral Dissertation, Tulan University.

Maler, Eve / El Andaloussi, Jeanne (1996) *Developing SGML DTDs: from Text to Model to Markup*. Upper Saddle River, NJ: Prentice Hall.

Prince, Gerald (1987): *A Dictionary of Narratology*. Lincoln: University of Nebraska Press.

Propp, Vladimir Ia. (1968): *Morphology of the Folktale*. Trans. Lawrence Scott; ed. Alan Dundes. Austin: University of Texas Press.

Renear, Allen / Mylonas, Elli / Durand, David (1996): "Refining our Notion of What a Text really is", in: ? <http://www.stg.brown.edu/resources/stg/monographs/ohco.html> [date of last visit].

Rumelhart, David E. (1975): "Notes on a Schema for Stories", in: Bobrow, D.G. / Colins, A. (eds.): *Representation and Understanding: Studies of Cognitive Science*,. New York: Academic Press 211-236.

Smirnov, IU. I. (1993): *Russkie Narodnye Skazki o Macekhe i Padceritse*. Novosibirsk: Nauka.

Unsworth, John (year?): "What is Humanities Computing?", in: *Computerphilologie* 2 <http://computerphilologie.uni-muenchen.de/jg02/unsworth.html>.