In Search of Cinderella: A Transaction Log Analysis of Folktale Searchers

Dolf Trieschnigg
University of Twente
Enschede, The Netherlands
d.trieschnigg@utwente.nl

Dong Nguyen
University of Twente
Enschede, The Netherlands
d.nguyen@utwente.nl

Theo Meder
Meertens Institute
Amsterdam, The Netherlands
theo.meder@meertens.knaw.nl

ABSTRACT

In this work we report on a transaction log analysis of the Dutch Folktale Database, an online repository of extensively annotated folktales ranging from old fairy tales to recent urban legends, written in (old) Dutch, Frisian and a variety of Dutch dialects. We observed that users have a preference for subgenres within folktales such as traditional legends and urban legends and prefer stories in standard Dutch over stories in Frisian. Searches are typically short and aim at large groups of stories (from the same subgenre or collector), or specific stories with the same main character. In contrast, search sessions are relatively long (median of around 2 minutes) and many result pages are viewed (average: 3.4 pages, median: 2 pages). Based on the observations we propose a number of improvements to the current search and browsing interface. Our findings offer insight into the search behavior of folktale searchers, but are also of interest to researchers and developers working on other e-humanities collections.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Measurement, Human Factors

1. INTRODUCTION

The Dutch Folktale Database contains over 42,000 folktales from different subgenres (such as fairy tales, legends and urban legends), from different time periods (from the Middle Ages to present day), and in different languages (Frisian, Dutch and a large variety of Dutch regional dialects). It has both an archival and a research function: it preserves a part of the Dutch cultural heritage and it allows researchers to investigate the oral tradition of telling stories. The Dutch Folktale Database is maintained by the Meertens Institute in Amsterdam which studies and documents the language and culture in the Netherlands. The initial, offline, folktale database was created in 1994, and in 2004 the database became available online at `http://www.verhalenbank.nl` (currently only in Dutch). A variety of users access the online collection, ranging from folktale researchers, students writing for school projects, journalists investigating urban legends, storytellers expanding their repertoire to a general audience interested in stories related to their local region. Search is the primary mechanism to access the collection and the homepage presents a simple search box to find stories in the collection (see Figure 1).

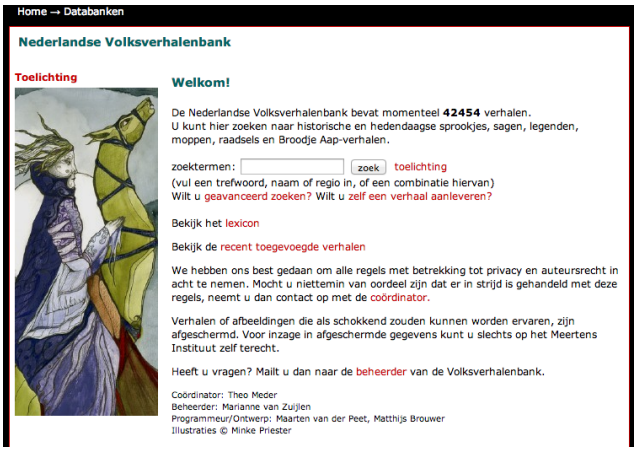


Figure 1: Homepage of the Dutch Folktale Database, including a simple search form.

It is unclear how folktale searchers actually use this search functionality. Given the broad range of users, is the search behaviour comparable to web searchers? Or do they have their own typical search behavior?

In this paper we investigate how users of the Dutch Folktale Database use its search functionality. By means of a transaction log analysis we would like to get more insight on how they search and what they search for. Based on the study we give a profile of folktale searchers and suggest improvements for the current database.

The overview of this paper is as follows. In section 2 we briefly review related work in the area of transaction log analysis. In section 3 we describe the Dutch Folktale Database and its users in more detail based on an online user questionnaire. In section 4 we describe the collection and preprocessing of the transaction log. In section 5 we in-

SIGIR'13, ENRICH Workshop August 1, 2013, Dublin, Ireland.
ACM .

vestigate frequently accessed content of the collection and in section 6 we analyze the users' search behavior. In section 7 we summarize our observations, give suggestions to improve the current database, and present lessons learned for similar projects.

2. RELATED WORK

In this section we briefly review related work in the area of transaction log analysis. For a more comprehensive review see Jansen [4].

Transaction log analysis is an inexpensive way to unobtrusively collect information about a large number of users about their interaction with a web (search) system [4]. Early transaction log analysis primarily focused on the logs of general web search [6, 12]. Jansen et al. [6] analyzed over 50,000 queries on the Excite web search engine. They report that web searchers are uncomfortable using Boolean search operators and other advanced search options. Typically only the first result page is viewed. Silverstein et al. [12] carried out a large log analysis of a web search engine. They report an average query length of 2.3 terms and short search sessions with few queries and result page views.

With the rise in popularity of vertical search, research has focused on analyzing the logs in these specific domains. Mishne and de Rijke [10] analyzed the query log of a large blog search engine. They show that blog searchers primarily search for names and blog themes. Blog search sessions are typically short and only few search results are viewed. Jones et al. [7, 8] investigated a transaction log analysis of a digital library containing technical reports in the area of computer science. Short queries were observed here as well (2.43 terms on average), and most queries (two out of three) did not contain boolean operators. Again, results sets are reported not to be thoroughly inspected by users. Ke et al. [9] investigated the search behavior of Taiwanese users of the ScienceDirect portal which gives access to scientific and technical papers. They report an average query length of 2.3 terms, but do not report on the number of queries during a search session. Huurnink et al. [2] analyzed the transaction log of a audiovisual archive from which material can be ordered. Half of the recorded sessions are shorter than a minute; search sessions resulting in an order are considerably longer (7 minutes). Queries consist mostly of free text keyword search, but also date filters are used. In 9% of the queries the advanced search function is used. Weerkamp et al. [15] report on the analysis of a people search engine log. They propose a classification scheme in this domain at the query, session and user level, for instance by distinguishing queries for more or less popular persons. An average session length of 1.6 queries was observed. Islamaj Dogan et al. [3] investigated one month of query data from PubMed, which provides access to a large repository of biomedical citations. They conclude that PubMed users primarily search for authors, genes/proteins, and diseases and they frequently reformulate their queries. Search sessions consist of 4 queries on average and an average query consists of 3.5 terms. Park and Lee [11] analyzed the transaction log of a web-based IR system in science and technology. They report very short queries (1.4 terms on average) and relatively long users sessions in comparison to web searchers.

To the best of our knowledge no transaction log analyses have been carried out for folktale searchers.

3. THE DUTCH FOLKTALE DATABASE

The Dutch Folktale Database currently[1] contains 42,454 *folktales*. Folktales circulate among people in oral tradition and are part of our folkore and cultural identity. By definition folktales cover a broad variety of subgenres. The most important subgenres in the Dutch Folktale Database are traditional legends (stories with a known place and time and often containing supernatural elements such as witches or ghosts), saint's legends (religious tales about saints, sacred objects and miracles), jokes (short stories for laughter), urban legends (gloomy contemporary stories claimed by the narrator to have actually happened), riddles (question-answer stories) and fairy tales (adventurous stories, playing in an unspecified time and place, often containing magical items). Table 3(a) lists the distribution of the collection over the subgenres. Most of the folktale material has been collected in the nineteenth to twenty-first centuries, but stories from the Middle Ages and the Renaissance are present as well. The stories have been written down in a large number of languages including Frisian, Standard Dutch, 17th century Dutch, Middle Dutch, regional dialects and combinations of languages (also see Trieschnigg et al. [13]). A total of 196 unique language combinations is present in the metadata (based on 92 unique language names). Table 3(b) lists the distribution of the collection over the languages.

Another important metadata field is the type of the folktale. This field refers to international catalog numbers used for indexing folktales such as the Aarne Thompson Uther index [14] and the Brunvand classification of urban legends [1]. Using this field all variations of Cinderella, e.g. told in different languages or in different times, are conveniently grouped.

Other metadata fields include: a summary and keywords in standard Dutch; the geographic region in which the story was told; the name of the storyteller; proper names present in the story; the source where the story came from, for instance a book or received by e-mail; a title of the story (which is frequently not part of an orally transmitted folktale); and the corpus this story is part of. Table 8(a) lists all the metadata fields (as present in the advanced search function).

3.1 User Survey

To get an idea of the user demographics we posted a brief opt-in questionnaire on the homepage of the Dutch Folktale Database, to which 88 people responded between June 2012 and June 2013. A summary of the answers is listed in Table 1 (note that some questions were not answered by all respondents). A majority of the respondents indicated to be male (57%). Table 1(b) lists the indicated age ranges of the users. All age categories are present, but more than a quarter (26%) of the users is between 55 and 64 years old. More than half (57%) of the users is above 45 years old. Table 1(c) lists the highest education the users have received[2]. The users are highly educated: 56% of the respondents indicated to have a university diploma. To the multiple response question what they intend to do with the found information, 64% of the respondents indicated 'personal use' (see Table 1(d)). Another large group of respondents (30%) indicated to use the found information for storytelling. 17% and 14% of the

[1] June 2013

[2] For explanation of the Dutch education names see `http://en.wikipedia.org/wiki/Education_in_the_Netherlands`

respondents indicated scholarly and educative use, respectively. Only 3 respondents indicated to use the information for journalism. In the 'other' category respondents noted inspiration for making art and for developing a guided tour.

We can conclude that the average user of the Folktale Database is highly educated, between 55 and 64 years old, and interested in folktales for personal use or for storytelling (or both).

3.2 The Search and Browsing Interface

The Dutch Folktale Database provides a simple and an advanced search interface. The simple search interface, shown in Figure 1, provides a single search box which searches the keywords, proper names and region metadata fields. The results are shown in a list and are ordered by their id number (e.g. 'ABIJMA22', which consists of a string prefix indicating its collection and a number).

The advanced search interface, shown in Figure 2(a), allows the user to enter separate query terms for each of the metadata fields. The entered values are combined with a Boolean AND. Using check-boxes next to the input field the user can indicate which fields should be shown in the result view. Also the advanced search results are ordered by id number.

When clicking a search result, the user is directed to an overview page of the folktale (shown in Figure 2(b)). This page lists all the metadata fields. The full-text of the story is available through a separate link. The overview page also links to a description of the story type (a catalog page), lists of stories by the same storyteller, lists of stories of the same type, and a map of stories of this type.

4. COLLECTION AND PREPARATION

We carried out a transaction log analysis of the Dutch Folktale Database. We followed the methodology described by Jansen [5] who distinguishes between collection, preparation and analysis of the transaction log. We extracted the transactions from an Apache web server log recorded between April 2010 and Jan 2012, a period of 21 months. From each line we used the following information: the date and time of the request, the IP address or hostname from which the request was issued, the requested page (URL), the user agent (typically the name of the browser) and the response code from the server.

A total of 3,870,947 requests was logged in this period. After removing requests from Web crawlers based on user agent (amounting to 70% of the logged traffic), and removing requests for style sheets and images which are part of the website template, a log of 502,893 requests remained.

We used a simple method to identify sessions in the transaction log: requests from the same IP address or hostname with no longer than one hour between two consecutive requests are grouped into a session. A similar method was used by Weerkamp et al. [15] in the context of people search. We realize that in case of a shared or public computer this can result in erroneously grouped requests. However, given the modest number of users, we expect few errors to occur.

In the next two sections, we first analyze frequently accessed content and then we look into sessions and simple and advanced queries.

Table 2: Request types

Request type	Percentage
Story overview	24%
Story text	17%
Homepage	13%
Simple search result	13%
Lexicon entry	11%
List stories of type	5%
Advanced search result	4%
List stories from storyteller	2%
Recent additions	2%
List lexicon entries	2%
Advanced search page	2%
Catalog page	2%
Map	1%
About page	1%
Multimedia	1%
Bibliography	$< 1\%$
Insert story	$< 1\%$

5. POPULAR CONTENT

In this section we first analyze the type of requests made to the website, then we describe which type of folktales was frequently accessed.

5.1 Request Types

We categorized page requests according to request types. In most cases this came down to classifying the URLs after removing all parameters. If, for instance, a user visits the entry page of the website, this is a request of type 'Homepage', which offers the possibility to search. Entering a (simple) search and pressing enter, results in a 'Simple search result' request, clicking the link to the advanced search form in a 'Advanced search page' request. Requests for (the first or later) search result page are labeled 'Advanced search result'. A complete list of the request types is listed in Table 2. 24% of the requests are for story overview pages, these show the metadata fields in a single view. The second most popular request is for the full text of a story, which is accessible from the overview page. Note that these requests include users who use a general web search engine and directly access a page. The simple search interface is accessible only from the homepage and a search result page is equally often requested as the homepage itself. Of the requests for simple search result pages, 73% is a request for the first page, whereas 27% is a request for the second or later page. The requests for advanced search result pages shows a higher percentage of requests for later pages (44%), indicating that during advanced searches more search results are viewed.

We conclude that basic content pages are most frequently requested, including story summaries, story full-text and lexicon entries. More advanced content pages, such as multimedia items, catalog pages and maps are less popular. The most popular way to find a story (on the site) is through a simple search, but also browsing via stories of the same type and stories from the same author is popular.

Table 1: Summary of responses to the user questionnaire (n=88)

(a) Gender

Gender	Frequency	
Male	49	57%
Female	37	43%

(b) Age

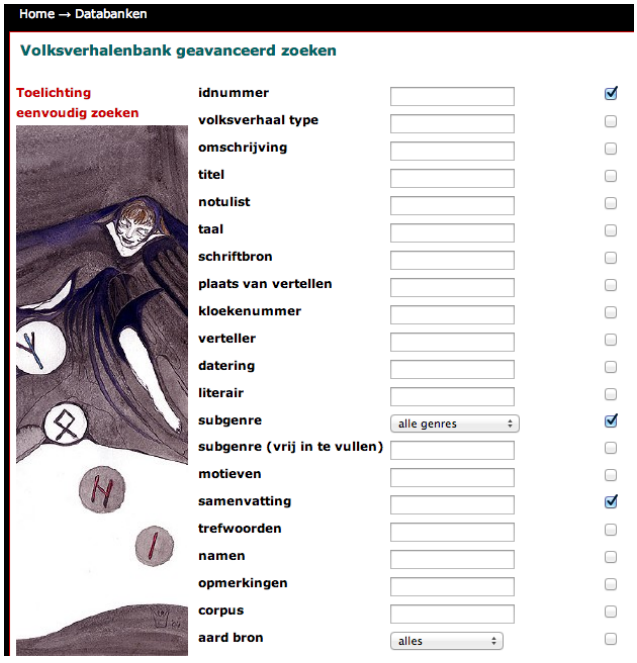
Age range	Frequency	
<15	9	10%
15–24	11	13%
25–34	8	9%
35–44	10	11%
45–54	8	9%
55–64	23	26%
65+	19	22%

(c) Education

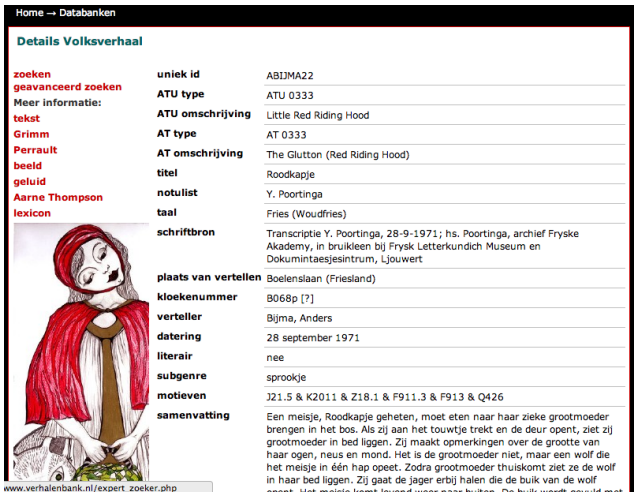
Education	Frequency	
No education, elementary school	6	7%
LBO, VBO, VMBO (secondary school)	4	5%
MAVO, first 3 years HAVO/VWO (secondary school)	3	4%
MBO (vocational learning)	7	8%
HAVO/VWO last 2 years (higher secondary school)	17	20%
HBO/WO-bachelor (bachelor)	29	34%
WO-doctoraal, master (master)	19	22%

(d) Usage

Usage	Frequency	
Personal use	56	63.6%
Scholarly use	15	17.0%
Education	12	13.6%
Journalism	3	3.4%
Storytelling	26	29.5%
Other	11	12.5%



(a) Advanced search page



(b) Story overview page

Figure 2: Screenshots of the search interface

5.2 Accessed Content

A total of 201,129 story requests were in the log. The most frequently requested story amounts to 1.1% of these requests, and is an urban legend about the vanishing hitchhiker. The second and third most frequently requested stories are a traditional legend about being "born with the helmet" and an urban legend about tampered food. Most of the requests for these items originate from popular searches on web search engines. For instance, the urban legend about the vanishing hitchhiker is the first hit when searching the database for a Brunvand catalog number, which is linked from several Google search results.

Tables 3(a) and 3(b) show that the story requests are not evenly distributed over the languages and subgenres present in the database. Traditional legends are frequently requested (45%) and also form the largest subgenre in the collection (55%). However, stories with urban legends and fairy tales, which amount to only a small fraction of the collection, receive many more requests. Most of the stories in the database are in Frisian (42%), but Standard Dutch is the language of most of the requested stories (64%).

In a redesign of the website these statistics can be taken into account to provide easy access to frequently accessed subgenres and languages, or to highlight stories which are currently not frequently requested.

6. SEARCH ANALYSIS

We first look into the characteristics of search sessions. Then we analyze simple and advanced searches in more detail.

A total of 140,136 of sessions is identified, of which 20,162 contain one or more simple or advanced searches. Table 4 lists the session statistics. A session consists of 3.6 requests and takes 3.5 minutes on average; however, half of the sessions consists of only a single request (sessions with only a single request were counted as a duration of 0 seconds). 20,162 sessions (14%) contains at least a single search.

Search sessions have an average duration of more than 10 minutes, but given the large standard deviation this value is deceptive: half of the search sessions is shorter than 2 minutes. Search sessions are quite long: on average 13 requests are made (with a median of 6). Search sessions involving an advanced search (median of 7 minutes) take more than three times longer than simple search sessions.

The sessions are quite long in comparison to search in an audiovisual collection. Huurnink et al. [2] reported a median of half a minute, where folktale searchers use 2 minutes. Also the percentage of advanced searches is higher in this collection (13% vs. 8%). The duration of an advanced search session is comparable to audiovisual searches leading to an order (both 7 minutes), which also could be considered an advanced search. A possible explanation of the differences could be the average age of the users of the folktale database: first-time visitors might need more time to get used to the search system. Users who are more familiar with the website and who use the advanced search function might search faster and more similar to professional audiovisual searchers.

Table 5 lists the querying and viewing statistics of simple and advanced search sessions. On average, 2.4 queries are issued during a simple search session. This is slightly more than the reported number of queries for people search [15] but fewer than an average PubMed search [3]. Twice as many queries are issued during an advanced search session. The number of viewed result pages is surprisingly high: 3.4 result pages are viewed on average (median 2.0). This means that on average 34 search result snippets are viewed. In contrast, the number of viewed story overview pages is relatively low: on average 3.1 (mean 1.0) overview pages are viewed. The full-text of the documents is visited for 71% of the viewed summaries. During advanced search sessions more unique queries are issued, and more result pages, overview pages and full-texts are viewed.

Table 5: Per session statistics

	Simple			Advanced		
Number of	Avg.	Std.	Med.	Avg.	Std.	Med.
Unique queries	2.4	4.4	1.0	4.9	9.8	2.0
Result pages viewed	3.4	7.3	2.0	7.7	15.0	3.0
Summaries viewed	3.1	10.6	1.0	7.9	19.0	2.0
Full-texts viewed	2.2	6.7	0.0	5.2	13.8	1.0

One possible explanation for the large number of viewed result pages might be the fact that the results are not ranked by relevance, but by id number. This might give incentive to look further for relevant documents.

6.1 Simple Queries

Table 6 shows the most frequent simple queries and a description of what it searches for. It is apparent that the most frequent queries are expected to give wrong results or no results at all. Searches for particular collections (e.g. 'cornelius bakker' and its abbreviation 'cb'), subgenre (e.g. 'sage', 'sagen', 'legende', 'stadssage', 'broodje aap') are not supported by the simple search interface, which only searches in the keywords, proper names and region metadata fields. Another frequently issued query is the empty query (yielding no results). This could indicate that visitors do not have a specific information need, but simply want to browse the collection.

Table 7 shows the number of terms per query. On average, a single query consists of 1.4 terms (standard deviation 0.9), but most of the queries (75%) consist of only a single term. In comparison to web queries (Silverstein et al. [12] reports an average of 2.35 terms per query), this is rather short. This might be explained by the fact that the collection is relatively small (around 42,000 documents); also short queries result in a manageable number of results. The same average number of query terms is reported by Park and Lee [11] for searching scientific and technical information.

6.2 Advanced Queries

Advanced queries can be submitted from the advanced search page (see Figure 2(a)), and allows to search in specific fields.

On average, the advanced queries (1.95 terms per query) are slightly longer than simple queries. Table 8(a) lists the use of fields. Advanced searches involving a (part of a) story id are most popular; these are typical known item searches, or searches for known collections[3]. Similar to the most frequent simple searches, searches for particular subgenres and

[3]The story id has a collection-specific prefix

Table 3: Story access per language/subgenre in comparison to occurrence in the database.

(a) Subgenre

Subgenre	Accessed		Database
Traditional legend	91,311	45.4%	54.5%
Urban legend	39,037	19.4%	7.2%
Fairy tale	34,687	17.2%	3.7%
Joke	20,086	10.0%	24.5%
Personal narrative	6,008	3.0%	2.5%
Saint's legend	3,704	1.8%	1.0%
Not assigned	2,370	1.2%	0.6%
Riddle	1,830	0.9%	5.0%
Animal tale	533	0.3%	0.1%
Song	496	0.2%	0.1%
Other	1,067	0.5%	0.8%

(b) Language

Language	Accessed		Database
Standard Dutch	128,759	64.0%	36.8%
Frisian	34,111	17.0%	42.0%
Standard Dutch mixed	12,886	6.4%	3.9%
17th century Dutch	5,058	2.5%	5.7%
Gendts	2,942	1.5%	0.3%
Noord-Brabants	2,874	1.4%	1.7%
Gronings	2,176	1.1%	2.1%
Middle Dutch	2,022	1.0%	1.6%
Flemish	1,577	0.8%	2.4%
Waterlands	1,031	0.5%	0.4%
Other	7,693	3.8%	3.2%

Table 4: Frequency and duration statistics of different types of sessions

		Duration (s)			Number of requests		
	Freq.	Avg.	Std.	Median	Avg.	Std.	Median
All sessions	140,136	213.9	1792.9	0.0	3.6	17.3	1.0
Search sessions	20,162	622.0	1801.3	120.0	13.1	31.3	6.0
Sessions with a simple search	18,923	608.1	1801.4	118.0	12.9	31.0	6.0
Sessions with an advanced search	2,586	1431.0	3104.0	428.5	32.6	61.6	12.0

Table 6: Most frequent simple queries

Query (translation)	Count	Description
cb	1,308	Collection
roodkapje (red riding hood)	816	Main character
sage (traditional legend)	516	Subgenre
sagen (traditional legends)	401	Subgenre
empty	396	
cornelius bakker	331	Collection
legende (saint's legend)	328	Subgenre
bokkerijders (buckriders)	284	Main character
witte wieven (white women)	224	Main character
stadssage (urban legend)	222	Subgenre
broodje aap (urban legend)	221	Subgenre
limburg	216	Region
sprookjes (fairy tales)	198	Subgenre
legenden (saint's legends)	193	Subgenre
moppen (jokes)	173	Subgenre
project (project)	154	
sprookje (fairy tale)	143	Subgenre
sinterklaas	141	Main character
kerst (christmas)	135	Theme
heks (witch)	126	Character type
Other	39,149	
Total	45,675	

Table 7: Number of terms per query

Query length	Frequency	
1 term	34,144	74.8%
2 terms	7,934	17.4%
3 terms	2,014	4.4%
4 terms	646	1.4%
>4 terms	937	2.1%

keywords are popular. The summary field is only infrequently used for searching. Adding this field to the database is quite expensive: an archivist has to read the full-text and write a summary by hand, which is a time-consuming process. It is not used for matching stories, but it is used for displaying search results. Another notable difference in frequency is between region and Kloeke number. The region is a free-text field indicating the place where the story was told, the Kloeke number is a unambiguous identifier indicating a region on the map in the Netherlands and Flanders. Despite the fact that such an unambiguous identifier is present, users prefer to search using a free-text description of the location. A plausible explanation is that searchers do not understand how to use the Kloeke numbers for searching.

Table 8(b) lists which fields are combined in multi-field searches. Most popular are combinations involving a subgenre (e.g. fairy tale, joke or traditional legend). Surprisingly many searches combine a story id with a subgenre, which might seem strange: stories with a similar story id (i.e. from the same collection) can be in multiple subgenres, making an additional filter on subgenre useful. Other useful combinations are a subgenre with one or more keywords,

Table 8: Use of fields in advanced searches

(a) Fields used in advanced searches

Field	Frequency		Examples
story id	4,564	30.0 %	cb, cd, esopet
subgenre	2,885	19.0 %	traditional legend (34%), saint's legend (13%), fairy tale (11%), urban legend (9%)
keywords	1,767	11.6 %	heks (witch), haas (hare), spook (ghost)
type	853	5.6 %	sage, legende, sprookje, at, brun
region	761	5.0 %	Amsterdam, Friesland, Groningen
storyteller	611	4.0 %	cornelius bakker, cb, km
names	575	3.8 %	bartje poep, herk ooievaar, Rotterdam
source type	533	3.5 %	oral (74%), book (8%), article (4%)
description	421	2.8 %	duivel (devil), heks (witch), boom (tree)
collector	403	2.7 %	Koman, Jaarsma, Sophie van Setten
language	291	1.9 %	Dutch (47%), Frisian (14%)
title	290	1.9 %	Roodkapje (Red riding hood), Assepoester (Cinderella)
summary	268	1.8 %	heks (witch), poema (puma)
date	213	1.4 %	1911, 2001, voor 1900 (before 1900)
corpus	162	1.1 %	Overbeke, USA, Jaarsma
source	151	1.0 %	Algemeen Dagblad, Overbeke
motifs	140	0.9 %	k083p, vrouw (woman)
remarks	128	0.8 %	beeld (image/statue), Roosendaal
Kloeke number	104	0.7 %	k150p, k083p, g196p
literary	86	0.6 %	ja (yes), grimm, nee (no)
Total	15,206	100.0 %	

(b) Fields used in multi-field searches

Fields	Frequency	
story id, subgenre	654	21.6 %
keywords, subgenre	333	11.0 %
source type, subgenre	309	10.2 %
region, subgenre	99	3.3 %
storyteller, subgenre	70	2.3 %
keywords, region	68	2.2 %
all fields	60	2.0 %
keywords, names	55	1.8 %
keywords, source type, subgenre	55	1.8 %
subgenre, summary	51	1.7 %
subgenre, type	47	1.5 %
keywords, story id	46	1.5 %
description, subgenre	42	1.4 %
corpus, keywords	41	1.4 %
keywords, language	39	1.3 %
story id, storyteller	34	1.1 %
collector, keywords	34	1.1 %
names, subgenre	33	1.1 %
collector, region	31	1.0 %
subgenre, title	30	1.0 %
Other	903	29.8 %
Total	3,034	100.0 %

and a subgenre with the source type (e.g. book, oral, e-mail). The examples show that searchers have difficulties to understand the advanced search interface. Commonly used queries in the 'type' field are in fact subgenres, which will result in no results.

7. CONCLUSION

In this paper we carried out a transaction log analysis of users of the Dutch Folktale Database. Additionally, we reported on the results of a short survey to determine basic demographics of its users.

Summary of Results

The survey indicated that most folktale searchers are 45 years or older, are highly educated and use the found information for personal use or storytelling.

Basic content items such as story overview pages and story full-text are most frequently accessed. More advanced pages such as maps and catalog pages receive few requests. The most frequently accessed content does not correspond to the distribution of subgenres and languages in the collection. Urban legends and fairy tales are relatively popular under searchers, whereas these subcollections are relatively small. Stories in standard Dutch are more popular than stories in Frisian, despite the large number of Frisian stories in the collection.

Users of the Dutch Folktale Database clearly have their own search behavior. Simple search queries are short (1.4 terms on average), but search sessions are relatively long and many search result pages are viewed. The click-through ratio from result page to story overview page is low, but when a story overview page is viewed frequently also its full-text is requested. Most simple searches aim to find subgenres and collections, but also main characters are frequently searched for.

Advanced searches lead to even longer search sessions with more viewed result pages and story pages. Advanced searches focus on particular stories or collections, subgenres and story types. Multi-field searches typically include a subgenre.

Based on the analyzed searches we hypothesize about two types of users. The first is someone who is familiar with the database and poses advanced queries on the collection to find a known story based on its id, collection or story type. The second is a user who is new to the collection and wishes to explore it. These users are characterized by simple queries on a particular subgenre which results in long result lists. These users are also characterized by empty queries, with the intent to retrieve the complete collection, but which currently returns an empty list.

Recommendations for the Dutch Folktale Database

Based on the results we propose the following improvements for the current Folktale Database.

It turns out that the most popular searches in the simple search interface are semantically incorrect: searching for a subgenre in the simple search interface does not search the subgenre metadata field. Also some of the advanced searches are unnecessarily complex. The search interfaces should more clearly communicate its functionality. Since searching for subgenre is so popular, this should be part of the simple search interface.

Given the broad (subgenre and collection) queries, there is a need for a browsing mechanism of the collection. This browsing mechanism should at least include subgenres, types and languages. The browsing mechanism could also be used

to promote less frequently requested parts of the collection, such as stories in dialects, or riddles and songs.

The free-text region metadata field is more frequently used for searching than the unambiguous Kloeke number, but it can be expected that searching the free-text region field gives undesirable results because of mismatches (e.g. a user searches for the name of the region rather than the village it was annotated with). A more advanced geographic search system which for instance allows querying on a map would make the geographic labeling of more use.

Lessons for Cultural Heritage

If we view our results in the context of cultural heritage in general, we can summarize the following lessons.

The most general advice we can give is to know your user and his needs. Obviously, the search interface should be geared towards these users and frequent search patterns. In the case of the Folktale Database users frequently search for certain subgenres and types which should be clearly accommodated. Search sessions are relatively long: users are willing to put considerable effort in their searches to achieve their goal. The search interface should accommodate these long search sessions, for instance by providing means to reformulate queries, view related stories and visualize search results using different perspectives. Another concrete suggestion would be the use of a basket in which relevant documents can be stored, but which also keeps track of seen stories. Additional tooling could be used to analyze the basket of relevant stories.

But perhaps just as important as the user and his needs is the goal of the access system itself. The goal of the Dutch Folktale Database is to allow the general public to access and get acquainted with a part of the Dutch Folklore. We observed that large subcollections of the Folktale Database are not frequently accessed. We think that this is because the user doesn't know about the existence of the material. In addition, we observed a group of users who does not have a clear information need but wants to explore the collection. Therefore an important lesson is that the access system can be used to put information of interest "on display", analogue to a museum which varies exhibitions to attract different target audiences. For the Folktale Database exhibitions can be stories about a particular region, in a particular dialect or about a specific theme or main character. Using such exhibitions, the system is not only a searchable archive but also a virtual museum which can be browsed and explored.

8. ACKNOWLEDGEMENTS

This research was supported by the Folktales as Classifiable Texts (FACT) project, part of the CATCH program funded by the Netherlands Organization for Scientific Research (NWO).

References

[1] J. H. Brunvand. A type index of urban legends. In *Encyclopedia of Urban Legends.*, pages 741–765. 2012.

[2] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *Journal of the American Society for Information Science and Technology*, 61(6):1180–1197, 2010.

[3] R. Islamaj Dogan, G. C. Murray, A. Neveol, and Z. Lu. Understanding PubMed(R) user search behavior through log analysis. *Database*, 2009, Nov. 2009.

[4] B. J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library and Information Science Research*, 28(3):407–432, 2006.

[5] B. J. Jansen. The methodology of search log analysis. In B. J. Jansen, A. Spink, and I. Taksa, editors, *Handbook of Research on Web Log Analysis*, pages 100–123. IGI Global, Hershey, 2009.

[6] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, Mar. 2000.

[7] S. Jones, S. J. Cunningham, and R. McNab. Usage analysis of a digital library. In *Proceedings of the third ACM conference on Digital libraries*, pages 293–294, New York, NY, USA, 1998. ACM Press.

[8] S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3 (2):152–169, 2000.

[9] H. Ke, R. Kwakkelaar, Y. Tai, and L. Chen. Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library and Information Science Research*, 24(3):265–291, Jan. 2002.

[10] G. Mishne and M. de Rijke. A study of blog search. In *ECIR 2006*, pages 289–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[11] M. Park and T. Lee. Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, 31(1):123–140, 2013.

[12] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, Sept. 1999.

[13] D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder. An Exploration of Language Identification Techniques for the Dutch Folktale Database. In *Adaptation of Language Resources and Tools for Processing Cultural Heritage workshop (LREC 2012)*, Istanbul, Turkey, May 2012.

[14] H. J. Uther. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson*, volume 1-3. Suomalainen Tiedeakatemia, Helsinki.

[15] W. Weerkamp, R. Berendsen, B. Kovachev, E. Meij, K. Balog, and M. de Rijke. People searching for people. In *SIGIR 2011*, pages 45–54. ACM Press, 2011.