Crowdsourcing Metadata for Audiovisual Cultural Heritage: Finnish Full-Length Films, 1946–1985

Hannu Salmi[1,4], Kimmo Laine[2,3], Tommi Römpötti[2], Noora Kallioniemi[1] and Elina Karvo[1]

[1] University of Turku, Dept of Cultural History
[2] University of Turku, Dept of Media Studies
[3] University of Oulu, Dept of Art Studies and Anthropology
[4] Turku Group for Digital History
hannu.salmi@utu.fi, kimmo.laine@utu.fi, tommi.rompotti@utu.fi, noora.kallioniemi@utu.fi, elina.karvo@utu.fi

Abstract. This paper is based on a crowdsourcing project which was realised at the School of History, Culture and Arts Studies of the University of Turku between the years 2013–2018. The idea was to develop a format through which long-term crowdsourcing could be integrated into the humanities curriculum. The project was realised in close cooperation with the National Audiovisual Institute (KAVI) in Finland. The aim was to help KAVI in developing its open database for Finnish cinema, Elonet, by engaging both graduate and postgraduate students in producing keywords, genre characterisations, plot summaries and other relevant fields of information for Finnish cinema. In total, the project produced metadata for 572 full-length films, both fiction films and long documentaries that had their theatre release between the years 1946 and 1985. The amount is substantial considering that, to date, around 1,600 full-length films have been released in Finland. At the same time, it produced a successful model for drawing on crowdsourcing in the classroom.

Keywords: Metadata, Digital Humanities, Film History, Audiovisual Culture, Audiovisual Heritage.

1 Introduction

Since 2006, the National Audiovisual Institute in Finland (KAVI), formerly known as the Finnish Film Archive, has been working to create an open database, Elonet (http://elonet.finna.fi). Its content is to a large extent based on the filmographic work done during the twelve-volume publication of the Finnish National Filmography between the years 1989 and 2005 [1]. The database includes information on imported productions and short domestic films, but the major goal is to provide users with basic filmographic information on all full-length films that have premiered in a movie theatre in Finland, starting from the first Finnish fiction film, *The Moonshiners* (*Salavi-inanpolttajat*, 1907). The focus has been on full-length work (i.e. a running time of 37 minutes or more or a length of 1000 meters of 35 mm film or more), although database also includes shorter fiction films created before the breakthrough of long fea-

tures in the 1910s. By the end of 2017, the *Finnish National Filmography* included 1,540 titles, 245 of which were documentaries [2]. The distribution of films in the collection by year of creation is shown in Figure 1. The so-called studio era, which lasted from the mid-1930s to the beginning of the 1960s, featured the most productions, although the 2000s are a close second and this period is already a substantial part of Finnish audiovisual heritage.

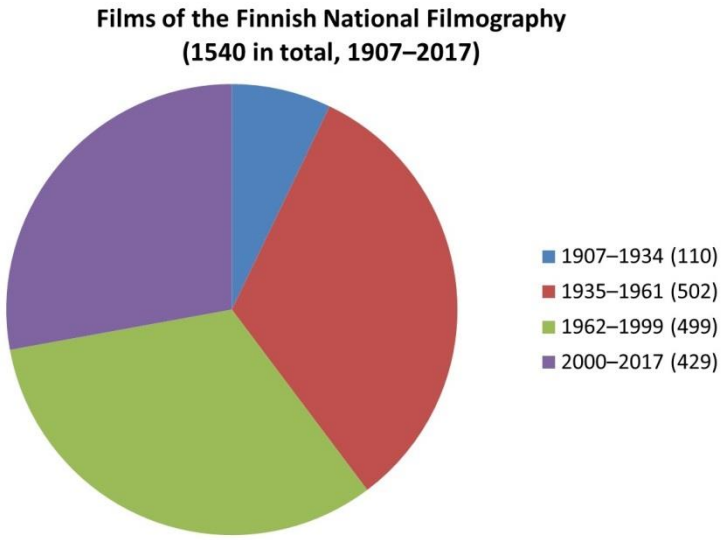

Fig. 1. Distribution of Finnish cinema in the *Finnish National Filmography* by year. Source: http://elonet.finna.fi.

The aim of the Elonet database is to provide information about the actors and staff of each film as well as production data; all accumulated data on the release history of the film, including later re-releases and television broadcastings; a description of the content; a summary of the press coverage of the film; important background information on the making of the film; soundtrack information; and details on possible censorship actions. The database also includes a short plot summary, genre categorisation and keywords that help the user to find relevant information.

Previously, keywording of fiction films was laborious and little had been done for the period after World War II. To help in this effort, in 2013, KAVI and the University of Turku launched a crowdsourcing project to investigate the possibility of producing missing metadata for the Elonet database. At the same time, basic information about each film was double-checked in order to identify possible mistakes or contradictions in the metadata.

Jeff Howe has defined crowdsourcing as an ‘act of taking work once performed within an organisation and outsourcing it to the general public through an open call for participants’ [3]. The *Oxford Dictionary of English* states that Howe was the first to define the word, but it adopts a slightly different definition: ‘The practice of obtaining information or services by soliciting input from a large number of people, typical-

ly via the internet and often without offering compensation.'[4] In our case, we did not make an open call but merely aimed to create a method in which metadata could be produced in a university classroom following the principles of crowdsourcing. Only about twenty students participated each year, but by the end of the five-year project, a large number of participants had contributed to the creation of metadata.

In recent years, several books and articles have been published on crowdsourcing for cultural heritage purposes, including aspects related to metadata [5]. An example of a metadata project is the BBC's crowdsourcing project, which aimed to produce descriptive tags for radio programmes and correct machine-generated tags [6]. In our project, in turn, we integrated crowdsourcing into the humanities curriculum and concentrated on plot summaries, keywords and genres.

2 Format and Material

The crowdsourcing project started in the 2013–2014 academic year as a course entitled Film History Research Group (10 ECTS). In total, 21 students from different departments in both graduate and postgraduate programmes were enrolled. The timeframe was restricted to 1946–1952 to keep the amount of films manageable; the *Finnish National Filmography* includes 127 titles for that period, which means that there were approximately six films for every participant. The goal was to cover all released films in the chosen period, not only providing basic filmographic details but also performing a close qualitative reading of each film for the benefit of subsequent research and, possibly, for the enrichment of metadata. The course included five parts: (1) background lectures on Finnish film production from 1946–1952; (2) discussion of the agenda, including guidelines for plot summaries and genre characterisations as well as principles for keywording; (3) distribution of the basic information form to be filled out for each film; (4) establishment of a set of research questions for a separate research form to be completed for each film; and (5) preparation of thematic essays by the participants in teams.

The groups gathered qualitative information about the content of the movies in part 4, but the crowdsourcing of metadata was emphasised in the basic information form in part 3. The participants had to fill out the form, which required them to write a short summary of the film (200–700 characters) to be later used in the Elonet database and double-check the longer description of the film that had been produced while editing the *Finnish National Filmography* to find possible mistakes. Each film was assigned a genre from a pre-set list as well as 10–20 keywords that capture the content of the film.

After the first year, the process was repeated four times. The five-year period ensured that a substantial number of Finnish films could be processed and metadata could be produced systematically. The format was refined along the way, but the basic premises remained the same. The timeframes and distribution of films were as follows:

1946–1952, 126 films (2013–14, PIs: Laine, Römpötti, Salmi)
1953–1956, 101 films (2014–15, PIs: Laine, Römpötti, Salmi)
1957–1961, 92 films (2015–16, PIs: Laine, Römpötti, Salmi)

1962–1970, 96 films (2016, PIs: Kallioniemi, Salmi)
1971–1979, 67 films (2017, PIs: Kallioniemi, Karvo)
1980–1985, 90 films (2018, PIs: Kallioniemi, Karvo)

As the previous table shows, throughout the five-year project, 572 full-length films were processed. Figure 2 shows how the period of 1946–1985 relates to the overall volume of Finnish cinema.

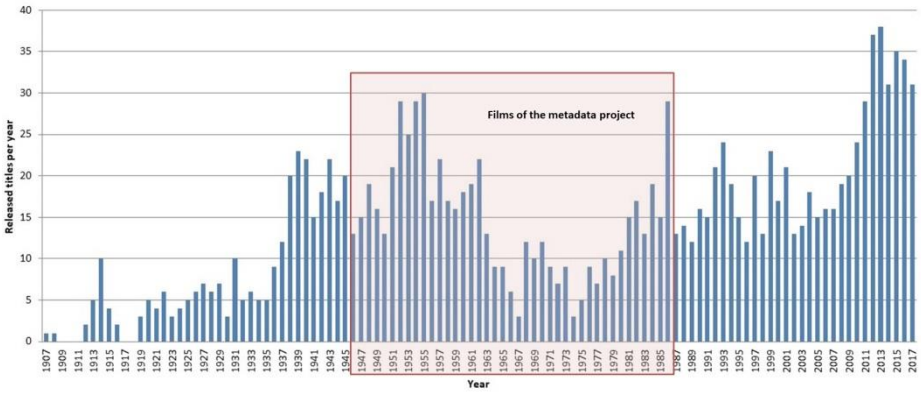

Fig. 2. The films processed in the metadata project in relation to all Finnish films. Source: http://elonet.finna.fi

The films processed in the project were provided by KAVI, which is the copyright holder for the major film studios of the 1940s, 1950s and early 1960s. Since KAVI could make digital copies of films, the participants had easy access to them. In some cases, especially for films made the 1960s and 1970s, the films had to be watched at KAVI in Helsinki.

3 Metadata in the Making

The project focused on three aspects of metadata:

(1) Plot summaries. A short summary was created for the title page of each film in Elonet so that a reader could obtain a quick but comprehensive view of the film. The guideline for the project suggested that each summary should include productional details so that the film makers, especially director, would be mentioned as well as the name of the film, its genre or type (animation, melodrama, etc.) and the background of the production (for example, if it is a literary adaptation). Furthermore, each summary should crystallise the content of the film, milieu (time and place), main outline of the plot and major roles, including the names of the principal actors.

(2) Keywords. For each film, 10–20 keywords were created to improve one's ability to search for it in the database. The aim was to characterise the phenomena, themes and motifs in the story and the cinematic space, including material artefacts, events, agents, time and place. The project drew on the General Finnish Thesaurus (YSA), which 'covers all fields of research and knowledge and contains the most

common terms and geographical names used in content description' [7]. In 2019, YSA was replaced by the General Finnish Ontology (YSO) and its extension for place names, YSO Places [8].

(3) Genre. A set of genres for both fictions and documentaries was provided and refined and added to during the project. The set included established branches of film, such as crime movies, westerns or melodramas, as well as 'Finnish' genres, including the *rillumarei* movies of the 1950s. The Finnish film industry was organised according to the Hollywood model from the 1930s until the early 1960s, and while there was no real genre system like in the United States, there were certainly films that applied elements of American genre films.

The major challenge of crowdsourcing is remaining as consistent and congruent as possible. In this project, in addition to guidelines and orientation sessions, consistency was ensured by the editorial participation of the PIs. KAVI double-checked the results before adding them to the database. Despite this, difficult problems were faced throughout the project, especially regarding keywords and genres. For example, the General Finnish Thesaurus is not designed to be employed to describe fictitious material. Keywords are useful for enhancing the discoverability of particular themes and phenomena, but they always have a discursive dimension. In the complex aesthetics of fiction films, many themes are touched upon not only through discussion or by showing particular items or places but also through symbolic communication, such as visual effects or references in the soundtrack. Thus, keywords can only provide a partial view of the content of an audiovisual product. Compared to the richness of the work, 10 or 20 keywords offer only a few points of reference for metadata. The researcher is thus faced with the challenge of selecting the most descriptive and illuminating keywords. Another problem is that some themes are so common, like alcohol use or falling in love, that they might become redundant and eclipse other themes. Despite these difficulties, however, the five-year project produced a relatively consistent vocabulary of keywords that significantly improved the searchability of Finnish films.

In the case of genres, it was obvious from the start that many films were characterised by more than one genre. If the aim of metadata is to increase discoverability, rather than to categorise films, then two or three genres can be assigned to a film to better describe it. A good example of a difficulty related to characterisation is the musical in Finnish film. The term 'musical' does not fit well with Finnish productions, as Finland never had musicals in the style of Hollywood. However, there were various types of films that featured music. In the late 1950s and early 1960s, a group of films described as *iskelmäelokuvat*, or Schlager films, were produced. These were compilation films comprising a series of musical numbers, often based on recently released records. This problem was solved by introducing *iskelmäelokuvat* into the list of genres and by giving also other characterisations.

4 Future Potential

During its five-year project, the Film History Research Group was able to meaningfully contribute to the construction of the Elonet database. In addition, the project produced new avenues for future research on Finnish cinema, such as the ability of

keywords to indicate major changes in the content of Finnish cinema. Our project included films published from 1946 onwards, a period that saw an important change in film production due to the gradual shift from wartime to peace. Many of these films dealt with concerns about social problems, for example the abuse of alcohol and the increasing crime rate. If the keywords associated with Finnish fiction films produced from 1946–1948 are clustered (i.e. those keywords that tend to occur in conjunction are placed together), thematic points of emphasis after the war can be identified. Figure 3 shows these narrative focal points from 1946–1948.

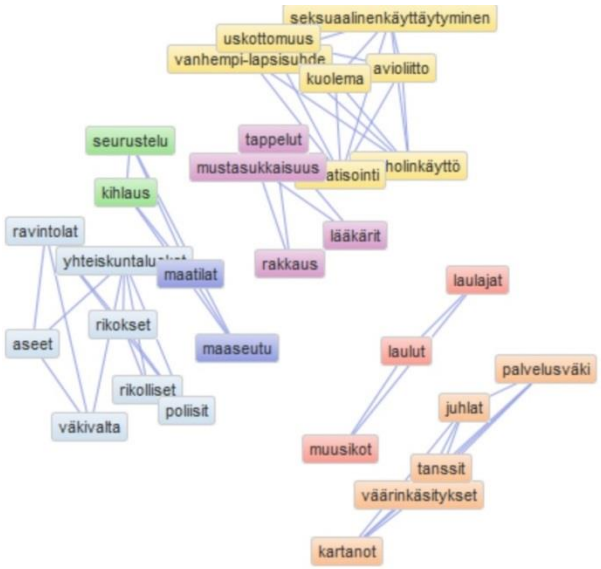

Fig. 3. Clusters of keywords for Finnish fiction films from 1946–1948. Source: Hannu Salmi/Elonet.

The Finnish-language keywords in Figure 3 refer to seven different groups of films, which are separated in the graph spatially and by colour. The light blue cluster on the left, for example, refers to crime movies and includes keywords such as ‘restaurants’, ‘guns’, ‘social classes’, ‘crimes’, ‘criminals’, ‘police’ and ‘violence’.

In this example, only keywords from 1946–1948 have been used, but if similar network graphs were produced for each three-year period after the war, larger changes in cinematic contents would become visible. Keyword analysis enables one to understand changes in narrative structures and topics on a large scale. There is much potential in this regard, but it was out of the scope of this project. During the crowdsourcing endeavour, in addition to keywords and plot summaries, we gathered qualitative information, too, about the themes of each film in the part 4 (research form). This produced hundreds of pages of content descriptions that can be text-mined for further analysis of the whole timespan from 1946 to 1985.

Another possible avenue for future research is to extract information from the audiovisual content itself through speech recognition and audiovisual content analysis. As Taylor Arnold and Lauren Tilton recently argued, distant viewing techniques can

be effectively used for ‘extracting semantic metadata from images’, and in fact from all audiovisual content [9].

5 Conclusion

This paper has provided an overview of a crowdsourcing project that represents successful cooperation between an academic institution and a memory organisation. In contrast to Jeff Howe’s definition, it never aimed to outsource metadata work to the general public; rather, it experimented with a format that combined academic research with social engagement. This was highly motivating for those who participated in the project since it added to their qualifications within the fields of cultural heritage studies and media studies. The participants added their contribution to the Eloent database.

Crowdsourcing might run the risk of being too wide and too uncontrolled. In our case, these problems were avoided by repeating the course five years in a row and enhancing its format. Each year, the group of participants was rather small, which ensured high-quality results. In sum, the project succeeded in finding a successful strategy for integrating crowdsourcing into academic teaching. It produced metadata for 572 full-length films, which is a substantial contribution to the study of Finnish cinema.

Acknowledgments

This work was supported by the National Audiovisual Institute, Finland, and the School of History, Culture and Arts Studies of the University of Turku. Special thanks to Jorma Junttila (National Audiovisual Institute) for cooperation.

References

1. Suomen kansallisfilmografia. 12 vols. Finnish Film Archive, Helsinki (1989–2005).
2. Elonet, http://elonet.finna.fi, last accessed 2020/01/14.
3. Howe, J.: The Rise of Crowdsourcing. Wired, June (2006), http://www.wired.com/wired/archive/14.06/crowds_pr.html, accessed 30 September 2019. See also Hopkins, R.: What is Crowdsourcing? In: Sloane, P. (ed.) A Guide to Open Innovation and Crowdsourcing: Advice from leading experts, p. 15. London: KoganPage, London (2011).
4. Crowdsourcing. In: Oxford Dictionary of English. Third edition (2013), https://www.oed.com/view/Entry/376403?result=2&rskey=BoihnO&, last accessed 2019/09/30.
5. Ridge, M. (ed.) Crowdsourcing our Cultural Heritage. Routledge, London (2016). See also, for example, Sounds of the Netherlands. In: Heritage in Motion, https://heritageinmotion.eu/himentry/sound-of-the-netherlands, last accessed 2019/09/30. See also Botto, R.: Crowdsourcing for Filmmakers: Indie Film and the Power of the Crowd. Routledge, New York (2017).

6. Raimond, Y., Smethurst, M., Ferne. T.: What we learnt by crowdsourcing the World Service archive. In: BBC Research & Development (2014), https://www.bbc.co.uk/rd/blog/2014-08-data-generated-by-the-world-service-archive-experiment-draft, last accessed 2019/09/30.
7. YSA, the General Finnish Thesaurus, https://www.kansalliskirjasto.fi/ en/node/167, last accessed 2019/09/30.
8. General Finnish Ontology YSO, https://www.kansalliskirjasto.fi/en/services/expert-services-of-data-description/general-finnish-ontology-yso, last accessed 2019/09/30.
9. Arnold, T., Tilton, L.: Distant viewing: analyzing large visual corpora. In: Digital Scholarship in the Humanities (2019), published online 15 March 2019. https://distantviewing.org/pdf/distant-viewing.pdf, last accessed 2019/09/30.