The Folktale Database as a Digital Heritage Archive and as a Research Instrument

Theo Meder

Abstract: When I first started building a Dutch Folktale Database at the Meertens Instituut in 1994, it was a rather simple stand-alone database, not only containing folktale texts, but meta data with information about the source, the narrator, the date, the genre, motifs and types as well. In 2004 a web version of the Dutch Folktale Database went online, and today the database contains over 42,000 folktales: fairy tales, jokes, traditional and contemporary legends. Part of the folktale material comes from the library and the archive of the Meertens Instituut. Smaller parts come from recent fieldwork and from digital media – websites, forums, blogs and even Facebook and Twitter. Since most of the folktales from the Meertens archive have never been published, the online database is an ideal digital archive for immaterial heritage, in particular Dutch narrative culture in past and present.

The Folktale Database can be used as an instrument for comparative research, for instance to study narrative variation in certain tale types over time, or between different regions. It could get even more interesting if we were able to perform international queries in several folktale databases. Such databases are being developed in Flanders, Catalonia, Portugal, Georgia and Mecklenburg. It would be a good idea to build a harvester that can retrieve information from an international set of databases.

The Dutch database could easily contain 100,000 folktales, if it was not such an arduous job to add all of the metadata by hand. In order to solve this problem, we found funding for a program called FACT: Folktales As Classifiable Texts. The tools are supposed to recognize language and genre, extract names, add keywords and write summaries. A PhD student will work on a tool to identify folktales as tale types from the Aarne-Thompson-Uther catalogue and other catalogues. Furthermore the PhD student is going to look for new ways to classify folktales by computer, using all sorts of clustering techniques. A second project that has been funded is called Tunes & Tales, dealing with melodies and folktales as sequences of motifs. Main focus is motif variation in oral tradition. A PhD student will go into the question what a motif is, research how a computer program can learn how to identify motifs, how some motifs tend to stay in place and other motifs move around or disappear, and finally how strings of motifs can form the DNA of a story and even of a whole cluster of stories. A post-doc researcher will construct a model that explains variation in oral transmission in both tunes and tales.

The world we live in is rapidly becoming digital, and the same thing can be said about science. At the very start of my own scientific career, I wrote my first articles on a typewriter. Soon, however, as a PhD student, I started writing my dissertation on a Commodore computer without a hard disk. All it had was two floppy disk drives. Just when we thought that CD-ROMs would be the data carriers of the future, the Internet came online. Today, data can be studied and exchanged 24/7 from all over the world, thanks to servers and the Internet, websites and databases, PCs, laptops, tablets and smartphones.

The amount of data online ... We do not count it in kilobytes, megabytes or gigabytes anymore, but rather in terabytes and petabytes. That sounds like a lot, but as yet only a small percentage of our libraries and archives have been digitized. Most of our sources still need to be consulted on paper or parchment. In the humanities, we have only just started to build our scientific databases, which requires a lot of funding. Digitalization is happening here today, for instance, in Rostock, at the Wossidlo archive, with all this wonderful folklore and folktale material from Mecklenburg-Western Pomerania in particular. In fact, the movement towards e-humanities is becoming prominent enough to organize a symposium like "Corpora Ethnographica Online" about it.

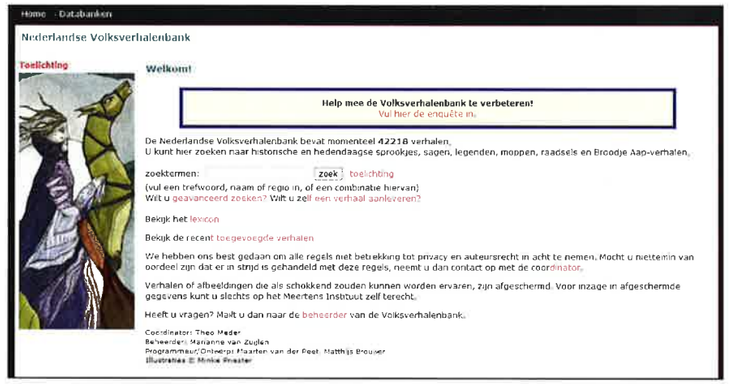

Fig. 1: Screenshot of the online Dutch Folktale Database.

Being a folktale researcher, I am especially interested in online folktale databases around the world. I have my own online Dutch Folktale Database at the Meertens Instituut in Amsterdam. But how many folktale databases are there world wide? As far as I can see, not that many. Of course, there is this wonderful international project called Project Gutenberg, which includes many editions of fairy tales and legends, but which is – with all due respect – nothing more than an online library for e-books. Worldwide there are many websites to be found with books being scanned and OCRed, but I would not call them folktale databases. If I were looking for the fairy tale of the fisherman, his discontent wife and the enchanted fish, searching for "ATU 555" would not help me a bit. Still, ATU 555 is the internationally acknowledged type number for "The Fisherman and his Wife" in the catalogue by Antti Aarne, Stith Thompson and Hans-Jörg Uther called "The Types of International Folktales". Supposing we would exclude plain text sites and book sites like the Project Gutenberg or Google Books (to name just another big one), and search for folktale databases that enable you

to use the ATU tale type system, how many websites would then remain? Only a few, I am afraid. These are the databases I know of:

1. The Dutch Folktale Database (by Theo Meder)[1] containing some 42,000 folktales, including fairy tales, traditional legends, saints' legends, jokes, riddles, and urban legends.
   A database that very much resembles the Dutch one is
2. The Flemish Folktale Database (by Stefaan Top) containing some 48,000 folktales.[2] This database mostly contains legends, and not so many fairy tales. Unfortunately, initiator Stefaan Top is retired and currently, there is no replacement at the University of Leuven. So the database has been "parked" on a server and left unattended.
   Then there are
3. The Archive of Portuguese Legends (by Paulo Correia and Isabel Cardigos) containing some 3,500 tales.[3]
4. RondCat: Catalan Folktales Search Engine (by Josep M. Pujol and Carme Oriol Carazo) containing some 6,000 folktales.[4] As far as searchable metadata are concerned, this Catalan database makes a very good impression.
5. The Georgian Folklore Database (by Elguja Dadunashvili)[5]: again a rich source with some 29,000 tales and related folklore material.
   And hopefully there will soon be
6. The Wossidlo Folktale Database (by Christoph Schmitt)[6].

There may be more folktale databases in the world, but because of my limited knowledge of languages I cannot always tell. I could not tell if such databases are searchable on metadata, nor if they contain enough folktale material to make a search worthwhile.[7] So five searchable folktale databases up and running ... that is not a lot. If you know more folktale databases, please let me know.

A look at the databases makes clear that the initiatives come from small countries, or from smaller regions within larger countries. Large countries like Ger-

1 http://www.verhalenbank.nl.
2 http://www.volksverhalenbank.be.
3 http://www.lendarium.org.
4 http://www.sre.urv.cat/rondcat.
5 http://titus.fkidg1.uni-frankfurt.de/database/folkarch/query.htm#inpform. Also see: http://folktreasury.weebly.com. Cf. the article of Elguja Dadunashvili in this volume.
6 Cf. the article of Holger Meyer, Alf Christian Schering and Christoph Schmitt in this volume.
7 For instance Terry Gunnell's "Sagnagrunnur", see: https://notendur.hi.is/terry/database/sagnagrunnur.htm; Pavel Kats' "WikiProverbs", see: http://www.wikiproverbs.com; Jeanmarie Rouhier-Willoughby's "Russian Folk Religious Imagination", see: http://rfri.rch.uky.edu.

many or France[8] are not represented (yet). Apart from the fact that "smaller" folklore repertoires are cheaper to digitize, it almost looks as if regions and small countries also feel more need to preserve their own (narrative) culture against the threat of larger and dominant national cultures. It almost looks like regions and smaller countries fear for the loss of their cultural identity in a united Europe, in a globalizing and standardizing world, or in a world where radical Islam has ambitions to expand. By displaying their narrative heritage these regions and small countries seem to underline they have a language and a culture of their own, worth preserving and researching.

Furthermore, it seems like the databases have a slight preference for legends over – for instance – fairy tales. This means that the focus lies on folk belief rather than on folk fiction and fantasy. Perhaps legends are considered to be more determinative for the local culture and identity of a specific group, whereas fairy tales are considered to be a more international genre ...

When I first started building a Dutch Folktale Database at the Meertens Instituut in 1994, it was a rather simple stand-alone database, not only containing separate folktale texts, but metadata with information about the source, the narrator, the date, the genre, motifs and types as well. Researchers should be able to see *when* a specific story was told, *where* it was told and by *whom*. Furthermore, researchers should be able to perform a genre query – searching for either fairy tales or legends or jokes or riddles – and researchers should be able to establish in what type of source – written or oral – a specific tale can be found. The main reason to build a digital database was, and is, to compare and understand variants of the same story in time and space. Some parts of folktales remain the same, whereas other parts start to change, for instance because morals and tastes may alter over time. Conducting comparative studies to research variation in oral (and written) tradition is one of the core businesses of the folk narrative researcher.

Just to give you a brief and simple example: there is a distinct difference in how the fairy tale of Little Red Riding Hood is told in the 17th and in the 19th century. In the 17th-century version of Charles Perrault the little girl gets eaten by the wolf, and that is the end of the story. Being eaten is the punishment for Red Riding Hood's imprudence: she should have been more careful, and now it is too late. End of story. In the 19th-century version of the Grimm Brothers, Little Red Riding Hood gets a second chance. The belly of the wolf is cut open and out come Red and her granny! So even though she made a huge mistake, Red Riding Hood is forgiven and awarded a second chance. You may remember

[8] In July 30, 2012, I received an e-mail from Marike van der Horst announcing that a group of amateurs and professionals will be starting a French Folktales Database soon. However, this is a regional initiative, not a national one: several French departments will join forces and put their folktales online. There will be no support from the French government.

that in the Grimm story Red is confronted with another wolf and she does not make the same mistake twice. The two versions of the story make clear how morals concerning punishment and forgiveness have shifted over time.

When dealing with folktales in the oral tradition, one can always find variation. One of the reasons for this is – of course – our incapability to literally remember and retell a story (unless we memorize it like a song), so that we need to improvise. But the other reason is because creative narrators tend to change stories to their liking (and the liking of their audience) on purpose. So two other features needed to be added to my folktale database: (1) the ability to identify similar stories and distinguish between different stories and (2) the ability to show how stories consist of a sequence of motifs or narrative building blocks. This is where the types and motifs kicked in. I already mentioned the catalogue called "The Types of International Folktales" by Aarne, Thompson and Uther.[9] This catalogue assigns numbers to fairy tales and anecdotes, so for instance Little Red Riding Hood is internationally known as ATU 333. For all kinds of legends – including modern urban legends – other catalogues are needed, but in the end a fair number of folktales can be identified as internationally well-known stories.

Every story consists of smaller building blocks we call motifs. There is an extensive catalogue of 45,000 motifs by Stith Thompson called the "Motif-Index of Folk-Literature".[10] The motif 'Animal Swallows Man (Not Fatally)', for instance, is catalogued as number F911.3. A motif is considered to be a smaller narrative element in a tale having a power to persist in tradition. On the level of motifs, folktales vary a lot over time and place: motifs can trade places, they can double or triple, disappear, be substituted et cetera. All meaningful variation takes place at motif level. It is, however, a tremendous job to find all these motifs and to add them to the database manually. For the sake of speed, I stopped adding motifs after a while. Because of new research today, we started to add the motifs again. In a small experiment we recently performed, five scholars made an overview of all the motifs present in just one version of the fairy tale of Cinderella. The story consisted of only 124 sentences, and altogether we managed to find no less than 68 small and large motifs. In average, this means one motif on every two sentences, which was a bit of a surprise to me.[11] On the other hand, there were only three motifs all of us had found and that we all considered to be rather essential:

[9] Uther, Hans-Jörg: The Types of International Folktales. A Classification and Bibliography (= FFC 284–286). Helsinki 2004 (3 vols.).

[10] Thompson, Stith: Motif-Index of Folk-Literature. A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends. Copenhagen 1955–1958 (6 vols.).

[11] See http://www.verhalenbank.nl/detail_volksverhalen.php?id=EFTDG01.

S31 Cruel stepmother
F823.2 Glass shoes
H36.1 Slipper test

Back to the beginning: slowly but surely, the Dutch Folktale Database expanded with new stories and new modules, containing folktale catalogues, editions of the Grimms' and of Perrault's tales, some occasional audiovisual files, and a lexicon explaining the history and meaning of all kinds of tale types to the general public. In 2004 a web version of the Dutch Folktale Database went online, and – as mentioned earlier – today the database contains over 42,000 folktales. The database is being visited by scholars, students, journalists, professional storytellers, relatives of deceased narrators and lovers of folktales in general.

As far as I am concerned, the Dutch Folktale Database serves two major purposes:

1. a digital heritage archive
2. a digital research instrument

First the digital heritage archive. A large number of the digitized stories stem from the library and the archive of the Meertens Instituut. Smaller parts come from, for instance, recent fieldwork and from digital media – websites, forums, blogs and even Facebook and Twitter. Since most of the folktales from the Meertens archive have never been published on paper, the online database would be an ideal digital archive of immaterial heritage, in particular of Dutch narrative culture in past and present. It contains all the folktales that were collected around 1900 by philologist Gerrit Jacob Boekenoogen and physician Cornelis Bakker (about a thousand stories). At the moment we are busy digitizing and adding metadata to a huge collection compiled by the Meertens Instituut in the 1960s and 1970s. No less than 32,000 (mainly) legends were collected by 24 collectors all over the Netherlands. We owe half of this collection to the Frisian collector Dam Jaarsma who recorded 16,000 folktales. First the tales were collected for the sake of a folklore atlas, later on for the sake of folk narrative research. These collections are being complemented by Dutch folktales from the Middle Ages until today, stemming from medieval romances, chapbooks, jest books, almanacs, chronicles, children's books, newspapers, magazines, up to television broadcasts and digitally-born tales. Although we have already built an extensive database, it goes without saying that there is still a lot of work to be done. The Dutch Folktale Database has the ambition to grow into the ultimate and representative digital archive on Dutch narrative culture, online accessible 24/7 all over the world for free. We are working on an English version of the

database and on an automatic translation of the database content. Still, the database wants to be more than just a collection.

Fig. 2: Screenshot of the metadata about a Frisian version of Little Red Riding Hood in the Dutch Folktale Database.

Secondly, the Dutch Folktale Database can be used as a research tool, more particularly as an instrument for comparative research, for instance to study narrative variation in certain Dutch tale types over time, or variation between different regions. It could get even more interesting if we were able to perform international queries in several folktale databases. It would be a good idea to build a harvester that is able to retrieve information from an international set of databases. Several initiatives have been made to standardize data exchange in the humanities, like the Dublin Core Metadata Initiative (DCMI), the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and CLARIN. It would be even more convenient to develop more international standards in folkloristics, for instance not just a type-index of fairy tales and anecdotes, but a much-needed international type-index of legends as well. And a type-index of modern genres too, like contemporary legends and jokes. Furthermore, consensus needs to be established on the matter of motifs. Which narrative building blocks will be acknowledged as motifs? Are we going to use the Motif-Indexes

of Thompson and Baughman[12]? Will distinguishing some 55,000 motifs be enough or too much? Are these the memes our stories are built of? Further international standardization is essential for comparative research to succeed. Comparative studies in variation can tell us a lot about the development, the meaning and the purpose of stories in past and present society. It can even tell us more about the very mechanics of oral tradition and about the structural characteristics of narratives.

More than 42,000 folktales in a database sounds like an awful lot, but it actually is not. The Dutch database could easily have contained 100,000 folktales by now, had it not been such an arduous job to add all of the metadata by hand. In order to solve this problem, we sought (and found) funding for a project called FACT: Folktales As Classifiable Texts. In this project, a postdoc and a programmer are going to build tools that will automatically process the input of folktales. The tools are supposed to recognize language, extract names, add keywords and write summaries. So if I put the text of a folktale into the computer, the computer should be able to tell me if it is Dutch or Frisian, if there is a Red Riding Hood involved or a Flying Dutchman, if the story is about wolves or ghosts, and how the plot goes. A PhD student will work on a tool that is meant to distinguish genres and identify folktales as tale types from the Aarne-Thompson-Uther catalogue and other catalogues. So again, if I give the computer a folktale, it should be able to tell me if it is a fairy tale or a legend, and if it is ATU 333 or ATU 555. Furthermore the PhD student is going to look for new ways to classify folktales by computer, using all sorts of clustering techniques. After all, our traditional classification system originated a century or more ago and was never designed for computer automation. Maybe the computer can distinguish completely different and equally meaningful classes of tales.

A second project that has been recently funded is called Tunes & Tales. This project deals with both melodies and folktales as sequences of motifs. Main focus is motif variation in oral tradition. A PhD student will go into the question what a motif is, research in what way a computer program could learn how to identify motifs, see how some motifs tend to stay in place whereas other motifs move around or disappear, and finally understand how strings of motifs can constitute the DNA of a story or even of a whole cluster of stories. One of the difficult tasks is, of course, to teach the computer how to analyse stories, by using tools like natural language processing and machine learning. Tools need to learn how to find both named and non-named entities in stories (most wolves do not have names), how to recognize the actions in the plot, how to identify the scenery (are we in a castle or a farmhouse?) and how to piece all of these elements together. Once this is possible, the computer can assign existing plot

[12] Baughman, Ernest W.: Type and Motif-Index of the Folktales of England and North America (= Indiana University Folklore Series no. 20). The Hague 1966.

motifs to tale characters in action. A postdoc researcher will construct a model that explains variation in oral transmission in both tunes and tales. Hopefully this research will give us more insight into how stories are built and how they mutate, how narrative grammar works and if there is, perhaps, such a thing as a universal narrative grammar (just like some linguists suppose there is a universal grammar for language).[13]

$$\text{A B C} \uparrow \text{D E F G} \; \frac{\text{H J I K} \downarrow \text{Pr-Rs}^{n} \text{ L}}{\text{L M J N K} \downarrow \text{Pr-Rs}} \; \text{Q Ex T U W}^{*}$$

Fig. 3: The structuralist formula describing the morphology of Tales of Magic by Propp.

Formalist and structuralist research into narrative grammar has been done before, by Vladimir Propp and Algirdas Greimas[14] to mention the most famous names. As you may know, Propp constructed a formula to capture the basic morphology – or if you like: syntax – of a specific group of fairy tales called the Tales of Magic. Greimas designed a general scheme into which all basic oral narrative should fit.

Neither Propp nor Greimas were ever able to use computers to analyse large amounts of text and metadata. Today computer techniques have finally become advanced enough to perform research tasks in the field of the analysis of the grammar of tales.

[13] The Meertens sub-departments DOC Volksverhaal and DOC Lied, and the projects of the Dutch Folktale Database, the Dutch Song Database, FACT, Tunes & Tales (along with Louis Grijp), COGITCH and Oral Transmission have, together with several partners outside the Meertens merged into a so-called eLab Oral Culture. See http://www.elab-oralculture.nl/. The cooperating external partners for FACT and (Tunes &) Tales are the University of Twente (Mariët Theune, Franciska de Jong, Djoerd Hiemstra), the University of Nijmegen (Antal van den Bosch), and the Fryske Akademy (Arjen Versloot). The employees for FACT are: Dong Nguyen (PhD student), Iwe Muiser (programmer), Dolf Trieschnigg (postdoc), and for (Tunes &) Tales: Folgert Karsdorp (PhD student) and Peter van Kranenburg (postdoc).

[14] Propp, Vladimir: Morphology of the Folktale. Second edition. Austin 2009; Greimas, Algirdas J.: Réflexions sur les modèles actantiels. In: Greimas, Algirdas J.: Sémantique structurale: recheche de méthode. Paris 1966, pp. 172–191; Greimas, Algirdas J.: Narrative Grammar: Units and Levels. In: Modern Language Notes 86 (1971) 6, pp. 793–806.

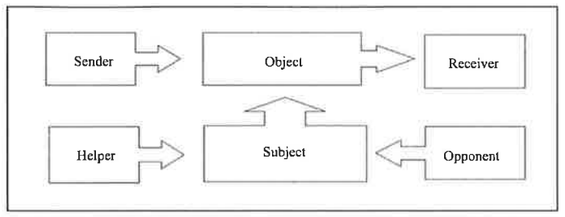

Fig. 4: The scheme of basic narrative structure by Greimas.

Thanks to research of this kind, the Dutch Folktale Database will be more than just a digital museum of stories. The large data set can be put to use for fundamental research as well, on a scale that has never been seen before, with sophisticated tools that have never been used before. If we allow our databases to join forces and build an international harvester, I am convinced that we can – no, will – make a meaningful leap forward in philological, ethnological and comparative science.

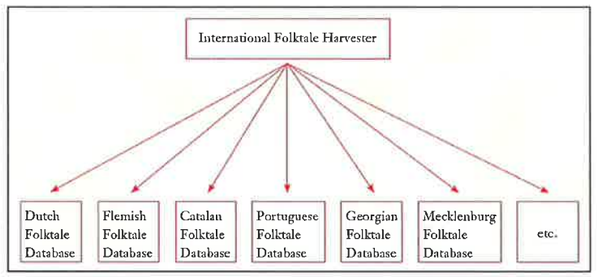

Fig. 5: Plan for an International Folktale Harvester enabling simultaneous queries in multiple regional and national folktale databases.

Webplattform der vergleichenden Erzählforschung
Ein virtueller Raum für die internationale Kooperation

Elguja Dadunashvili

Abstract: This article presents the "Web Platform of Comparative Folk Narrative Research" as an instrument of international fairy-tale research. The platform will gradually transform the Aarne-Thompson tale type index into a flexible database and will pave the way for: a consolidation of resources and data within a global framework; a unification of several trials of systematization and by this, enable its users to compare fairy-tales on an international level and create a full picture of the variations of a single tale type both within the ethnical repertoire and throughout the world.

The platform has different modes of operation for internal and external users. It consists of four functional areas: (1) Search engine (with the following search boxes: ethnical origin of the repertoire, type according to the Aarne-Thompson-Uther (ATU) tale type index; identification number of the text contained in the database, index of words used for the motif description); (2) Chart for quantitative information in regard to the tale type (namely: frequency of the type within the repertoire, detailed information about the ethnical and regional inventory of fairy-tales, emerging combinations of several types); (3) Inventory of the tale type as a flexible scheme of the textual elements of the type description. As a result, the elements can be dragged and dropped and new elements can easily be cut or inserted; (4) Chart with details about the text (place of record, narrator, language, dialect and so forth).

Detailed information about the "Web Platform of Comparative Folk Narrative Research" and access to the database via: http://www.folktreasury.ge/Folklore/.

Die „Webplattform der vergleichenden Erzählforschung“[1], kurz: „Webplattform“, stellt einen virtuellen Forschungsraum dar, der es sich zum Ziel setzt, mit Hilfe der durch ihn zur Verfügung gestellten Methoden und Instrumentarien die Vergleichende Analyse des nationalen und internationalen Märchenrepertoires zu vereinheitlichen und zu konsolidieren.[2]

Die theoretische Grundlage der Plattform beruht auf der von der „Finnischen Schule“ entwickelten „geographisch-historischen“ Arbeitsmethode. Zudem basiert die Funktionsweise der Webplattform, die als Raum für internationale Kooperationen dient, auf der Institutionalisierung dieser Methode. Der jüngere Versuch einer solchen Institutionalisierung ist das Langzeitprojekt „Enzyklopädie

---

1 Siehe http://www.folktreasury.ge/Folklore/. Georgischer Titel der Plattform: ხალხური პროზის კომპარატიული ანალიზის ელექტრონული პლატფორმა.

2 Dieses seit 2010 im Folklorearchiv Rustaweli-Institut für Georgische Literatur laufende Projekt wird von der Rustaweli-Stiftung für Wissenschaft Georgien gefördert. An dem Projekt sind folgende Wissenschaftlerinnen, Wissenschaftler und Hilfskräfte beteiligt: David Makalatia, Marine Turashvili, Meri Khukhunaishvili-Tsiklauri, Ether Intskirveli, Elene Gogiashvili, Darejan Toloraia und Magda Turashvili.