DOI:10.1145/2209249.2209267

A searchable meta-graph can connect even troublesome house elves and other supernatural beings to scholarly folk categories.

BY JAMES ABELLO, PETER BROADWELL, AND TIMOTHY R. TANGHERLINI

Computational Folkloristics

THE STUDY OF folklore, or folkloristics, is predicated on two premises: traditional expressive culture circulates across and within social networks, and discrete "performances" of traditional expressive forms serve an important role as the locus for the ongoing negotiation of cultural ideology (norms, beliefs, and values). The underlying assumption is that folklore emerges from the dialectic tension between individual members of a cultural group on the one hand and the "traditions" of that group on the other. This ongoing tug-of-war ensures that traditional expressive culture is constantly changing, adjusting to the needs of the individuals within a group.

The goal of any folkloristic investigation is to understand how traditional expressions create meaning for the people who create and receive them; studies range from a consideration of fairy-tale telling by 19th-century peasants (such as in the classic works of the Grimm brothers) to analysis of rumor propagation in the aftermath of disasters (such as Hurricane Katrina in 2005 and the 2011 tsunami in Japan). By exploring the traditional expressions of group members across time and space, scholars can develop a sophisticated understanding of the complex processes of culture development and change.

Since inception of the field in the 19th century, folkloristics has been based on a three-part process:

- Fieldwork consisting of the identification of a folk group and the collection of its traditions;
- Archiving, editing, and publishing these collections; and
- Analyzing the collected folklore based on a single or combination of multiple theoretical perspectives.

For the most part, folklorists limit their studies to small, well-defined collections or subsets of much larger collections. A study corpus is often selected based on criteria like genre or topic. The "variants" in the resulting study corpus are then subjected to "close readings"; from an ethnographic perspective, close readings focus on analysis of the symbolic aspects of the expression as an important meaning-making process for storytellers and their audiences. Conclusions drawn from close readings are subsequently abstracted to make more general comments about the changing contours of the cultural ideology of the group in question. This approach has worked well for generations of folklorists, particularly in the context of the relatively small size of accessible collections and the general alignment of research

» key insights

- **The field of computational folkloristics weds algorithmic approaches to classic interpretive problems from the field of folklore.**
- **A multimodal network representation of a folklore corpus (hypergraphs) liberates folklore exploration from the limitations of existing classification schemes.**
- **Imagine a system in which the complexities of a folklore corpus can be explored at different levels of resolution, from the broad perspective of "distant reading" down to the narrow perspective of traditional "close reading."**

questions with existing archival-finding aids.

As folklore collections are increasingly digitized, and as Web-based social media sites become established loci for the circulation of traditional expressive forms, earlier methods of folklore scholarship have begun to falter for several reasons: First, it is increasingly difficult to determine the boundaries of a study corpus, not only because existing "finding aids" that rely on hand-indexing have not kept up with the growth of digital resources but also because modern research questions no longer align with the original ideas underpinning the early aids. Second, the size and scope of folklore corpora have increased so dramatically it is impossible to apply close-reading methods to even a fraction of the items in a study corpus.

Due to this dramatic change in the scholarly landscape, where issues of access no longer pose absolute limits on the scope of a study, scholarship in folklore, as in the humanities as a whole, must adapt to a new environment where many thousands, if not millions, of texts are readily available.

At the same time, canonical approaches to corpus selection that guaranteed high precision but also resulted in low recall are more difficult to defend intellectually. Whereas a scholar could, until recently, reasonably propose a study focused exclusively on, say, the several hundred folksongs indexed in the Child collection of the Anglo-American ballad,[8] the availability of many more related texts from the same time period or the same cultural area necessitates that the scholar acknowledge this greatly expanded domain. This fundamental change in the accessibility and size of target domains is a welcome development but also requires that folklorists augment existing methodologies to take advantage of the change in scope.

Along with increased accessibility of machine-readable primary-source materials in the form of digitized archives (such as The Danish Folklore Archive),

ILLUSTRATIONS BY J.F. PODEVIN

born-digital archives (such as The Shoah Foundation Visual History Archive), and enormous informal repositories of personal interaction (such as Facebook and Twitter), there has been an equally rapid increase in methodologies for search and discovery in large, poorly labeled corpora (such as Google). However, many of these new algorithmic approaches for search, discovery, and analysis developed for the Web do not readily translate to the disconnected realm of most collections of literature or folklore. On the other hand, many of the research questions scholars want to address—generally inconceivable prior to wide-scale availability of large digital corpora—demand more targeted approaches than those developed for the biological and physical sciences, scientific co-citation networks, and e-commerce. Availability of substantial digital folklore corpora ultimately speaks of the need for computational folkloristics.

As a sub-discipline of folklore, computational folkloristics encompasses:

- Digitization of existing resources or collection of new resources in machine-actionable form;
- Development of extensible data structures for description and storage of these resources;
- Novel methods for classifying folklore data based not only on existing classification schemes and metadata but also on surface-level phenomena (such as the linguistic features of texts);
- Domain-sensitive methods for search and discovery; and
- Algorithmic methods for corpus study, including visually rich approaches that fuse statistical representations of the data with appropriate historical maps,[35] as well as combinatorial graph analytical approaches (such as network-based role discovery).[30]

These methods augment, rather than supplant, earlier close-reading methods prevalent in folklore, allowing for a more consistent approach to the selection of study materials for a given research question. Ideally, one should be able to move seamlessly between a bird's-eye view of the overall study corpus and the complex interconnections among people, places, and artifacts that underpin the corpus at one end of the spectrum and the close reading that has characterized a great deal of prior folklore scholarship at the other.

Distant Reading

In 2000, Franco Moretti[25] proposed the theoretically rich concept of "distant reading," anticipating the profound changes in humanities scholarship presaged by developments in information technology. In Moretti's analysis "Distance... is a condition of knowledge: It allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems."[25] Distant reading is thus a complementary approach to the close-reading approaches that have characterized humanities research for centuries. It is a particularly apt approach for folklore, since folklorists are not only interested in the particular features of a discrete text but also the much broader picture of how that discrete text (or performance) fits into the wide range of traditional expression through time and space.

We endorse this view of the need to fuse close reading and distant reading. The backbone of computational folkloristics must leverage current technological developments that allow us to semantically interconnect large amounts of data and explore them relatively easily in integrated analytical and visual environments inexpensively. Here, we outline some of the components of an ideal computational folkloristics system and the main elements that provide quick classification of a large, poorly labeled corpus coupled with intuitive visual navigation. This system would allow a researcher to explore the complex relationships among tradition participants (the people in the tradition group), the historical and geographic environment in which they lived and worked, and the folklore they created, and also allow for incorporation of pertinent scholarly feedback annotations. A central component of such a system can be based on "hierarchical graph maps,"[21] an algorithmic counterpart to distant reading. Hierarchical graph maps fuse two of the most common abstractions of visual navigation—graphs and maps. They also align with Moretti's conceptualization of methods suited for the implementation of distant reading in the humanities.[23]

Limitation of Classification Schemes

Classification is a vexing first-order problem in folklore. Since the field's inception in the 19th century, folklorists have been concerned with the classification of texts, devising numerous classificatory systems. They have evolved into a series of overlapping and relatively shallow ontologies. Genre classifications play an important role in the organization of most folklore collections, with, say, narrative folklore being sorted into genres (such as myth, folktale, legend, rumor, fairy tale, and ballad).[3] Such classifications are problematic as a model of folklore production, given the vast array of native genre distinctions from target cultures that do not align with scholarly categories.[12] Other efforts have concentrated on organizing folklore by topics within a single genre or group of closely associated genres (such as folk tale and legend). Perhaps best known is "The Types of International Folktales" commonly referred to as the ATU index.[37] The underlying indexing philosophy of folklore type indices, which exist for specific national traditions[9] or specific genres,[10] are best summed up as "one story/one classification." A slightly more fine-grain index that allows for greater overlap at the unit of "tale" focuses on "motif," or constitutive narrative element of a tale. In it, a tale consists of a series of motifs, and it is the motifs, rather than entire tales, that are indexed, then cross-referenced to published collections.[13,36] In terms of search and retrieval, these traditional classifiers provide a high degree of precision but, due to that precision, sacrifice a great deal of recall. Recall

is so low that searches based on the indices miss many potentially interesting variants. Although these classificatory schemes have served folklorists for so long, the systems fail dramatically when suddenly addressing tens of thousands or even hundreds of thousands of texts from published and unpublished collections.

High-precision indices (such as standard folklore type and motif) have a second problem—cost of implementation. When target collections are relatively small—dozens of examples to perhaps hundreds—classification systems can be applied through manual methods. But when target collections expand to 10,000 or more examples, they become much too costly to implement. This cost is amplified by the need to revise and update existing indices and re-index existing collections, a problem that is particularly acute when the archives are open-ended and continuously growing (such as blog posts about uprisings in the Middle East).

As a simple, historical example from the Danish materials, no one has yet classified (according to the ATU index) the several thousand fairy tales in the collections of the Danish Folklore Archive (http://www.dafos.dk), nor does it seem anyone ever will. We may well be nearing the end of the useful lifespan of single-classifier schemes, given their implementation costs and inability to describe the complexity of tradition. So how does a scholar perform search and discovery to answer research questions in folklore? What is needed is a system for the automatic indexing of a collection that is flexible enough to allow for shifting priorities of folklore researchers and that can adapt to a sudden increase in the scale of the target domain. In the following sections, we describe one such approach to problems in search and discovery based on a network model of a folklore collection. As our test corpus, we use a subset of the Tang Kristensen collection of Danish folklore housed at the Danish Folklore Archives in Copenhagen.

Ghost Stories?

Evald Tang Kristensen, a schoolteacher in Jutland, Denmark, is today widely regarded as the most prolific folklore collector ever.[35] Over the course of his active collecting career, 1865–1923, he collected folklore from approximately 4,000 named individuals whose ages and life histories he meticulously recorded. His collection spans more than 24,000 manuscript pages, including nearly a quarter-million stories, songs, proverbs, riddles, rhymes, and descriptions of everyday life. Approximately one-third of this collection has been published in a series of printed volumes, organized first by genre and subsequently indexed according to Tang Kristensen's own idiosyncratic topic categories; these categories are not consistent across the 80 volumes in the published collection.

Since 1999, we have been digitizing and transcribing not only the published collection but also an unpublished collection consisting largely of field-collection manuscript pages. We have extracted from Tang Kristensen's memoirs, letters, and personal papers the names of his informants and the places these people lived. These places, along with places mentioned in their stories, have been geo-coded using historical gazetteers. We also extracted Tang Kristensen's collecting routes through Denmark indexed to

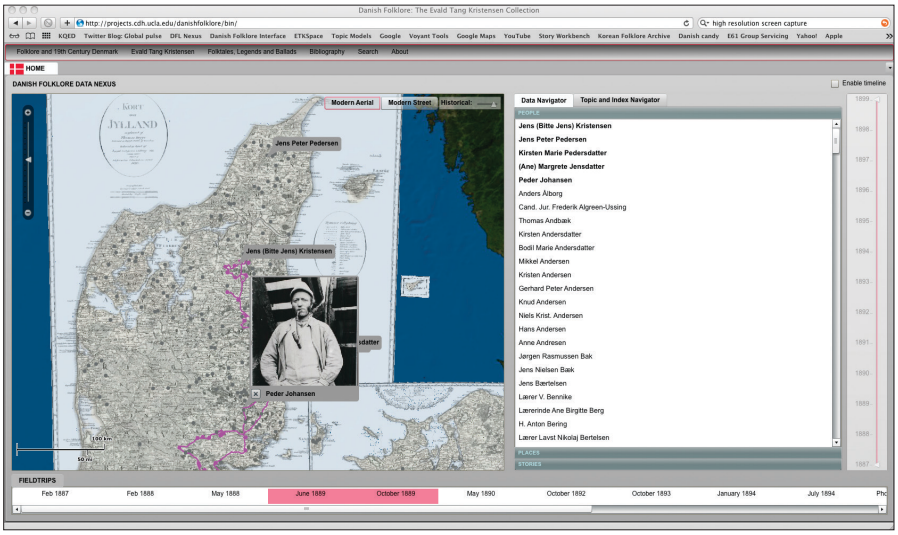


Figure 1. A simple GUI for exploring a subset of a folklore archive, here the Evald Tang Kristensen collection, includes historical maps and the ability to drill down to images of the collector's field notes.

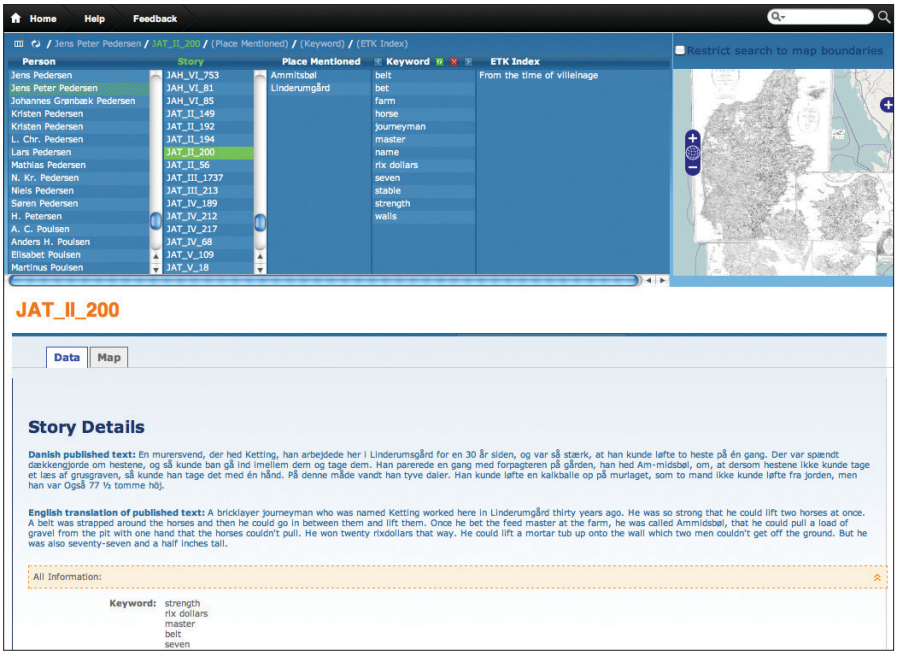


Figure 2. The faceted browsing system ETKspace (based on the mSpace faceted browser) allows users to quickly select a subset of the overall corpus based on multiple criteria (such as storyteller, places mentioned, places of story collection, and date of collection).

Table 1. Attributes for each story text.

Attributes / Notes	Term frequency of keywords	Shallow-ontology	Evald Tang Kristensen index	Place Names	Personal Names
Description	A term frequency count of keywords selected from the vocabulary of the entire corpus; keywords can be bigrams.	A shallow, hierarchical ontological representation of each text in the corpus. The nine top categories—Actions or Events; Animals; People; Places; Resolution; Stylistics; Supernatural Beings; Time, Season, Weather; Tools, Items, and Conveyances—along with 125 second-level categories, offer a hierarchical overview of the content elements of the story.	A topic index to each of Tang Kristensen's published collections.	A geo-referenced representation of places mentioned in a story and the place where a story was collected.	An index of the people mentioned in a story and the storytellers.
Theory	The underlying theory of this approach is known as "bag of words," where the vocabulary of a given text is recognized as meaning bearing.	The underlying theory of this approach is predicated on Vladimir Propp's *Morphology of the Folktale*[28] and, later, Alan Dundes's refinement of that work.[12] Dundes proposed that individual elements of a story (allomotifs) often fulfill a structural role in the narrative (motifemes). The shallow ontology first developed in Tangherlini[34] catalogs the stories' content on multiple levels, allowing comparison across stories even when the vocabularies of the stories are divergent.	Nearly all published folklore collections include their own topic indices based on the collector's or editor's evaluation of "what the story is about." In Tang Kristensen's publications, he separated the collections first according to genre, then according to "top level" topics (such as "Witches and their Games"), then into more specific categories (such as "Driving with three wheels").	Folklore theory has always been concerned with the relationship between traditional expression and the physical environment.[22] Inclusion of geo-referenced place names here allows one to explore the geographic location of stories in regard to attributes related to content.	Recent folklore theory emphasizes the role of storytellers in shaping tradition as part of their own creative and ideological expression.[31] People mentioned in stories help situate the stories in the local social environment.
Computed?	Yes. Keywords were identified using AutoMap. The vocabulary of each individual story was subjected to a stop word filter (articles, pronouns, prepositions, conjunctions) and rudimentary stemming (Snowball stemmer). A term-frequency count for each individual story was computed based on this limited vocabulary.	Not yet.	No.	Partially. The identification and alignment of place names with historical place name gazetteers is a semi-supervised operation; the de-duplication of place names in the corpus was assisted by DDupe.[4]	Partially. Named-entity detection implemented in Mallet was used to generate a list of potential personal names from the corpus.[24] This list was aligned with the partial list of personal names in one of Tang Kristensen's indices.
Benefits	Can be computed quickly.	Allows for comparison of stories that share ontologically similar content, even when the vocabulary is divergent.	These are preexisting indices that represent the collector's or editor's historical view of the collection; folklorists are familiar with these indices.	Places mentioned reveal association between topics and places in the real world made by storytellers.	The emphasis in folklore theory on the role of the individual in the creation of tradition and the highly localized and historicized nature of much folkloric expression can be captured by this index.
Challenges and drawbacks	Although Danish is not highly inflected, there is a need to at the very least stem the vocabulary. Umlaut and limited cases of syncope introduce inaccuracies into the stemming. Without stemming, inflected forms of words are counted as discrete lemmata. Bag-of-words approaches discard all important grammatical information that contributes to meaning making.	For this work, we applied the shallow ontology by hand and are working on automating the process using DanNet, the Danish version of WordNet.	The indices fall prey to the "one-text-one-classification" pitfall of early folklore classification systems.	Many of the places mentioned in stories are difficult to resolve to existing place-name gazetteers. Some place names have changed over the years, some have ceased to exist, and others are only known in local culture. Small orthographic variations in place-name spelling also contribute to some uncertainty.	Danish names are complex. A century of at times contradictory laws have led to a situation where many people have a last name based on a patronymic ending in -sen, with a limited number (approximately 116) possible first components (such as Pedersen from Peder's son) and a second last name derived entirely from place names. The former phenomenon makes it difficult to align people mentioned with external archival resources (such as a census), while the latter makes it difficult to easily discern between place names and personal names.

places, people, and time. Moreover, we have searched census data and church records to situate his informants in both space and time. As part of this effort, we developed two basic modules to address corpus navigation: The first is a relatively simple study environment appropriate for exploration of a subset of a larger corpus incorporating rich biographical and interpretive data, along with historically accurate maps (see Figure 1). The second is a faceted browsing application intended to help select corpus resources that can be ported to more sophisticated study environments (see Figure 2).

Although the work is far from complete, this growing window into the daily life of the rural population of late-19th-century Denmark offers an intriguing glimpse into the conflicting agendas—cultural, economic, intellectual, and political—of scholars, amateur folklore collectors, and storytellers, as their country struggled toward greater democratic freedom and free-market organization. It also makes an excellent testbed for developing and testing tools for computational folkloristics.

Table 2. Attributes associated with DS IV 650, the misclassified ghost story.

ETK Index	Manor lords, ladies and mistresses
Places Mentioned	Skårupgård, Todbjærg
Personal Names	[none]

keyword	*frequency*
bed	2
church	1
counselor	1
death	3
door	4
farm_hand	3
headless_horse	1
night	1
north	1
old	1
rafters	1
riding	1
stall	3
torment	1

Shallow ontology	*Entry*
Resolution	Negative
Animals	Farm animal
Animals	Horse
Places	Barn
Places	Church
Actions or Events	Death
People	Farmhand/Shepherd
Supernatural Beings	Ghost/Revenant

One of Tang Kristensen's published collections of legends includes the following story: "It was the old counselor from Skårupgård who came riding with four headless horses to Todbjærg church. He always drove out of the northern gate, and by the gate was a stall. They could never keep the stall door closed. They had a farmhand who closed it once after it had sprung open. But one night, after he'd gone to bed, something came after him and it lifted his bed straight up to the rafters and crushed him quite hard. Then the farmhand shouted and asked them to stop lifting him up there. 'No, you've tormented us, so now you'll die...' I heard that's how two farmhands were crushed to death. He wanted to close the door and then they never tried to close it again."[32]

The story concerns a vengeful haunt, yet Tang Kristensen classified it as being about unnamed manor lords. Consequently, a researcher interested in the relatively simple topic of ghost stories in late-19th-century Denmark would be unable to discover the story through existing finding aids. Even in the digital realm, the discovery of the story as a ghost story is a challenge. Since it makes no overt mention of ghosts, simple keyword searching on Danish terms for ghosts (*spøgelse*, *genganger*) would fail. Experiments with supervised text classifiers, in this case a naïve Bayes learner trained on 200 stories labeled "ghost stories" by Tang Kristensen, also fails to classify the story as a ghost story. Indeed, only a relatively high degree of domain expertise would allow a researcher to find the story, recognizing that headless animals are often related to hauntings. Unfortunately, headless animals are also related to portents and the devil in Danish tradition; thus the recall of stories for a keyword search on the Danish term for "headless" (*hovedløs*) would trigger the opposite problem with far greater recall than is probably helpful. The story is not an anomaly in the collection but rather illustrative of the underlying problem of existing classification schemes for the vast majority of folklore collections.

Folk Story Hypergrahs

Since folklore can be classified according to numerous criteria, connecting items together based on them allows information to flow throughout the corpus. That is, building a multimodal network representation of the corpus offers a stark contrast to the compartmentalization of information represented by a singly classified printed text model. As a bonus, a network model of the corpus opens up the applicability of multi-resolution network analysis. The intention is to represent folklore as a dynamic hypergraph of people connected in time and space to places and their stories. This representation is extensible as more resources become available. One can thus imagine a system that stores the social networks of tradition participants, the academic networks of scholars and collectors, communication and transportation networks linking geographical places, repertoire networks linking storytellers to their stories, and finally linguistic and semantic networks that link stories. The underlying theory is that it is difficult to find something if it is not connected to something else. By connecting the three main actors of the folklore equation—people (storytellers and scholars), places (where stories were collected and mentioned in stories), and stories (or folkloric expression in general)—a researcher can discover not only specific texts but patterns in the network data not readily apparent if the collection is presented in disconnected fashion. The resulting model of a folklore collection rests on a series of graphs, each describing a mode of the complex folklore hy-

pergraph or meta-network. This hypergraph model of folklore offers the potential for productively spanning the distant-reading and close-reading spectrum explored earlier.

Connections

Connecting texts to one another is one of the most complex tasks of the hypergraph-generation process. Networks describing the connection of stories to their tellers and stories to the places they mention are relatively easy to generate since they are mainly an alternative representation of the corresponding databases. More complex is generating <story to story> networks based on multi-objective criteria. As a case in point, one criterion of our work is to not abandon earlier systems of classification (such as the tale type, motif indices, and collection-specific indices) but incorporate the information from them into the story representation. More generally, the chosen story representation should offer mechanisms for incorporating accumulated scholarly knowledge. Consequently, our approach is to model each story as an attribute-valued vector, where the values of some attributes are known a priori and others are computed. Some attributes may take as values either simple scalars or "links" to more complex structures or processes. Other attributes may have associated with them time-varying functions recording a sequence of attribute values over time, or a time series. We did not incorporate time-varying attributes in our initial limited experimental dataset of stories; Table 1 lists the attributes for stories in the corpus, and Table 2 lists the attributes of a single story.

The dataset includes 342 storytellers and 942 stories, along with Tang Kristensen's topic index for two collections—*Danske sagn*[33] and *Danske sagn, ny række*[32]—as the basis for the network. We supplemented this information with two additional weighted graphs: The first was based on a simple "shared keyword" weight computed for each pair of stories. Using AutoMap we generated a set of 1,201 keywords shared among all stories across the corpus, after eliminating pronouns, prepositions, articles, and conjunctions.[7] This bottom-up approach gave us an important view of the corpus based on shared vocabulary across texts. The second was a top-down approach that categorized stories according to a shallow ontology for the corpus developed specifically for the realm of Danish folklore.[35] Such shallow ontologies depend on domain expertise, providing a view of the corpus attuned to the "tradition dominants" of a particular tradition group.[15,21,26] The use of natural language processing (NLP) tools in concert with DanNet, the Danish version of WordNet, may allow us to develop further the first-generation shallow ontology for the Danish folklore corpus.[27,29] Topic-modeling methods (such as Latent Dirichlet Allocation and Latent Semantic Analysis) could be used as additional methods for generating a <story to story> graph.[6,11,14] The advantage of the hypergraph model, or multimodal network, is that new network representations can be added to the overarching model as their applicability to the study of folklore is further understood.

Our hope is that the end result is a system with applicability for researchers working with a range of folklore and ethnographic collections and ultimately for any collection or series of collections of culturally expressive forms. In our own earlier work, we explored development of a shallow ontology for the Danish folklore collection as a part of a method for revealing narrative tendencies in the corpus based on four attributes of the storytellers: gender, occupation/class, age when the stories were told, and education.[35]

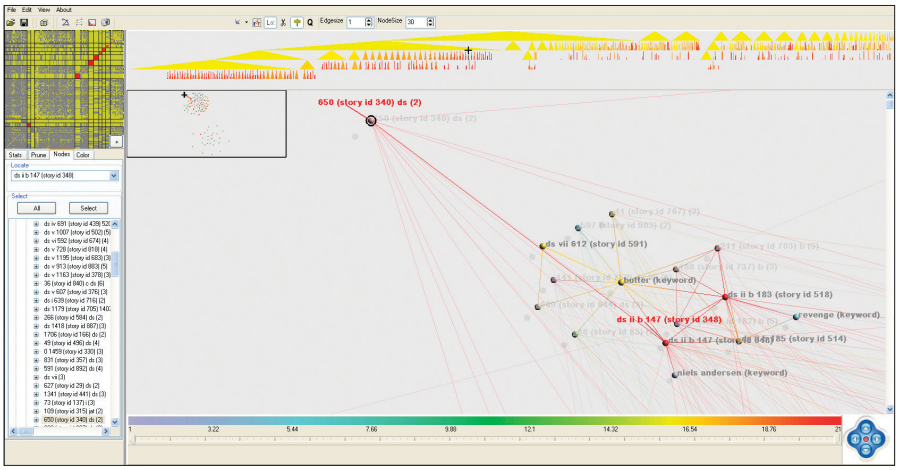
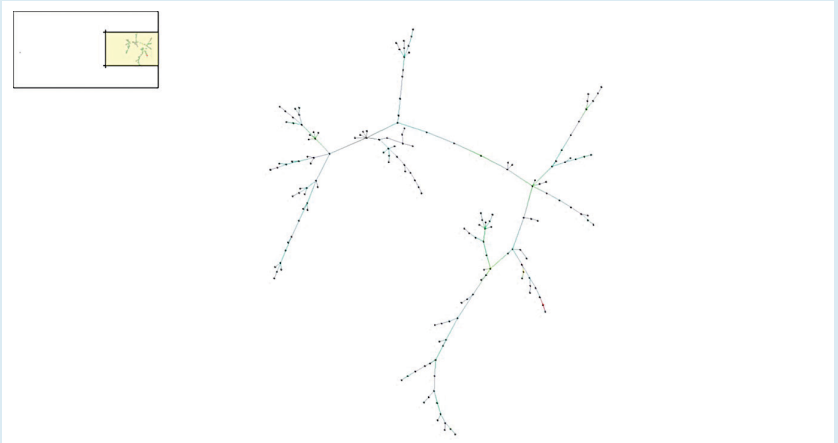


Figure 3. Overview of the hypergraph browser for computational folkloristics. The target ghost story (story id 340) is upper left, identified with a landmark. The story of the house elf (*nisse*) is right of center, also identified with a landmark (story id 348). The upper navigation screen shows a tree view of the organized hypergraph; the left panes provide navigation and selection; the lower right includes a graph-navigation toggle. This hypergraph browser is based on Abello et al.[2]

Figure 4. Overall story space for the test corpus, with a cluster selected for further inquiry.



In 2000, when digitizing large collections became practical, we revisited the coding of the stories from the 1994 study[35] and began refining the story-attribute set into a consistent shallow ontology. We further refined the shallow ontology in 2008–2009 to develop a geo-referenced topic browser for a collection called the "Danish Folklore Nexus" (http://projects.cdh.ucla.edu/danishfolklore) (see Figure 1) and a faceted browsing system for Danish folklore still under development (http://etkspace.ucla.edu) (see Figure 2). Over 2008, the initial test corpus was coded with attributes from the shallow ontology in a semi-supervised system. Tangherlini, an expert in Danish folklore, checked automatic assignments made by AutoMap's thesaurus function for accuracy.[7] This iterative process of expert checking of automatic assignments also yielded further refinement of the ontology.

Typical Folklore Tasks

The resulting network for our test corpus includes 2,973 nodes with 52,663 edges, and the system's design aims to address two questions emblematic of regular research problems in folklore:

- Given a poorly labeled target story (such as the ghost story mentioned earlier), how can the system place the target story in a neighborhood of "similar" stories?; and
- Can the system suggest candidate stories that would likely be of interest to a folklore researcher?

We addressed them through the following basic computational modules of a system:

Hypergraph Constructor → Hypergraph Projector → Hierarchical Graph Map Constructor[1] → Graph Navigator[2] → Visual and Textual Selectors → Landmark Annotators → Web Browser Finder → Answer Verifier → Report Generator → Hypergraph Curator

The system interface lets users interact with each computational module through textual input/output, textual search, drop-down menus and tabs, mouse selectors, visual sliders, zooming, a navigation wheel, a landmark annotator, and statistical reports (see Figure 3).

Initial user interaction amounts to specifying a target story descriptor—choosing from a story identifier, collection of key words of interest, or known scholarly categorization. The system then brings up both a visual network representation of a "cluster" of stories related to the target story, together with a Web link, to a color-coded textual representation of the related stories and other pertinent scholarly information (such as authors, places mentioned, and historical context or period where the story is placed) (see Figure 4). Based on this information, a researcher might decide to place a landmark tag in one of the story's elements, get a projection of the meta-network on a particular subset of attributes, and/or provide feedback on the perceived quality of the system's suggested answer (see Figure 5). At this point, the researcher can visually navigate to clusters of stories adjacent to the system-provided cluster, select

Figure 5. Close-up view of a selected story cluster focusing on stories concerning economic threat.

Note the clear semantic divide between terms in the upper half of the cluster graph and terms in the lower half, with those in the upper half focusing on historical figures and those in the lower half focusing on supernatural figures.

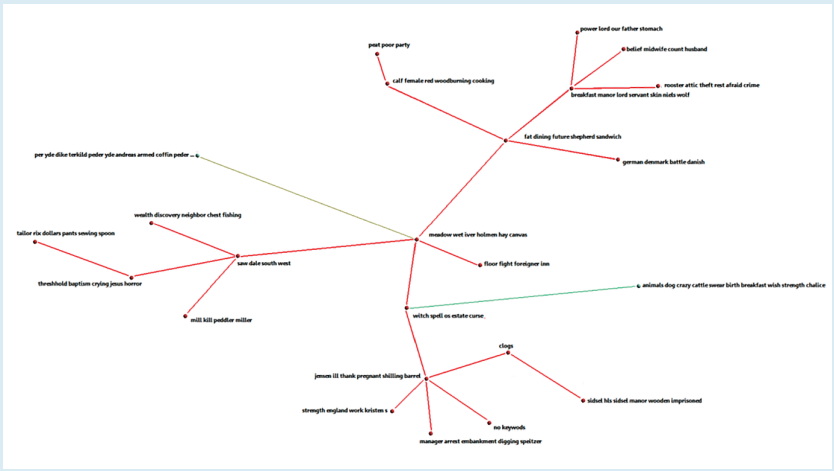
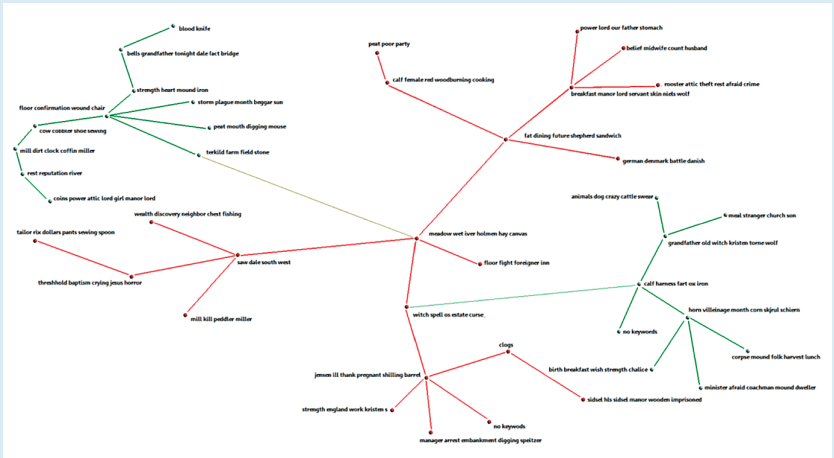


Figure 6. Three related story clusters, expanded for discovery of candidate stories for further exploration.

The two related clusters are related more closely with supernatural threats. The cluster on the left is bifurcated, with stories at the top related to robbers, theft, and murder, and those at the bottom related to Satan. In the cluster on the right, the stories at the top involve witches and animals related to Satan, with both categories representing significant economic threat. The stories at the bottom involve interaction with mound folk, a supernatural group closely related to economic issues.



another story, and iterate the process (see Figure 6). The system keeps the user-selected landmarks as aggregated counts that can be used by authorized story "curators" to update the existing hypergraph for future analysis. This incorporation of scholarly feedback is essential for the system's performance, as it allows for the aggregation of accumulated expert knowledge, which in turn helps guide researchers as they navigate the hypergraph.

The graph-theoretic techniques we applied to generate the final hierarchical graph map involve four major tasks:

- Define similarity measures in stories;
- Construct a weighted graph among stories, where the weight of the connection between two stories is obtained through order-preserving transformations of their similarity measures;
- Decompose the graph(s); and
- Compute a "hierarchy tree" for the obtained graph(s).

The similarities measures we used are based on several algorithms, including: Hellinger distance,[23] weighted Jaccard coefficient,[19] cosine similarity,[18] and scholar-defined weights. The measures are transformed through Gaussian kernels and other standard kernels[20,30] that assign to every pair of stories that share at least a minimum set of attributes a corresponding similarity weight. In general, the graphs obtained through these transformations vary depending on the type of similarity measure used. Consequently, we obtain a final weight between two stories by computing a norm of their similarity-weights vector. Because the obtained graph has different density "regions," we follow the approach described in earlier work by Abello et al.,[2] decomposing the collection of stories into maximal sub-graphs according to their inherent k-connectivity. Each obtained sub-graph is then clustered using an adapted version of Markov Clustering.[2] The entire process is encoded in a data structure called a "hierarchy tree" visually represented in the user interface as a hierarchical graph map (see Figure 3),[2] constituting one of the central modes of interaction between user and story corpus. Worth noting is that each cluster includes a unique associated label string that encodes its "semantic" placement in the hierarchy tree. These cluster labels can be used to help provide and receive feedback from scholars studying the corpus, as well as judge the effectiveness of the suggested classification.

The trouble with house elves begins with their unpredictable nature, which, for folklorists, makes them difficult to track through the landscape of the story corpus.

Performance on the Two Folklore Tasks

The first task we included in the system was to place the target story in a neighborhood of closely related ghost stories. The system succeeded admirably, moving the story from a group of stories about manor lords in the original classification scheme to a neighborhood of stories about ghosts and other supernatural beings that haunt workers in farm buildings and barns—a new implicit category that did not exist in Tang Kristensen's original indices. Although we left success criteria for this work fuzzy, we placed the target story in the same part of the organized meta-graph as several ghost stories exhibiting certain topical similarities; for example, another story where a haunt makes it impossible to use part of the farm appears nearby in the hypergraph: "A woman died in a farm in Dokkedal... and she showed herself every noon at a specific place in one of the rooms and always as a black shape. Because of that, the room stood empty for a long time, since nobody dared go in there..."[33]

Perhaps more interesting is the system's ability to meet the second task—suggest candidate stories, given the researcher's interest in a target story. In this case, the visual representation of the meta-network reveals a close affiliation of the target story with a story not classified by Tang Kristensen as a story about ghosts but as a story about house elves (*nisse*), a category of supernatural beings not usually considered thematically related to ghosts: "When they got home, the farmhand was happy because now he'd gotten something to use for feed, and afterward *nis* [a house elf] could go and feed the animals just as much as he wanted to. Then they got another farmhand, and he didn't want to let him go on like that. But he got lifted up in his bed and all the way up to the rafters, so he lay there dead when people got up the next morning."[33]

The intersection between the target story and this latter story, likewise not

classified as a ghost story and completely lost to the researcher using earlier finding aids to find ghost stories in the corpus, suggests the broad applicability of the method. A researcher with a specific story in hand can find stories that relate to that story on a thematic level, while another researcher, without a particular target story in hand, can discover neighborhoods of "like-minded" stories. By adding or subtracting network modes or by testing different weights for edges in different projections of the overall meta-network, a researcher can quickly reorder the folklore meta-graph for ongoing discovery of information links not easily discovered through existing finding aids. This information can be added back into the now dynamic representation of the folklore corpus, allowing for more sophisticated analysis of stories and story patterns.

To the best of our knowledge no well-established criteria exist for judging the performance of such a computational system. Evaluation requires specification of a series of well-defined tasks that engage potential users in a "natural manner." The two tasks we defined earlier are typical of folklore researchers, and the system performs well in answering the folklorist's questions. We also imagine other natural tasks that might be used to evaluate and refine the system. Two that seem fundamental are "multiple key attribute searching" and "cluster labeling." Because folklore research is predicated on understanding the "context" of a cultural expressive form, the visual representation of the search results must be a projection onto the overall hierarchical-graph-map view of the corpus. Answers must be provided at human-interactive rates, a performance criterion our current system meets after the initial hierarchical graph map for the corpus has been computed.

Cluster labeling is an important task for researchers in light of this new interactive environment for the study of a folklore corpus. Hierarchical graph maps are presented to the researcher visually and include the string-labeled representations of the clusters in the hierarchy tree; the system computes the cluster labels. Researchers should also be able to provide textual feedback about the labels for their selected clusters, a feature we have not yet implemented. The system could then incorporate this researcher-generated feedback in its labeling algorithms. The more expert feedback the system incorporates, the more definitive and useful the cluster labels it could generate. Researchers' inability to provide a short textual description of a given cluster could suggest the need for revision of similarity measures and cluster methods. Such interaction between a folklore system and researchers should be central to improving the quality of this type of classification system.

Extending the System

The goal of a scalable multilingual folklore system necessarily relies on NLP. Folklore has been collected in many countries and languages and from many eras. Each language poses special challenges, from transcription and representation in machine-readable form to morphological and syntactic complexity.[a] Higher-level problems (such as named-entity detection and disambiguation, extracting structural elements, and sentiment detection) complicate the representation of texts significantly. Our system does not yet include narrative structures as attributes of a story, as explored by Elson and McKeown[14] and Finlayson,[16,17] Our own future work will include coding of narrative structure into the attribute set of a given story, as it should add additional insight into the target folklore corpora. Finally, our system does not include sentiment detection, though it would be a useful extension to be able to label texts or parts of texts as expressing doubt or uncertainty or characterizing the emotional aspects of characters in stories, but it lies beyond our current capabilities.

One remaining yet potentially rewarding challenge is the large-scale digitization of folklore archives around the world. Many folklore collections are based on recordings of spoken dialects or languages with poorly described or nonexistent written grammars; not surprisingly, many of them also have

inconsistent orthographies, at times varying from collector to collector. Digitization will require close collaboration among archivists, librarians, computer scientists, and domain specialists. Indeed, one reason we chose the Danish folklore archive as our initial target corpus was our team's representation in all these fields. Finally, many folklore collections exist almost entirely as audio and visual recordings, posing significant challenges for digital archiving and requiring significant advances in NLP technology before they are accessible to a computational folklore system.

The system's generalizability depends on these advances and developers' ability to automatically recognize and code narrative features and narrative structures. As part of the effort to extend our work to other corpora, domain experts could develop shallow ontologies for other tradition areas and other types of expressive forms. Shallow ontologies for, say, Korean or Mexican folklore would have to include types of supernatural beings beyond those currently described in the system; for example, the Korean *tokkaebi*, a blue gremlin with a particularly nasty disposition, and the legendary Mexican *chupacabra*, feared for its habit of attacking and sucking the blood of goats, would have to be incorporated into ontologies describing these traditions. The advantage of shallow ontology is that it is extensible at the level of specific manifestations (such as *nisse*, *tokkaebi*, *duende*, and *chupacabra*) while also gathering these supernatural beings into a single higher-level category as well. Researchers can further extend the ontological categories to describe other cultural expressive forms, from literary works (such as the novel)

a Our group is developing a system for automated lemmatization and morpho-syntactic tagging for Old Icelandic (NSF # BCS-0921123); modern Icelandic inflection is fully described, allowing for more rapid lemmatization of modern Icelandic texts.[5]

to narrative expression (storytelling) in social media.

Conclusion

In rural Denmark in the 19th century, stories would reflect the trouble farmers had with everything from ghosts to house elves. The trouble with house elves begins with their unpredictable nature, which, for folklorists, makes them difficult to track through the landscape of the story corpus. This exploration of the contours of computational folkloristics and our description of preliminary experiments in multimodal network classification for folklore corpora offer not only the possibility of being able to track house elves but promising directions for future work.

One troubling aspect of working with a large amount of humanities data is that scholars often cannot see the forest for the trees, to borrow a folk expression. Moretti spoke convincingly of distant reading, a corrective to the long-standing tradition in the humanities of very close reading.[25] Distant reading allows scholars to "see the forest," discovering patterns that might otherwise be obscured by too close attention to the detail of a text or performance; the same can be said of the methods outlined here.

Fortunately, with these computational methods, researchers are able to combine distant reading with close reading. In so doing, they can interrogate the relationship between folk expressive culture and the individuals who created and perpetuated these expressions in time and place. Computational folkloristics offers an opportunity to read and interpret culture in a more holistic fashion than ever before.

Acknowledgments

We wish to thank Nischal Devanur for help in processing the data. We also thank colleagues at the Institute for Pure and Applied Mathematics (UCLA), participants in Rice University's "Technology, Cognition, and Culture" lecture series, and Indiana University's "Networks and Complex Systems" symposium for comments and suggestions on earlier versions of this work. The research was funded by an American Council of Learned Societies Digital Innovation Fellowship and National Science Foundation grant IIS-0970179. Many of the ideas were developed at the National Endowment for the Humanities Institute for Advanced Topics in Digital Humanities through "Networks and Network Analysis for the Humanities" NEH grant HT5001609. The work of James Abello is partially supported by the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers University, Piscataway, NJ, "Special Focus on Algorithmic Foundations of the Internet" NSF grant #CNS-0721113, and mgvis.com (http://mgvis.com), New Jersey.

References

1. Abello, J. Hierarchical graph maps. *Computers & Graphics 28*, 3 (June 2004), 345–359.
2. Abello, J., van Ham, F. and Krishnan, N. ASK-GraphView: A large-scale graph visualization system. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (Sept./Oct. 2006), 669–676.
3. Bascom, W.R. *The Forms of Folklore: Prose Narratives.* University of California, Berkeley, 1965.
4. Bilgic, M., Licamele, L., Getoor, L., and Shneiderman, B. D-Dupe: An interactive tool for entity resolution in social networks. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology* (Baltimore, Oct. 31–Nov. 2). IEEE, Piscatway, NJ, 2006, 43–50.
5. Bjarnadóttir, K. Um Beygingarlýsingu íslensks nútímamáls, 2009; http://www.lexis.hi.is/kristinb/umBIN.html
6. Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (Jan. 2003), 993–1022.
7. Carley, K., Columbus, D., Bigrigg, M., and Kunkel, F. *AutoMap User's Guide 2010.* Technical Report. Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2010.
8. Child, F.J. and Kittredge, G.L. *The English and Scottish Popular Ballads.* Houghton Mifflin Company, Boston and New York, 1882.
9. Choe, I.-h. *A type index of Korean folktales.* Myong Ji University Publishing, Seoul, 1979.
10. Christiansen, R.T. *The Migratory Legends.* Suomalainen Tiedeakatemia, Helsinki, 1992.
11. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 6 (Sept. 1990), 391–407.
12. Dundes, A. From etic to emic units in the structural study of folktales. *Journal of American Folklore 75*, 296 (Apr.–June 1962), 95–105.
13. El-Shamy, H.M. *Folk Traditions of the Arab World: A Guide to Motif Classification.* Indiana University Press, Bloomington, IN, 1995.
14. Elson, D.K. and McKeown, K.R. A tool for deep semantic encoding of narrative texts. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and Fourth International Joint Conference on Natural Language Processing of the AFNLP* (Suntec, Singapore, Aug. 3). Association for Computational Linguistics, Stroudsburg, PA, 2009, 9–12.
15. Eskeröd, A. *Årets äring. Etnologiska studier i skördens och julens tro och sed.* Nordiska Museets handlingar 26. Håkan Ohlssons boktryckeri, Lund, Sweden, 1947.
16. Finlayson, M.A. Deriving narrative morphologies via analogical story merging. In *New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy*, D. Gentner, K. Holyoak, and B. Kokinov, Eds. (Sofia, Bulgaria, July 24–27). New Bulgarian University Press, Sofia, 2009, 127–136.
17. Finlayson, M.A. Collecting semantics in the wild: The Story Workbench. In *Proceedings of the AAAI Fall Symposium on Naturally Inspired Artificial Intelligence* (Arlington, VA, Nov. 7–9). AAAI Press, Menlo Park, CA, 2008, 46–53.
18. Fogaras, D. and Rácz, B. Scaling link-based similarity search. In *Proceedings of the 14th International Conference on the World Wide Web* (Chiba, Japan, May 10–14). ACM Press, New York, 2005, 641–650.
19. Gasperin, C., Gamallo, P., Agustini, A., Lopes, G., and Lima, V.D. Using syntactic contexts for measuring word similarity. In *Proceedings of the ESSLLI Workshop on Semantic Knowledge Acquisition and Categorisation* (Helsinki, Aug. 13–17). ESSLLI, Helsinki, 2001, 18–23.
20. Hofmann, T., Schölkopf, B., and Smola, A.J. Kernel methods in machine learning. *Annals of Statistics 36*, 3 (June 2008), 1171–1220.
21. Jiang, X. and Tan, A.-H. Mining ontological knowledge from domain-specific text documents. In *Proceedings of the Fifth IEEE International Conference on Data Mining* (Houston, Nov. 27–30). IEEE Computer Society, Los Alamitos, CA, 2005, 665–668.
22. Krohn, K. *Die Folkloristische Arbeitsmethode.* H. Aschehoug & Co., Oslo, 1926.
23. Le Cam, L.M. and Yang, G.L. *Asymptotics in Statistics: Some Basic Concepts.* Springer-Verlag, New York, 2000.
24. McCallum, A.K. *MALLET: A Machine Learning for Language Toolkit.* University of Massachusetts, Amherst, 2002; http://mallet.cs.umass.edu
25. Moretti, F. Conjectures on world literature. *New Left Review 1* (Jan.–Feb. 2000), 54–66.
26. Paneva, D., Rangochev, K., and Luchev, D. Ontological model of the Knowledge in Folklore Digital Library. In *Proceedings of the Fifth HUBUSKA Open Workshop on Knowledge Technologies and Applications*, T. Urbanova, I. Simonics, and R. Pavlov, Eds. (Kosice, Slovakia, May 31-June 1). HUBUSKA, Kosice, Slovakia, 2007, 47–55.
27. Pedersen, B.S., Nimb, S., and Trap-Jensen, L. DanNet: Udvikling og anvendelse af det Danske WordNet. Nordiske Studier i leksikografi 9 (2008), 353–370.
28. Propp, V.I.A. *Morfologija Skazki.* Academia, Leningrad, 1928.
29. Sanfilippo, A., Tratz, S., Gregory, M., Chappell, A., Whitney, P., Posse, C., Paulson, P., Baddeley, B., Hohimer, R., and White, A. Ontological annotation with WordNet. In *SemAnnot 2005: Proceedings of the Fifth International Workshop on Knowledge Markup and Semantic Annotation*, S. Handschuh, T. Declerck, and M.-R. Koivun, Eds. (Galway, Ireland, Nov. 7). Ceur-WS, Aachen, Germany, 2005, 27–36.
30. Schölkopf, B. and Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, 2002.
31. Siikala, A.-L. *Interpreting Oral Narrative.* Suomalainen Tiedeakatemia, Helsinki, 1990.
32. Tang Kristensen, E. *Danske sagn, som de har lydt i folkemunde, udelukkende efter utrykte kilder. Ny Række.* Woels Forlag, Copenhagen, 1928–1939.
33. Tang Kristensen, E. *Danske sagn, som de har lydt i folkemunde, udelukkende efter utrykte kilder.* Aarhus Folkeblads Bogtrykkeri, Aarhus, 1892–1901.
34. Tangherlini, T.R. Legendary performances: Folklore, repertoire and mapping. *Ethnologia Europaea 40*, 2 (2010), 103–115.
35. Tangherlini, T.R. *Interpreting Legend: Danish Storytellers and Their Repertoires.* Garland Publishing, New York, 1994.
36. Thompson, S. *Motif-Index of Folk-literature. A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends.* Indiana University Press, Bloomington, 1955–1958.
37. Uther, H.-J. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson.* Suomalainen Tiedeakatemia Helsinki, 2004.

James Abello (abello@dimacs.rutgers.edu) is research professor in the Center for Discrete Mathematics and Theoretical Computer Science of Rutgers University, Piscataway, NJ.

Peter M. Broadwell (broadwell@library.ucla.edu) is a Council on Library and Information Resources postdoctoral fellow in the Digital Initiatives Department of the Charles E. Young Research Library at the University of California, Los Angeles.

Timothy R. Tangherlini (tango@humnet.ucla.edu) is a professor of folklore in the Scandinavian Section and the Department of Asian Languages and Cultures at the University of California, Los Angeles and a fellow of the American Folklore Society.