# The Technological Developments of the Dutch Folktale Database (1994–2016)

*Theo Meder*
Meertens Instituut
theo.meder@meertens.knaw.nl

Abstract
*In 1994 the Dutch Folktale Database started as a stand-alone database and went online in 2004. Since 2016, and after two major projects, all kinds of metadata can be added automatically and semi-supervised: languages, names, keywords, summaries, subgenres, motifs and tale types. To this end, the database went over to a new platform called Omeka that fits the needs of many databases in the humanities, and which can handle all kinds of plug-ins. The following techniques have been used: n-grams, language detection, named entity recognition, keyword extraction, summarization, bag of words, machine learning and natural language processing. Furthermore MOMFER, a search engine for motifs has also been added. The interpretation of data is facilitated by new means of visualisation: geographical maps, timelines, a network of similar tales, and word clouds. Since the database meets the requirements of Dublin Core, it can be connected to similar databases or a data harvester. Recently, a Trans-Atlantic Digging into Data application has been made to build a harvester called ISEBEL: Intelligent Search Engine for Belief Legends. The harvester should be able to search in a Dutch, Danish and German database simultaneously. Later, other databases can be added.*

Keywords
*Folktale; database; visualisation; harvester; e-humanities*

Estudis de Literatura Oral Popular, núm. 5, 2016, 45–69 | DOI: 10.17345/elop201645-69
ISSN: 2014-7996 | http://revistes.urv.cat/index.php/elop

Resum

*L'any 1994, la base de dades holandesa de contes populars va començar com una base de dades independent i es va posar en línia el 2004. Des de l'any 2016 i després de dos projectes importants, tots els tipus de metadades es poden afegir de manera automàtica i semisupervisada: idiomes, noms, paraules clau, resums, subgèneres, motius i tipus de contes. Amb aquesta finalitat, la base de dades va analitzar una nova plataforma anomenada Omeka que s'adapta a les necessitats de moltes bases de dades en les humanitats, i que pot gestionar tot tipus de connectors. S'han utilitzat les tècniques següents: n-grames, detecció del llenguatge, reconeixement d'entitats nombrades, extracció de paraules clau, resum, bossa de paraules, aprenentatge automàtic i processament de llenguatge natural. A més de MOMFER, també s'ha afegit un motor de cerca de motius. La interpretació de dades es facilita amb els nous mitjans de visualització: mapes geogràfics, línies de temps, una xarxa de contes similars i núvols de paraules. Com que la base de dades compleix els requisits de Dublin Core, es pot connectar a bases de dades similars o a un recol·lector de dades. Recentment, s'ha creat una aplicació de mineria de dades transatlàntica per construir un recol·lector anomenat ISEBEL: Intelligent Search Engine for Belief Legends (motor de cerca intel·ligent de llegendes de creences). El recol·lector ha de ser capaç de buscar en una base de dades holandesa, danesa i alemanya simultàniament. Més endavant s'hi poden afegir altres bases de dades.*

Paraules clau

*rondalla; base de dades; visualització; recol·lector; humanitats en línia*

Rebut: 29/09/2016 | Acceptat: 18/10/2016

Looking at the photo below, we can conclude that the media landscape of 1961 was much easier to survey than today's. At the time, I was one year old. We had radio and black-and-white television, which commanded all our attention.

Radio and TV

As to writing by hand, the dip pen and the inkwell were fighting a losing battle against the biro in the Netherlands. Binary numbers and computers did exist, but it would take another twenty years or so before they would come within reach of the average consumer. Professional writers used typewriters.

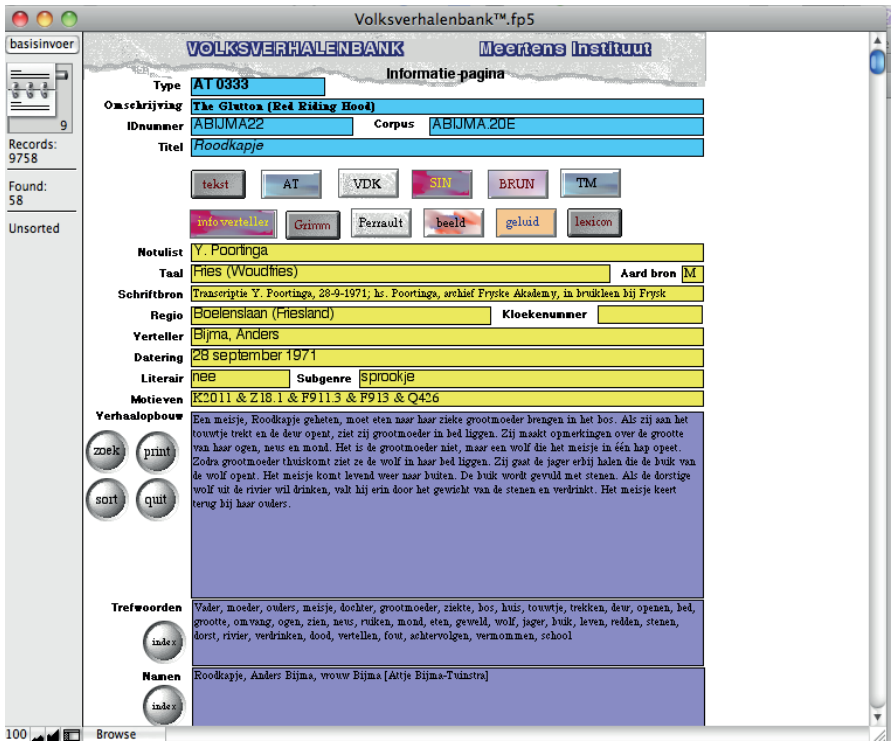
Olivetti

When I got a PhD position at Leiden University in 1986, the electronic typewriters by Olivetti had just been purchased on a large scale. They were noisy, bulky contraptions, capable of remembering an entire line, which they could erase again with the help of a correction ribbon, should the user wish so. When I was typing out my first scholarly article on just such an Olivetti machine, my colleague Bert van Selm walked in and exclaimed: "Don't tell me you are going to write your dissertation on that thing". When I asked him, rather sheepishly, what else I could do, he replied: "Buy yourself a personal computer. It is quite an investment at the start, but it will give you years of pleasure". And an investment it was for a mere student. I had to take out a loan to buy a Commodore PC10-II, with a monochrome green screen, no hard disk and two 5 ¼-inch floppy disk drives. First you had to load the operating system MS DOS, then the text editor WordPerfect before you could feed the lower disk drive with the floppy which was to contain your freshly-written dissertation. And I did it: I managed to finish my 707-page PhD thesis on the Middle Dutch performing poet ("sprookspreker") Willem van Hildegaersberch.


Commodore PC10-II · Sprookspreker in Holland

And oh, the convenience of it all: you could correct your text over and over again without having to insert yet another new blank sheet of paper. Not to mention the cutting and pasting to your heart's content. Faced by the PhD committee, my supervisor joked: "He never managed to find the delete button". He then introduced a new rule for all other PhD students: dissertations may be no thicker than 200 pages. Incidentally, in the next few years, the Olivettis at all Dutch universities would face a fate of dust gathering.

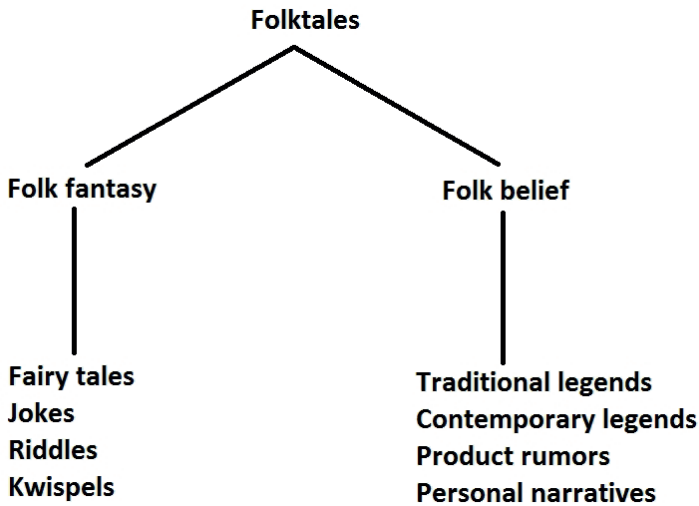
After that first PC, developments accelerated. Initially, it was believed that the CD-ROM would become the storage medium of the future, but when e-mail in the nineties was joined by the Internet, it became clearer and clearer that we were about to witness the construction of a worldwide web of interlinked servers and computers. While we were using Netscape, developments never stopped: soon there would be laptops, smartphones and tablets.




Database in FileMaker Pro

When I was employed by the Meertens Institute in 1994, one of my assignments was to build a database of Dutch folktales, following (more or less) the model of the Dutch Song Database, which had existed for several years by then. The first version of this database was built by myself, in FileMaker Pro, on a Mac. It was a database with texts, linked to a large amount of metadata. This version was still stand-alone and offline. It could only be used by myself and by people visiting the Meertens Institute.

The Dutch Folktale Database was specifically meant for folktales: tales that have been in oral circulation for varying periods of time among groups of people. These tales from the oral tradition also had to qualify as a form of narrative art. This requirement was meant to prevent the database from becoming an "oral history" collection, which would lead to a massive influx of just any kind of narrative material. The tales we were looking for were the ones (often) included in the well-known catalogues as internationally-known tales: fairy tales, traditional legends, jokes, contemporary legends, product rumours, riddles, kwispels[1] and personal narratives (like, for example, stories that are repeatedly told in families). In other words, we are dealing with folktales that can be classified either as folk fantasy (fiction) or as folk belief (non-fiction).




Folktales schema

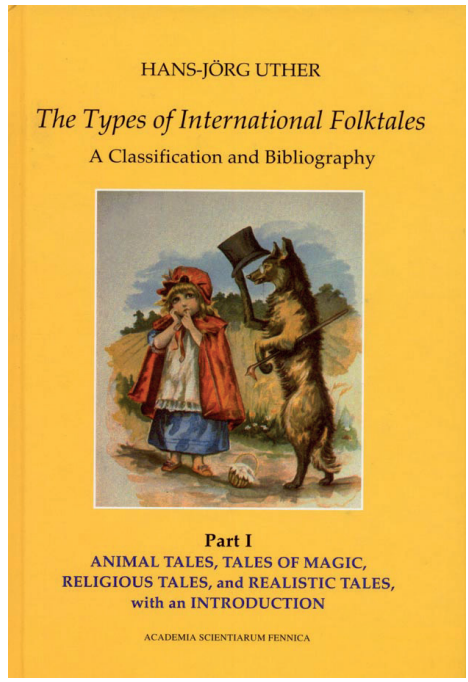
By 2004, there were enough folktales (approx. 20,000) in the database to go online, so it was decided to build a web version in MySQL. Since then, the Dutch Folktale Database has been accessible at <www.verhalenbank.nl>. In 2016, after the initiation of two long projects (FACT and Tunes & Tales), a new version was launched, using the Omeka platform. Omeka is pre-eminently suitable as a database format in the humanities, if only because it follows the guidelines of Dublin Core. This enables data to be exchanged with other (international) databases, provided these are also Dublin Core-based.

The core elements of the database are still the (variants of) folktale texts, but there is also a great deal of metadata: who the narrator is, who the collector was, where the stories were told, where they are set, what the source is, when they were recorded, what the subgenre is, in what language or dialect they appear, which names occur in them, which keywords could be attributed, what the summaries of the tales look like, the special features that are worth mentioning, which motifs occur in the tales, and what the tale type is? Other metadata included are

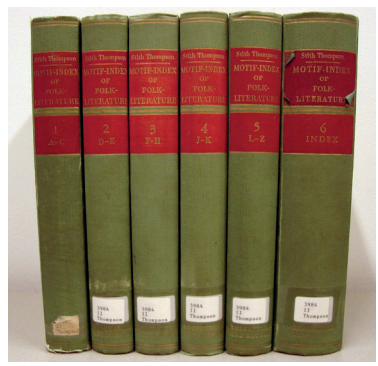
1. For this subgenre, see Meder & Burger (2006).

copyright details and explicitness of content (if applicable). The database may also contain images, and audio and video files.

The types and motifs are specific to the research of folktales. A particular folktale in all its possible variants is called a type. The most commonly used catalogue is the one by Aarne-Thompson-Uther: *The Types of International Folktales* (Uther 2004). This contains roughly 2200 catalogued international folktales (for example, ATU 333 *Little Red Riding Hood*).




*The Types of International Folktales* (Uther 2004)



*Motif-Index* (Thompson 1955–58)

Each tale type consists of one or several motifs: the smaller narrative building blocks. In his *Motif-Index*, Stith Thompson included no fewer than 45,000 motifs (Thompson 1955–58). Little Red Riding Hood, to mention just one example, contains the following motifs (and possibly more):

J21.5 – *"Do not leave the highway"*

K2011 – *Wolf poses as "grandmother" and kills child*

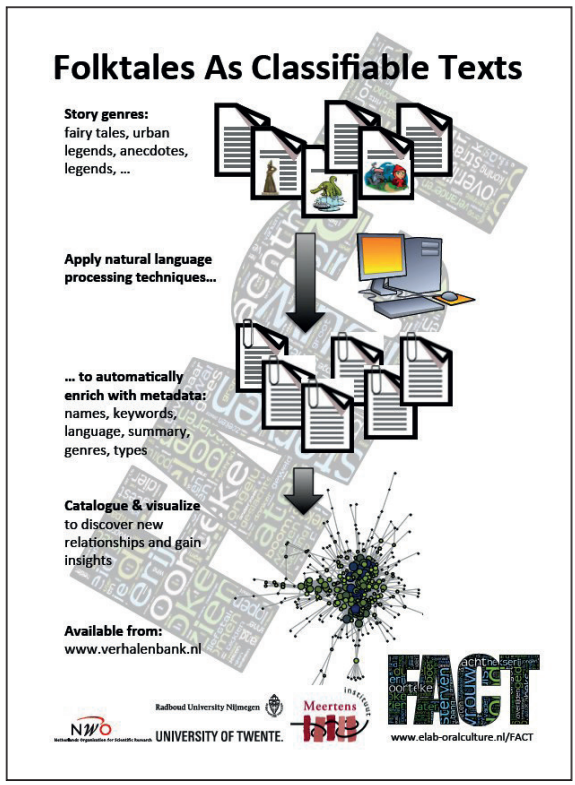
Z18.1 – *What makes your ears so big – To hear the better, my child, etc.*

F911.3 – *Animal swallows man (not fatally)*

F913 – *Victims rescued from swallower's belly*

Q426 – *Wolf cut open and filled with stones as punishment*

It is not so much including stories in the database that takes up an enormous amount of time and effort, but rather the adding of the metadata. I was keen to automate some of these tasks. Mariët Theune of the University of Twente and myself made an application for the project FACT: Folktales as Classifiable Texts, which was granted in 2012. The team was enriched with three technical experts from the University of Twente: Dolf Trieschnigg (postgraduate, 3 years), Dong Nguyen (PhD student, 4 years) and Iwe Muiser (scientific programmer, 4 years). The poster below displays what we had in mind.




Folktales as classifiable texts

Would it not be nice to upload folktale texts and instruct the computer to attribute the metadata pertaining to language, names, keywords, subgenre, summary, motifs and tale type by itself? FACT did not get involved with motifs, for this was a job to be tackled in another project, Tunes & Tales (see below). The automated attribution of the other metadata proved to be successful, albeit to varying degrees. Consequently, the attribution process must be semi-supervised at all times. The following techniques were implemented: n-grams, language detection, named entity recognition, keyword extraction, a summarization tool, bag of words, machine learning and natural language processing. The systems are subject to a continuing learning process based on new input. The computer is capable of identifying languages and dialects, especially in those cases in which there has been previous and frequent contact. It recognises Standard Dutch, Frisian, Middle Dutch and 17th-century Dutch almost without fault. Dutch city dialects, however, are frequently mistaken for Standard Dutch, which is not so very odd. The recognition of names is flawless too. The attribution of keywords is somewhat more problematic, since this system is mostly based on Standard Dutch. The software extracts the main nouns and verbs from the text, but if these are in Frisian or Middle Dutch, they will not be found. The software also has a problem attributing the more abstract keywords that do not actually occur in the text: one would expect the term "drowning" to be linked to the keyword "death", the word "soldier" to the keyword "military" and the term "priest" to the keyword "clergyman". Making this link manually (rather than by computer) will prove most successful. The technique of summarization mainly consists of deleting irrelevant sentences and retaining relevant sentences. One can opt for a sliding scale here. But, again, a summary in Middle Dutch or Frisian will not be very helpful. It requires a second analysis on top of the first one. The recognition of tale types is heavily dependent on the training material: if the computer has encountered Little Red Riding Hood, say, twenty times, it will not have any difficulty recognizing a twenty-first version of the same tale. If it has seldom encountered a version, it is likely to make a random guess. The computer does a reasonably good job of recognizing the subgenres, but it has not quite mastered distinguishing between modern legends and traditional legends yet. The more challenging assignments for people appear to be equally challenging for the software.

The scientific programmer took care of some visualisations, which I will elaborate on later. The PhD student published a dissertation on variability in folktales and posts on social media (Nguyen 2017).

All in all, the FACT project has engendered many important publications,[2] it made a considerable contribution to TweetGenie[3] (part of the TINPOT project) and it has led to a special form of valorisation, which I will discuss below.

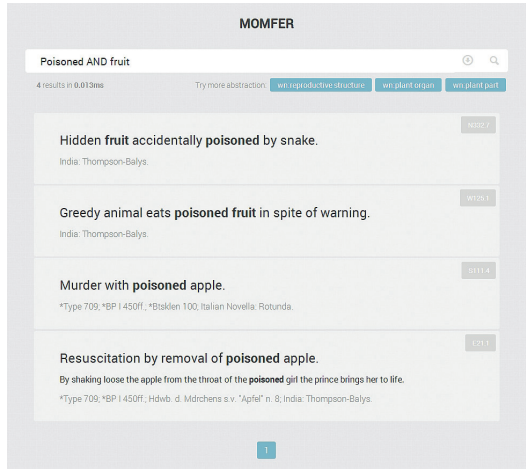
---

2. See Meder (2010, 2012, 2014); Meder, Nguyen & Gravel (2016); Meder, Karsdorp, Nguyen *et alii* (2016); Muiser, Theune & Meder (2012); Nguyen, Trieschnigg, Meder *et alii* (2012); Nguyen, Trieschnigg & Theune (2013); Nguyen, Gravel, Trieschnigg *et alii* (2013); Nguyen, Trieschnigg & Theune (2014); Nguyen, Trieschnigg & Meder (2014); Nguyen (2017); Trieschnigg, Hiemstra, Theune *et alii* (2012); Trieschnigg, Nguyen & Theune (2013); Trieschnigg, Nguyen & Meder (2013).

3. A web app that can predict the gender and age of Twitter users on the basis of tweets in Dutch; see <www.tweetgenie.nl> [Last access: October 2016].


The Tunes & Tales Team

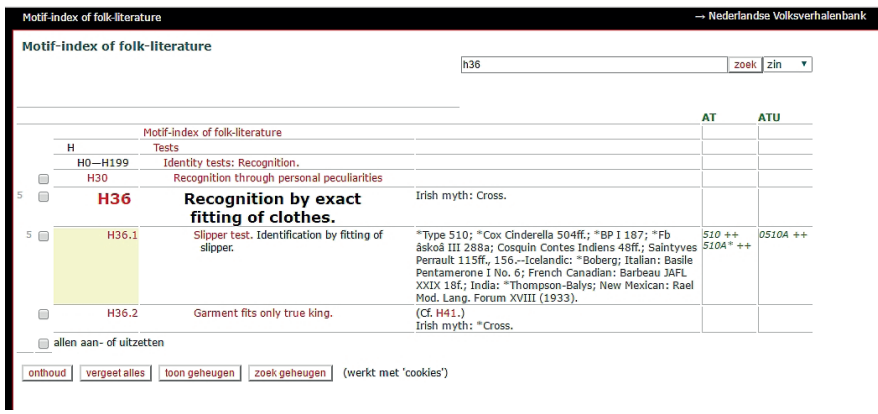
A second research project to receive a grant in 2012 – from the Royal Netherlands Academy of Arts and Sciences – was Tunes & Tales. The applicants for this project were Louis Grijp (1954–2016) and Theo Meder. The project encompassed research into the variability and stability of tales and melodies in the oral tradition, or, in other words, research into motif sequences in linguistic and musical performances. PhD student Folgert Karsdorp (left in the photo) was responsible for automating the *Motif-Index* and implementing automatic recognition of the motifs in the tales. He developed software that could recognise actors and actions in Dutch texts and match them to English motifs from the *Motif-Index*. He managed to do this by means of a parser, machine learning, natural language processing and WordNet. This software was integrated into the Dutch Folktale Database, and also put online independently under the name MOMFER (see www.momfer.ml). This motif search engine contains all 45,000 motifs from Thompson's *Motif-Index*, but takes a flexible approach to search queries thanks to WordNet. It allows searches on a higher, more abstract level: if the search term is "animal", WordNet will run through all specific animals. The example below illustrates how a Boolean search into "poisoned AND fruit" also resulted in the hit "murder with poisoned apple" (Motif S111.4). The results are ranked in order of importance.




MOMFER

Folgert Karsdorp wrote a computational-evolutionary PhD thesis on variability in transmission (Karsdorp 2016). The Tunes & Tales project also led to valuable publications in the field of folktale research.[4]

In the same period, Dirk Kramer, who was temporarily employed by the Meertens Institute, created yet another database based on the *Motif-Index*.[5] The advantage of this index is that it maintains the hierarchical structure applied to the *Motif-Index*. If a particular motif occurs in the Dutch Folktale Database as well, it is referenced by means of a link. It is also possible to search for motif numbers, like in the example below: H36 *Recognition by exact fitting of clothes*. This is followed by motif H36.1, *Slipper test. Identification by fitting of slipper*, well known from the Cinderella fairy tale. The motif number is marked in light yellow here, which means that it occurs in the Dutch Folktale Database.




Database based on the *Motif-Index*

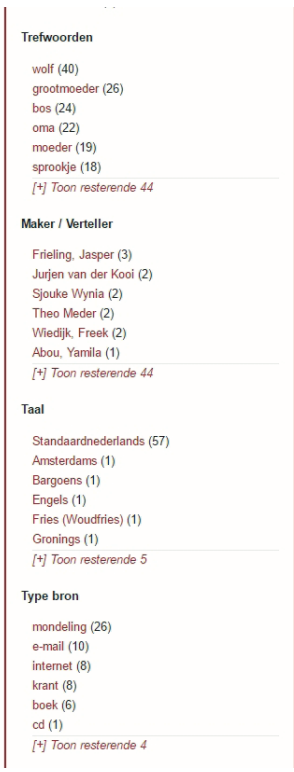
4. See Karsdorp, Van Kranenburg, Meder *et alii* (2012*a*, 2012*b*); Karsdorp (2013); Karsdorp & Van den Bosch (2013); Karsdorp, Van der Meulen, Meder *et alii* (2015*a*, 2015*b*); Meder, Karsdorp, Nguyen *et alii* (2016).

5. See <http://www.dinor.demon.nl/motif/index.html?index> [Last access: October 2016].

All this has led to considerable improvement in the functionality of the Dutch Folktale Database for researchers and those entrusted with data input. Hopefully, it will also allow the database to better serve its twofold purpose:

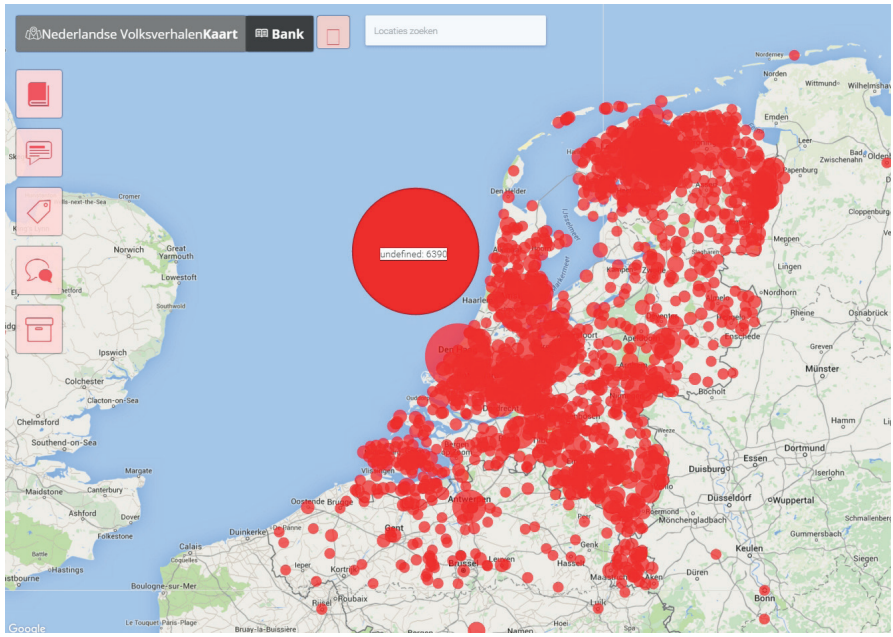
1. as a digital archive of collective intangible heritage
2. as an instrument for (e.g.):
   - comparative research in time and space
   - research into variability and stability
   - research into narrative patterns and structures

It is to be expected that the developments described above will result in the faster input of data, which will be to the advantage of visitors too. Analysis has shown that there are many visitors who do not access the Dutch Folktale Database by the welcome page, but by the search query on Google or some other search engine. The general user population includes fellow scholars, students writing their theses, journalists eager to check if a particular story classifies as a contemporary legend, (semi-)professional storytellers looking for repertoire, relatives of deceased narrators, and folktale lovers in general. The search terms used make clear that many users do not know how to perform queries or what to look for. On the welcome page, one can do a Google-style search, with Boolean operators like AND, OR, and NOT, or with a literal quote between quotation marks. The search capabilities of Omeka have improved considerably since the addition of Solr, which creates indexes. Furthermore, the system includes all sorts of advanced and very advanced search options. It is possible, for instance, to request the display of tales told in Rotterdam and within a 10-mile radius. For the average user, much simpler instructions will suffice, though. One search term will often produce quite a number of hits. At present, the Dutch Folktale Database automatically offers concrete filters that might be useful for reducing the number of hits. Left, you will find the example of Little Red Riding Hood: people can now reduce the set of hits by adding a keyword like "grandmother", by clicking on the name of a narrator, by selecting a specific language, by choosing a particular source (e.g. only oral recordings), et cetera.




Filters

Another useful instrument is a map of the Netherlands with dots indicating how many tales are linked to a particular place. People are inclined to begin their search by entering their place of birth or residence. Also in this case, the set of search results can be reduced. The big dot in the North Sea, by the way, represents the number of tales whose place of origin or setting is unknown (or at least unclear).

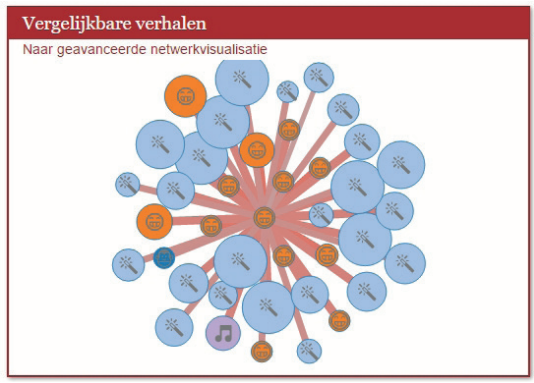



Map of the Netherlands

Another visualisation added to the tale pages are the maps showing where a tale was told, and what place the tale is about (if known). The example below refers to a tale which was told in Zutphen, but is set in Amsterdam.


Place of narrating, place of action.

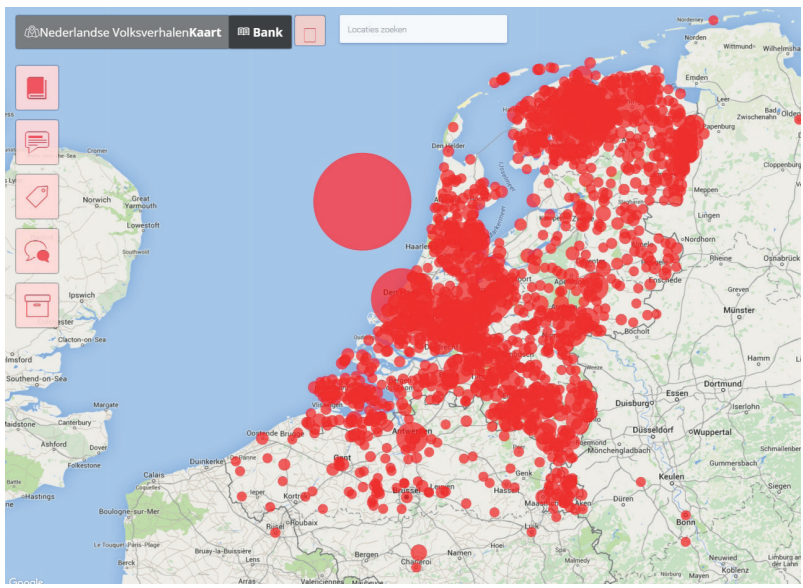
Tale pages also visualise the possible presence of related tales. This relationship is determined on the basis of the tale type number and/or the keywords. Each genre has its own colour and symbol. The example used here is Little Red Riding Hood once more: light blue circles with magic wands indicate fairy tales, orange circles with smileys refer to jokes, and purple circles with musical notes represent songs. The dark blue circle refers to an encyclopaedic entry containing background information. All the information required can be displayed by clicking the circle.




Little Red Riding Hood and related versions

All tale pages including metadata can be translated into dozens of languages thanks to Google Translate. These translations are far from perfect, but they are the best thing available at the moment.

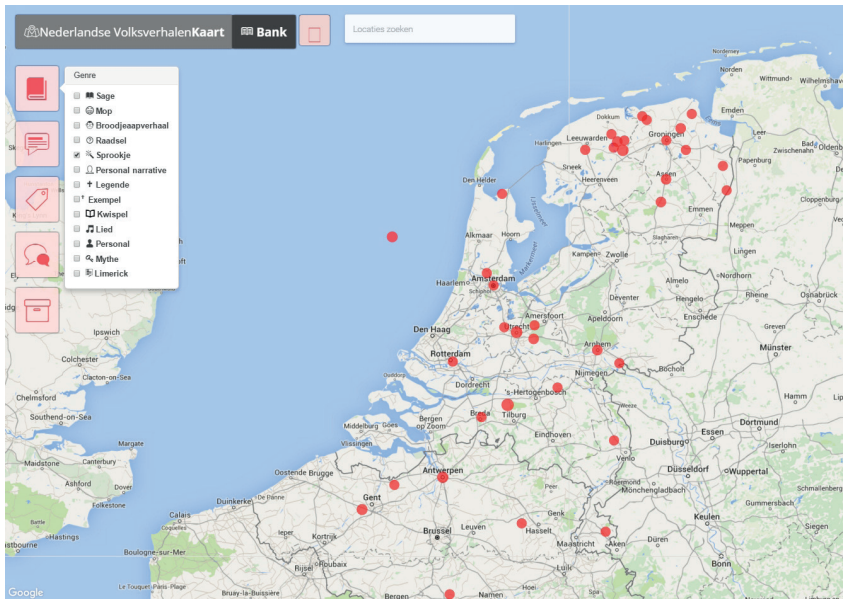
As of now, every search can be visualised. Doing so reveals, for example that, in the past, the Netherlands had many tales about witches, to mention just one example.



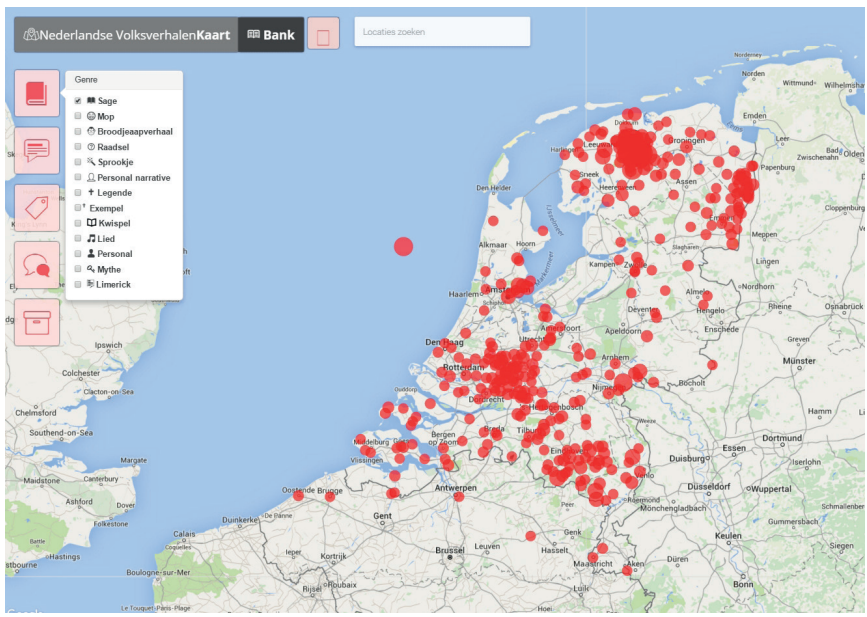

Tales about witches

The map is almost entirely red. The presence of gaps (and their locations) can be explained by the fact that material from these areas has not yet been entered, not that there are no tales. And again, the outcome can be filtered. Many people think witches mainly feature in fairy tales, but the map tells us otherwise. There are not many dots for witches in fairy tales.

Legends, however, are full of witches, as can be seen below.

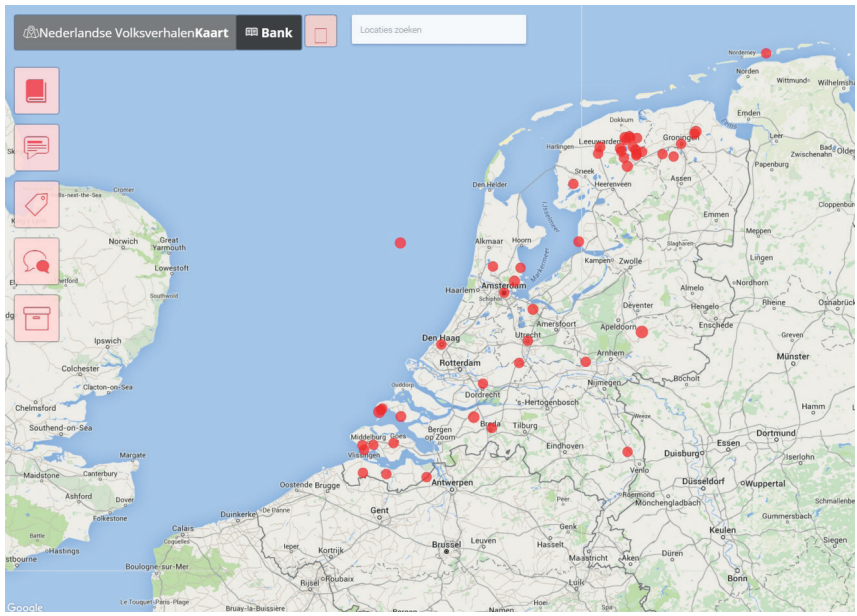



Witches in fairy tales




Witches in legends

The hotbeds are, again, an indication of intensive collecting activities rather than anything else. For whatever reason there are no apparent concentrations of witch tales. Searches for geographical distribution provide some more interesting maps. When the search term "mermaid" is used, a pretty clear coastline pattern of mermaid tales pops up (which is not too surprising, of course).

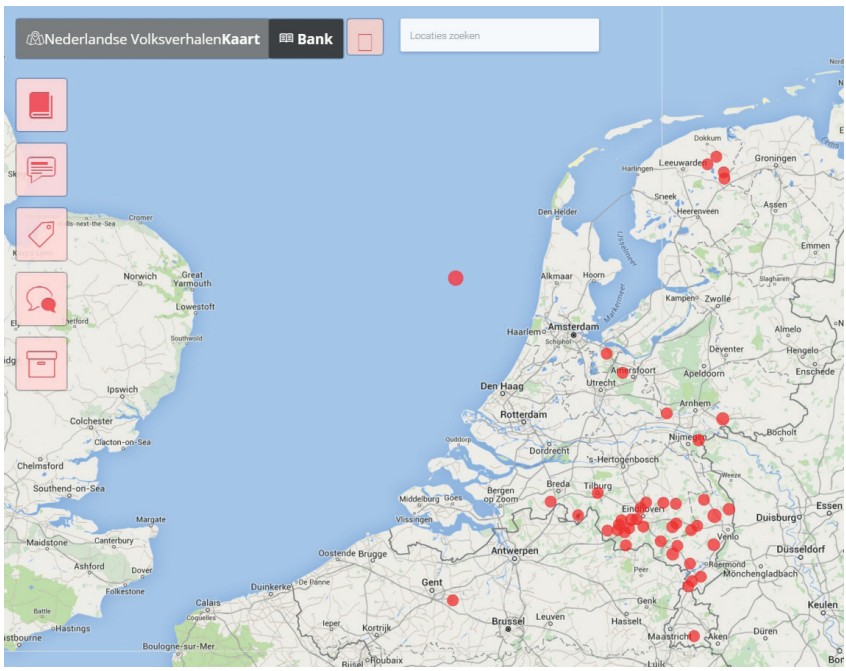


Mermaid

As one last example, I would like to present the map of the distribution of tales about the "vuurman" (fire man). In short, the fire man is a burning soul returning to earth to correct a crime. During his life, he displaced boundary posts to enlarge his property. Consequently, he is punished in his afterlife for this fraud. He is often seen dragging around a boundary stone aimlessly, unable to remember where to put it. In the Bible, people displacing boundary stones are warned repeatedly about their sin and threatened with damnation.[6] The map shows that the fire man occurs almost exclusively in the Dutch provinces of North Brabant and Limburg, the two Roman Catholic regions in the south of the country. In the north (province of Friesland), there is a small cluster too, but when one takes a closer look at the tales, it becomes apparent that in these areas, the fire man features only as a bogeyman: a fearsome creature ensuring the children's safe return home at night-time. The boundary stone tale only occurs in the south. The distribution is due to certain differences of opinion between Catholics and Protestants. Catholics believe in a heaven and a hell one cannot escape from, but allow for a purgatory in between. There is room for transfers here. The soul may end up going to hell anyway, but if enough time is spent, by the people on earth, praying and burning candles for the soul of the deceased, he or she can go to

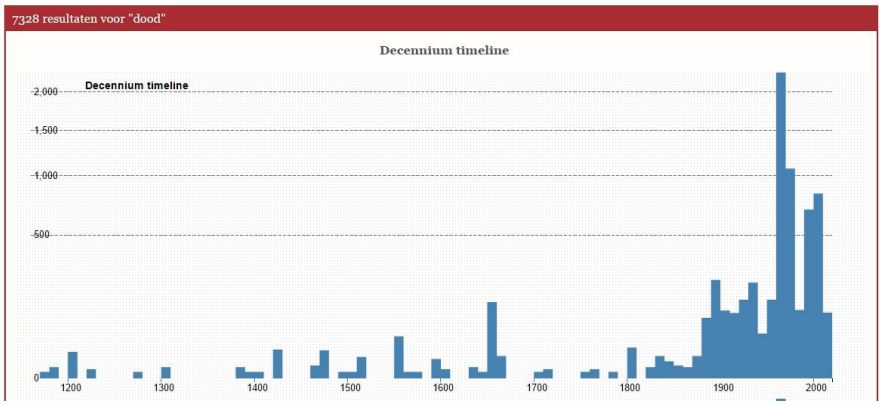
6. See Job 24:2, Proverbs 22:28, Deuteronomy 19:14 and, notably, Deuteronomy 27:17.

heaven after all. In the Catholic view, burning souls can – in special cases – return to earth in an attempt to remove the sin. Protestants do not believe in purgatory. In their view, returning from the hereafter is an impossibility. Ergo, there is no place for the fire man tale in Protestant circles.




Vuurman (fire man)

Another form of visualisation is the timeline. Dominant keywords in the Dutch Folktale Database – like "death"– can be marked off on a timeline (7,328 hits for "death"). This also gives us an insight into the collecting practice: most folktales were recorded in the 19th and 20th centuries, which explains why the keyword "death" reaches its peak around this time. But at the same time, a keyword like "death" makes clear that this motif occurs in folktales from the Middle Ages on.




Timeline

Yet another form of visualisation is the creation of a word cloud on the basis of a specific set of folktales (also in this case, Solr supports the rather limited possibilities of Omeka). A word cloud of all 44,000 folktales in the database leads to the result in the picture below.


Word cloud of all folktales in the Dutch Folktale Database

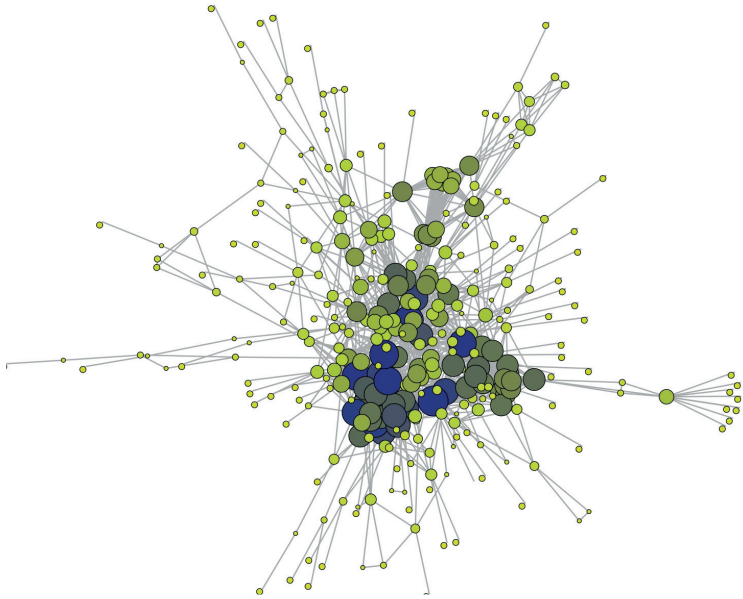
The bigger words are the more dominant keywords. The nature of the keywords makes clear that (pessimistic and supernatural) legends prevail over all other genres: death, man, woman, dying, omen, second sight, magic, sorcery, witch, haunting, devil, fear, etc. A word cloud of the (10,000) jokes in the Dutch Folktale Database presents an entirely different picture.


Word cloud of jokes in the Dutch Folktale Database

The two dominating concepts in the jokes are (not surprisingly) "sex" and "stupidity". Quite a large number of jokes, in some way or other, deal with sexual taboos, either in the heterosexual sense or in relation to homosexuality, masturbation, fornication, bestiality, rape, incest, etc. The prevalence of the stupidity motif can easily be explained from the infinite number of Dutch jokes about the ignorance of Belgians (or rather Flemings) and blondes. The next keywords in line here are "man", "woman", "money", "death", "farmer", "minister", "clergyman", "genitals" and "drinking".

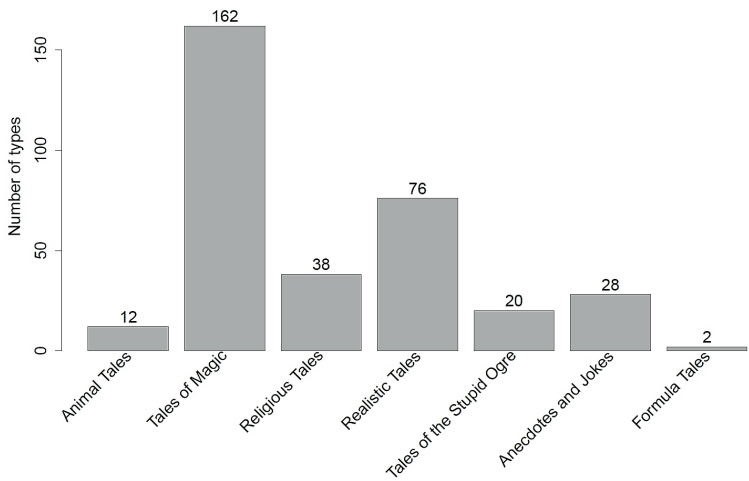
In the last few years, a great deal of computational research has been done into the Dutch Folktale Database and, for instance, catalogues. One of the important assumptions in the research into motifs was that, in principle, folktales (of the types listed in the ATU catalogue) freely interchange motifs. Meanwhile, we have been impelled to adjust this assumption. Firstly, we must acknowledge that the 45,000 motifs in the *Motif-Index* are insufficient to describe all folktales and sequences of motifs in folktales. Basically, the ATU catalogue includes only a small number of distinctive motifs and leaves out many other possibilities. In other words, it would be possible to distinguish many more motifs (as Baughman did [1966]], by adding another 10,000 motifs, which also proved insufficient). Secondly, there appear to be many short tales that consist of a single distinctive motif, which does not appear in any other tales. And thirdly, we must conclude that only certain categories of fairy tales interchange motifs frequently (like the wolf with the stones in his belly at the end of Little Red Riding Hood and The Wolf and the Seven Kids/Goats).[7] The following visualisation pictures the interchange of motifs.

7. ATU 333 and ATU 123 resp., with the motif Q426, *Wolf cut open and filled with stones as punishment*.



Interchange of motifs

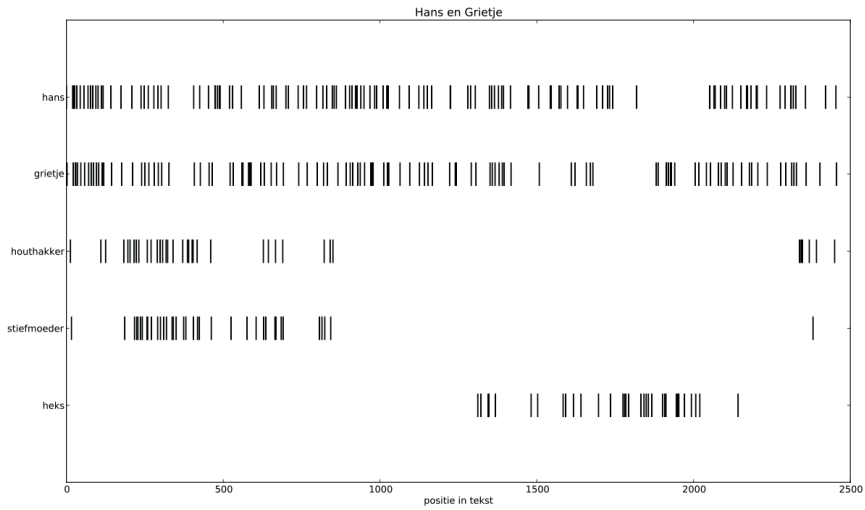
Tales containing only one distinctive motif that is not interchanged are left out in this visualisation, which would otherwise be full of small, separate islands. As we can see, there are quite a number of motifs that appear in only one other tale, if at all. The core of the interchange appears in the centre. The larger and darker the nodes, the more motifs are interchanged with other folktales (based on the information in the ATU catalogue). Focussing on this last group of tales to determine the segment they fall into, we end up with two subcategories.




Subcategories

The most frequent interchange of motifs obviously takes place in "Tales of Magic" and "Realistic Tales" (which are in essence the same kinds of fairy tale, although "realistic tales" do not contain magic elements). Anyone intending to do further research into the interchange of motif sequences should focus on these two subcategories.

Such research was performed by Folgert Karsdorp, who also focussed on the issues I am about to elaborate on below. In an attempt to find (universal) patterns, actors in folktales could be visualised in the form of a bar code. Using this technique with a version of Hansel and Gretel provides the following result.




Actors in a version of Hansel and Gretel

The above analysis appeared automatically by first having a parser distinguish the different actors and actions, and even identify every "he" and "she". Hansel and Gretel are the heroes, of course, so they are introduced first and are present throughout the plot, from beginning to end. Even Gretel's heroic deed can be identified, and recognised as a thickening in the bar code. After the heroes, the father and the stepmother are introduced. They are absent from the middle part of the plot. The father returns at the end, while the stepmother is just said to have disappeared (without any further explanation). The unmistakable opponent, the witch, does not appear on the scene until the second half of the tale, and even her downfall is more or less visible. Follow-up research into patterns is obvious. For example, do all other versions of Hansel and Gretel show a similar bar code? But also: do other fairy tales and folktales show similar patterns in terms of plot development with different actors in specific roles? This research would enable us to shed light on the question about the universality of human storytelling culture.

Let me get back to the valorisation project that ensued from FACT: the "SagenJager" (TaleTracker in English, not a literal translation). The Netherlands Organisation for Scientific Research (NWO) was prepared to provide additional funds if our scholarly endeavours were to develop something of social relevance.

It was decided to produce a so-called TaleTracker based on the Dutch Folktale Database. This mobile website contains walking and cycling trails for tourists from one folktale to the next. At present, there are 16 different trails: walking trails of approximately 3 to 6 miles, and cycling trails of some 20 miles.

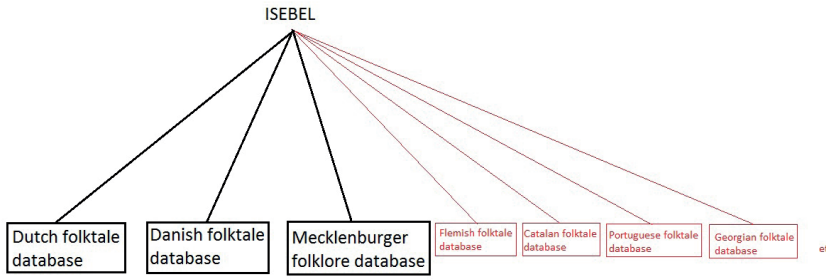

SagenJager

On smartphones or tablets, the tourist is represented by a blue dot that moves along the trail. For more information, see <www.sagenjager.nl>. Every signpost along the trail is linked to a tale with explanatory comments, in writing or audio. Additional information can be found through a link to the source text in the Dutch Folktale Database. The "SagenJager" has received ample media attention, and the trails are quite popular.[8]

Finally, we envisage a future of extended international cooperation. This may include the development of a harvester, which could perform searches in several folktale databases at the same time, with the requirement that these other databases are also Dublin Core-based. Recently, a large international research application was submitted for the development of ISEBEL: an Intelligent Search Engine for Belief Legends. The idea is that – to begin with – three databases are interlinked by the harvester ISEBEL: the Dutch Folktale Database, the Danish Folktale Database of Tim Tangherlini (UCLA) and the Northeast German folklore database of Christoph Schmitt (WossiDia, University of Rostock). This will lead to the visualisation of a wide coastline from the Netherlands through Denmark to

8 The cycling trail around Oostermeer (Friesland) includes a printed map, over 7000 copies of which were handed out in one year.

Mecklenburg-Vorpommern. Other folktale databases (from Flanders, Catalonia, Portugal, Georgia, etc.) could be added to the search capabilities of ISEBEL at a later stage.




ISEBEL

To conclude, we submitted an application to develop tools named "SagenChecker" (legend checker) and "TrustTheSource". These tools could be of particular interest to journalists who wish to prevent contemporary legends from finding their way into their reports. The detection tool will be fed with a large number of digital contemporary legends through machine learning. It should also develop into a self-learning system. An alarm should go off every time a journalist starts writing a contemporary legend. To complement this, "TrustTheSource" is supposed to check the reliability of posts on social media by tracing the source of the post and weighing the value of accounts, followers, tweets, retweets and replies. We intend to get this project off the ground in collaboration with the University of Groningen, Leiden University, Hanzehogeschool Groningen (University of Applied Sciences) and the University of Twente.

In the account above, I have reported the technological developments concerning the Dutch Folktale Database in the period from 1994 to 2016. In fact, the archiving and research activities related to the database are still in their infancy. We are on the brink of a period in which advanced computational research in the humanities is feasible and worthwhile.

## References


Baughman, Ernest W. (1966): *Type and Motif-Index of the Folktales of England and North-America*. The Hague: Mouton & Co.

Karsdorp, Folgert (2013): "Het is groen en leeft nog lang en gelukkig: Classificatie van volksverhaalgenres op basis van formules". *Tijdschrift voor Nederlandse Taal- en Letterkunde* vol. 129, no. 4: 274–288.

— (2016): *Retelling Stories: A Computational-Evolutionary Perspective*. Nijmegen.

Karsdorp, Folgert; Antal van den Bosch (2013): "Identifying Motifs in Folktales Using Topic Models". In *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning*, pp. 41–49.

Karsdorp, Folgert; Peter van Kranenburg; Theo Meder; Antal van den Bosch (2012*a*): "Casting a Spell: Identification and Ranking of Actors in Folktales". In F. Mambrini; M. Passarotti; C. Sporleder (eds.): *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*. Lisbon: Edições Colibri, pp. 39–50.

Karsdorp, Folgert; Peter van Kranenburg; Theo Meder; Dolf Trieschnigg; Antal van den Bosch (2012*b*): "In search of an appropriate abstraction level for motif annotations". In *Computational Models of Narrative workshop (LREC 2012)*. Istanbul.

Karsdorp, Folgert; Marten van der Meulen; Theo Meder; Antal van den Bosch (2015*a*): "MOMFER: A Search Engine of Thompson's Motif-Index of Folk Literature". *Folklore* no. 126: 37–52.

— (2015*b*): "Animacy Detection in Stories". In M. Finlayson; B. Miller; A. Lieto; R. Ronfard (eds.): *Proceedings of the Workshop on Computational Models of Narrative (CMN'15)*. OpenAccess Series in Informatics 45. Atlanta, pp. 82–97.

Meder, Theo (2010): "From a Dutch Folktale Database towards an International Folktale Database". *Fabula* no. 51 (1/2): 6–22.

— (2012): *Avonturen en structuren: op zoek naar de bouwstenen van volksverhalen*. Amsterdam: Meertens Instituut.

— (2014): "The Folktale Database as a Digital Heritage Archive and as a Research Instrument". In Meyer Holger; Christoph Schmitt; Stefanie Janssen; Alf-Christian Schering (eds.): *Corpora ethnographica online. Strategien der Digitalisierung kultureller Archive und ihrer Präsentation im Internet*. Münster: Waxmann, pp. 119–128.

Meder, Theo; Peter Burger (2006): "«A Rope Breaks. A Bell Chimes. A Man Dies». The Kwispel: A Neglected International Narrative Riddle Genre". In P. Catteeuw; M. Jacobs; S. Rieuwerts *et alii* (eds.): *Toplore. Stories and Songs*. Trier: Wissenschaftlicher Verlag Trier, pp. 28–38.

Meder, Theo; Folgert Karsdorp; Dong Nguyen; Mariët Theune; Dolf Trieschnigg; Iwe Everhardus; Christiaan Muiser (2016): "Automatic Enrichment and Classification of Folktales in the Dutch Folktale Database". *Journal of American Folklore* vol. 129, no. 511: 78–96.

Meder, Theo; Dong Nguyen; Rilana Gravel (2016): "The Apocalypse on Twitter". *Digital Scholarship in the Humanities* no. 31 (2): 398–410.

Muiser, Iwe; Mariët Theune; Theo Meder (2012): "Cleaning up and Standardizing a Folktale Corpus for Humanities Research". In *Second Workshop on Annotation of Corpora for Research in the Humanities.* Lisbon, pp. 63–74.

Nguyen, Dong (2017): *Text as Social and Cultural Data: A Computational Perspective on Variation in Text.* Enschede.

Nguyen, Dong; Dolf Trieschnigg; Theo Meder; Mariët Theune (2012): "Automatic Classification of Folk Narrative Genres". In *First International Workshop on Language Technology for Historical Text(s) (KONVENS 2012).* Vienna, pp. 378–382.

Nguyen, Dong; Rilana Gravel; Dolf Trieschnigg; Theo Meder (2013): "«How Old Do You Think I Am?»: A Study of Language and Age in Twitter". In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Publications, pp. 439–448.

Nguyen, Dong; Dolf Trieschnigg; Mariët Theune (2013): "Folktale Classification Using Learning to Rank". In P. Serdyukov *et alii* (eds.): *Proceedings of ECIR* 2013. Springer, pp. 195–206.

— (2014): "Using Crowdsourcing to Investigate Perception of Narrative Similarity". In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, pp. 321–330.

Nguyen, Dong; Dolf Trieschnigg; Theo Meder (2014): "TweetGenie: Development, Evaluation, and Lessons Learned". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations.* Dublin, pp. 62–66.

Thompson, Stith (1955-58): *Motif-Index of Folk-literature. A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*. 6 vols. Bloomington: Indiana University Press.

Trieschnigg, Dolf; Djoerd Hiemstra; Mariët Theune; Franciska de Jong; Theo Meder (2012): "An Exploration of Language Identification Techniques for the Dutch Folktale Database". In *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage (LREC 2012)*. Istanbul, pp. 47–51.

Trieschnigg, Dolf; Dong Nguyen; Theo Meder (2013): "In Search of Cinderella: A Transaction Log Analysis of Folktale Searchers". In *Proceedings of the First ACM SIGIR Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage.* Dublin.

Trieschnigg, Dolf; Dong Nguyen; Mariët Theune (2013): "Learning to Extract Folktale Keywords". In *7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH*. Sofia, pp. 65–73.

Uther, Hans-Jörg (2004): *The types of international folktales. A classification and bibliography based on the system of Antti Aarne and Stith Thompson*. 3 vols. Folklore Fellows' Communications 284, 285, 286. Helsinki: Suomalainen Tiedeakatemia.