A Preliminary Investigation into the Search Behaviour of Users in a Collection of Digitized Broadcast Audio

Lund, Haakon; Skov, Mette; Larsen, Birger; Lykke, Marianne

A preliminary investigation into the search behaviour of users in a collection of digitized broadcast audio

Haakon Lund, University of Copenhagen. Royal School of Library and Information Science, hl@iva.dk

Mette Skov, Aalborg University. Department of Communication and Psychology, skov@hum.aau.dk

Birger Larsen, Aalborg University. Department of Communication and Psychology, birger@hum.aau.dk

Marianne Lykke, Aalborg University. Department of Communication and Psychology, mlykke@hum.aau.dk

Abstract

An increasing number of large digitized audio-visual collections within digital humanities have recently been made available for users. Often access to digitized audio-visual collections is hampered by little and inconsistent metadata. This paper presents the preliminary findings from a study of the search log in a radio broadcast archive. Firstly, results in relation to the identified types of search terms show that the *Programme listing* category was the most frequently identified category followed by categories of *Person* and *Subject*. Secondly, users rarely apply advanced search operators but instead apply phrase or single word queries.

1 Introduction

An increasing number of large digitized audio-visual collections within digital humanities have recently been made available for users. Often access to digitized audio-visual collections is hampered by little and inconsistent metadata (e.g., Hollink et al., 2009). This results in a number of challenges for the use of digitized collections both in research and in daily life. The purpose of this paper is to present the preliminary findings from a study of the search log in a radio broadcast archive, the LARM.FM[1] – archive and accordingly gain insight into end users search behaviour in the context of digitized broadcast audio. In contrast to related previous studies (e.g., Huurnink et al., 2010) the present study focuses solely on access to digitized audio collections. The following research questions guided the study:

Q1: How are the users issuing queries to the LARM.FM – archive?
- What type of search terms can be identified?
- What types of search operators are used?

The LARM.FM–archive was launched in November 2012 as part of a joint initiative between the Danish national broadcasting corporation (DR), the State and University Library hosting the Danish Media Archive, and a consortium of Danish university humanities departments. The LARM.FM-archive provides streaming access to more than 1 million hours of Danish national, regional and local radio broadcasts from 1925 and onwards. In addition the archive is seen as part of a research infrastructure that enable researchers to search and annotate the many recordings of the radiophonic cultural heritage and to communicate about and interact with radio broadcasts.

---

[1] http://www.larm-archive.org

2   Related studies

Despite the growing number of digitized collections within audio-visual archives (Wright, 2007), there are only few studies of how end users search and interact with these collections. In a recent study Huurnink et al. (2010) analysed media professionals' use of an audio-visual archive through transaction log analysis. The results of the study show that (i) over half of the sessions were completed in less than 1 minute; (ii) nearly all of the queries contained a free text keyword search while almost a fourth specified a date filter. Program title was the most frequently occurring keyword search; and (iii) the advanced search options was used in only 9% of the queries. Huurnink et al.'s (2010) study took place within the context of a large broadcast archive containing both video and audio resources. They did not distinguish between the two media types, but the study revealed that less than 10% of the clicks and orders to the archive were for audio material. Accordingly, the results of the study must be skewed towards video as a media type. Therefore the results are not directly comparable to the results of the present study, but nevertheless considered relevant. In addition earlier studies have analysed written requests to audio-visual archives (Hertzum, 2003; Sandom & Enser, 2001) identifying users' underlying information needs but not search behaviour. Finally, only a very limited number of user studies focus solely on audio archives. An exception is Kim et al. (2003) reporting on an exploratory study of the criteria searchers use when judging the relevance of recorded speech from radio programmes.

3   Method and experimental design

The LARM.FM-archive has 731 registered users as of September 2013 (Table 1). Due to copyright restrictions the LARM.FM-archive is not open to the public but only to researchers, university students and some professionals (librarians, archivists etc.). Unfortunately the user registration system has changed since the launce of the LARM.FM– archive and only details about occupation and institutional affiliations from 427 users are available.

Table 1 : The user groups in LARM.FM. The 284 users with unknown occupation is due to the change in user registration system.

Occupation	Number of users	% of total
Researchers	133	18
Students	228	31
Professionals	86	12
Unknown status	284	39
Total	731	100

The high number of students is due to the participation of students in research projects using LARM.FM as well as the use of LARM.FM for teaching and therefore it is expected that the composition of the user groups will change according to on-going research activities or teaching activities.

The LARM.FM user interface is developed in Microsoft™ Silver Light, which in combination with the use of Google Analytics as logging systems limits the possibility to log all user actions in the interface. The implementation of Google Analytics in LARM.FM allows us to identify a user session but it is not possible to identify the user behind a user session. Subsequently it is not possible to connect sessions to users or their characteristics, e.g. user groups.

The radio programmes were imported into LARM.FM with sparsely populated metadata about each programme but the implemented metadata scheme allows users of LARM.FM to enrich the recordings (Lund, Bogers, Larsen & Lykke, 2013). As of September 2013 only 1086 of 578978 programmes have been enriched with user generated metadata.

Log analysis is an unobtrusive way to collect large amounts of data on user search behaviour (Jansen, 2008). In this study we have analysed the usage log for the period from 10th of August 2013 until 9th of September 2013 both days included resulting in approximately 15740 entries. To identify user entries containing a query formulation the usage log where parsed through a sed-script resulting in 6837 entries that included a search string. Of these 2827 were identified as unique user sessions where the user had performed a search in LARM.FM

To identify the types of search terms used in the extracted query strings a coding scheme was developed based on a scheme used by Huurnink et al. (2010). The scheme divided queries into 8 categories, of which 5 categories (Location, Person, Name, Genre, Subject) were taken from Huurnink et al. (2010) and 3 additional categories (Time/year, Title and Programme listing ID) were added to match the current study. Categorization of search terms was done independently by 2 of the authors of this paper. The results were then compared and decisions on category were decided accordingly.

4   Results

First we look at the type of search terms that can be identified. A total of 2827 search sessions were identified including 2470 unique queries and 357 queries where the same query formulation was repeated.

Table 2 : Queries according to list of categories

Type	# of queries	% of queries
Time/year	151	6
Title	211	9
Location	75	3
Person	292	12
Name	68	3
Genre	68	3
Subject	357	14
Programme listings	645	26
Unknown origin	689	28
Total # of occurrences	2556	
Total # of unique queries	2470	

Table 2 shows the frequency of the 8 categories identified in the 2470 unique user queries. The *Programme listing* category was the most frequently identified category (26%) reflecting a verificative search on a specific radio programme listing ID number. It is followed by the *Subject* (14%), *Person* (12%), and *Title* (9%) categories. Overall, *Location*, *Name* and *Genre* were the least frequently identified categories.

The LARM.FM front-end allows for entering a search string or a direct search on date (either single date or date range), content type (radio shows and program listings), user generated content (user enriched metadata) and broadcast channel. The search box for entering a search string does not provide any help feature about the search syntax used. The back-end of LARM.FM is built on Apache Solr[2] search platform and search phrases are delimited by using double quotes. The default operator in LARM.FM is OR.

Table 3 shows the type of delimiters and search operators used. The results show that the large majority of queries are issued as either a phrase search or a single word string, whereas use of advanced search features such as Boolean operators or truncation is limited. The 177 extracted word strings are probably meant to be phrase searches and are possibly due to users who are not aware of the correct search syntax.

---

[2] http://lucene.apache.org/solr/

Table 3 : Delimiters and operators used

	Result
Phrase search using delimiter	1202
Word (single word string)	1058
Word (multiple words)	177
Operator (NOT)	1
Operator (AND)	53
Truncation (*)	19

5　Discussion and conclusion

The findings correspond to previous results concerning end-user searching that end-user searchers still make simple, short queries with few free-text search terms and little use of advanced features.

The present paper presents the preliminary results regarding how users issue queries to the LARM.FM-archive based on log analysis. The preliminary results indicate that parallels can be drawn to the log analysis study by Huurnink et al. (2010). Firstly, in relation to applied search terms, the categories of *Person* and *Subject* are more frequently applied than *Location, Name,* and *Genre*. Secondly, users rarely apply advanced search operators but instead apply phrase or single word queries. Infrequent use of advanced search operators is earlier identified in relation to audio-visual material (Huurnink et al., 2010) and in search behaviour in general (Markey, 2007). The analysis of user queries further indicated a surprisingly interest in music related radio broadcasts. Future log analysis on user behaviour in LARM.FM will study a larger time span (one year) to determine whether the music related interest can be explained by on-going research activities or whether it is a more general characteristic. Future work should also include how queries evolve during a search session, how date and programme filters are used to limit searches, and an analysis of how users tag and annotate radio programmes.

References

Hertzum, M. (2003). Requests for information from a film archive: A case study of multimedia retrieval. *Journal of Documentation*, 59(2), 168–186.

Hollink, L., Schreiber, G., Huurnink, B., van Liempt, M., Rijke, M., Smeulders, A., Oomen, J. & de Jong, A. (2009). A multidisciplinary approach to unlocking television broadcast archives. *Interdisciplinary Science Reviews*, 34(2-3), 257- 271.

Huurnink, B., Hollink, L., van den Heuvel, W. & de Rijke, M. (2010). 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.

Jansen, B.J. (2008). The methodology of search log analysis. In B.J. Jansen, A. Spink & I. Taksa (eds.), *Handbook of research on Web log analysis* (pp. 99-121). Hershey, PA: Idea Group Inc.

Markey, K. (2007). Twenty-Five Years of End-User Searching, Part 1: Research Findings. *Journal of the American Society for Information Science and Technology*, 58(8). 1071-1081.

Lund, H , Bogers, T , Larsen, B & Lykke, M (2013), CHAOS: User-driven Development of a Metadata Scheme for Radio Broadcast Archives. *Proceedings of the iConference 2013*. (pp. 990-994), Urbana, Ill., IDEALS.

Sandom, C.J., & Enser, P.G.B. (2001). Virami—Visual information retrieval for archival moving imagery. In *Proceedings of the International Cultural Heritage Meeting* (pp. 141–152). Milan, Italy: Archives & Museum Informatics.

Wright, R. (2007). *Annual report on preservation issues for European audiovisual collections*. Retrieved September 17, 2013, from http://www.prestospace.org/project/deliverables/D22-8.pdf.

Table of Tables