MedieKultur | Journal of media and communication research | ISSN 1901-9726

Article – Theme section

How streamification challenges the Royal Danish Library's collection of cultural heritage

Andreas Lenander Ægidius

MedieKultur 2021, 70, 91-109

Published by SMID | Society of Media researchers In Denmark | www.smid.dk

The online version of this text can be found open access at www.mediekultur.dk

Abstract

Danish legal deposit legislation mandates that the Royal Danish Library must collect contemporary culture for the benefit of the public and researchers, now and in the future. In this article, I analyse how the move toward access models and the subsequent streamification of media content challenges the collection of cultural heritage. I draw on empirical data from two central activities. The main empirical data stems from archival research and interviews with the library's internal and external stakeholders. The second source of empirical data is the in-house testing performed by web curators when they analysed the collection of streaming-only content via an API from the Danish Broadcasting Corporation (DR). The analysis is followed by a discussion of the effects of streaming software and services on collection methods such as stream-ripping, screen capture (image or video), and research collaborations with the producers and distributors of born-digital content.

Keywords

Streamification, streaming services, video-on-demand (VOD), digital cultural heritage, collection methods, application programming interface (API), DR.

Introduction

The Royal Danish Library[1] collects, preserves, and makes cultural heritage available in the form of texts, images, audio and video, whether published on physical delivery technologies (e.g., CD, DVD) or online (web). Danish legal deposit legislation mandates the preservation of contemporary culture for the benefit of the public and researchers, now and in the future. The law mandates publishers of physical publications to deposit exemplars of their publications to the library. According to the law, any published content should be deposited regardless of the particular media used as its means of distribution (Kulturministeriet, 2005). The law states that the Royal Danish Library must collect content published online. Publishers must provide access to the published content. Content is increasingly published outside traditional media channels. The increasing amount of content that is published online/digitally first is known as "born-digital (Brügger, 2016). It is increasingly digital-only, and accessed via streaming services that operate across national borders. In the meantime, the mandate to document and collect digital cultural heritage at a national level is increasingly difficult to fulfil. The library simply cannot collect it all. Situated in this highly fragmented and dynamic media landscape, the web curators at the library are testing ways to collect digital-only streaming content.

In this article, I analyse how the move toward access models challenges the collection of cultural heritage. The development of access models and their effect on commercial and public sectors is leading to a renegotiation and a reformatting of the ways we can collect the cultural heritage. I term this development *streamification*, with reference to similar concepts that outline processes that help us frame the influence of software, platforms, and individual companies.

I draw on empirical data from two central activities. The main empirical data stems from archival research and interviews with the library's internal and external stakeholders. The second source of empirical data is the in-house testing done by *web curators* as they analysed the collection of streaming-only content via an application programming interface (API) from the Danish Broadcasting Corporation (Danmarks Radio - DR). Content previously published on DR as flow-TV is now only available through its video on-demand (VOD) software, *DRTV*.

The analysis will lead to a discussion of collections methods for streaming software and services. The analysis suggests that collection through API should be paired with, and is different from, other collection options, each with varying strengths and weaknesses. Building on the overview provided by Laursen et al. (2017), I discuss screen-capture (image or video), stream-ripping, collection by collaboration, and web harvesting.

The present analysis of existing collection practices and API-tests at the Royal Danish Library provides a unique opportunity to survey the ways that streamification challenges the collection of digital cultural heritage. The curation and inevitable selection of content

[1] In Danish, Det Kgl. Bibliotek.

is critical in order to provide the current and future service to the public and to media researchers.

Born-Digital

Curators subdivide the concept of digital content in their daily work. Brügger (2016) proposes a general typology of digitality, based on the *provenance* of the digital material. He identifies three types of digital material: digitised, born-digital, and reborn-digital material.

1. Digitised material is analogue material that has been digitised, typically print and tape.
2. Born-digital material is digital material that has never existed in any form other than digital. This includes all types of material on digital media, such as CD-ROMs and DVDs, or the web.
3. Reborn-digital material suggests that born-digital material has been collected and preserved, and has to a large extent been changed in the process of collecting and preserving it (Brügger, 2016).

This distinction is part of the daily discourse about content and collection practices in the Department of Digital Cultural Heritage at the Royal Danish Library, which I will explain in the case below. The content made available on DR flow-TV and *DRTV* is born-digital simply because the production process and distribution method is more effective than having to rely on analogue recording methods and digitisation processes. In the case of radio and TV the born-digital content and the digitised content exist side by side. When certain content predates digital production methods, DR will have had to digitise the content. This makes sense from the user's point of view, and we then see streaming merely as the transmission of digital content. For analytical purposes, I argue that a stream is the progressive download of a media file along with an amount of metadata. Parts of the metadata will be instantiated in a graphical user interface (GUI) for the user's orientation. The stream is born-digital, however, collection through an API can result in large media files and various types of metadata files. Sorting, validating, and preserving this data and subsequently providing access to it, means that there is a probability of the material being reborn-digital, which is not ideal.

Increasingly, publishers rely on a combination of digital-first and digital-only publication strategies as a way to safeguard against potential losses from the production of physical copies to a market that is mainly digital. The Danish population is adopting streaming to supplement their traditional television viewing, but the use flow-TV is decreasing. In 2019, based on a daily average of seven hours of media use, the general population in Denmark (15-75 years) watched twice as much flow-TV compared to streaming TV-content, films, and short clips. They spent 33 pct. of their time on flow-TV vs. 15 pct. on TV/video streaming. Other categories included were broadcast radio, streaming radio and music, reading online news, gaming online, social media, and print media. The average for youth and young adults (15-31 years) was the reverse, and heavily skewed towards TV/ video streaming: 10 pct. of their time was spent on flow-TV vs. 25 pct. spent on TV and video streaming (DR, 2019, p. 5). This indicates that we cannot expect or rely upon there ever being a physical copy of the digital content published. This is crucial for the library, as it shifts the burden of legally depositing physical copies from the publishers to the library, which now has to collect even more content published online. This reflects the shift from the purchase of books, CDs and DVDs to the increasing reliance on streaming services for the same content.

Collection versus selection

For the purposes of clarification, I define archiving as the practices involved in making archives. The process of archiving includes collection, preservation and providing access to the content. Collection is thus the first step in a three-part workflow that includes preservation and the subsequent accessibility of the digital material. This means that the decision to collect streaming content intertwines with decisions about how best to preserve the content for the future. At the same time, the content that constitutes our cultural heritage is only ever relevant when we can access it to learn about our historic and contemporary affairs. This is evident in the strategy of the Royal Danish Library ,which articulates how it aims to collect, preserve, disseminate, and research the Danish cultural heritage (KB.dk, n.d.-b).

Netflix, Spotify, YouTube, and other commercial streaming services see themselves as intermediary distributors. Many, if not all of them, supplement distribution with the production and publication of their own original or co-sponsored content. They operate at a global level, or they typically aspire to do so. DR and other public service media provide live and/or on-demand streaming services based on their own and licensed content. However, territory-specific copyright regulation limits how content can move across borders. Some if not all of the content is so-called *Danica*, meaning made by or about Danes, which makes it Danish intellectual and cultural resources and eventually heritage. In the meantime, the mandate to document and collect digital culture at a national level is increasingly difficult to fulfil. The library simply cannot collect it all. Librarians and researchers have described multidimensional disruptions, and argue that selection is a critical part of the documentation and collection of culture (Clifford, 2017; Mandel, 2019). This is especially true for content available as streams on social media, including YouTube. This leads to collection practices that aim to establish representative archives rather than complete archives. UNESCO recognises that the survival of digital heritage worldwide is much less assured than its traditional counterparts, and provides generalised guidelines for the identification and collection of significant digital heritage (Choy et al., 2016).

Streaming

Streaming is in technical terms a form of media content distribution called progressive downloading that allows playback before the entire media file downloads to the client computer. Contrary to purchase and subsequent download, the streamed content remains only temporarily in a hidden cache on the client computer.

The move from owning media files to accessing them is what essentially separates the era of download purchases and file sharing from the era of streaming. A streaming service will typically provide access through subscription-based or ad-based models that vary greatly in reference to its implementation in different markets, legislative frameworks, and IT-infrastructure. Streaming is the commodification of temporary access to content that we used to download via purchase or file sharing (Herbert et al., 2019). Beginning in 2010, DR enabled the live and on-demand streaming of a selection of its content at dr.dk/nu. The revamped DRTV launched in 2014 with dedicated apps and HbbTV support (Holm, 2014).

Streaming has moved from the periphery to the centre of internet development, and affected all traditional media industries in doing so, which resulted in a multitude of variants (Spilker & Colbjørnsen, 2020, p. 1215). It is worth noting that the remote control and the VCR predate streaming video on-demand as technological solutions that reconfigured user agency (Burroughs, 2019). In providing a cultural lineage of streaming, Burroughs (2019) rightly references the concept of flow (Williams, 1975), and following Sterne (2012), he notes how compression through the codec leads to the interoperability of streaming media across platforms and formats, and within digital culture.

From a user perspective, streaming services typically constitute access to a stream of text, sound or video. It is not uncommon for text-based content to be re-framed as a streaming product, especially if a given service provides access to more than one content type, i.e. books and audiobooks that were widely downloaded by purchase, or file-sharing before streaming caught on as a market venue for books (Have & Pedersen, 2020). In the streaming of video, for example, types of content vary within existing categories, as the concept of streaming has come to cover subscription-based video on-demand (movies and series), live or on-demand television (programs, news, films and series), as well as streaming short clips (e.g. YouTube). In these broad terms streaming tends to signify a highly dynamic and interconnected ecology of digital distribution, with its associated websites and apps that give access to, and even mix, VOD, flow-TV, and video-sharing that can be either formally commercial or informal, or even illegal (Lobato, 2019; Lobato & Thomas, 2015; Spilker & Colbjørnsen, 2020).

Streaming also exhibits adaptive powers in the way it settles into, or occupies, the existing digital media ecology. Spilker and Colbjørnsen (2020) analysed the dynamics of five dimensions of streaming as a way to outline its evolution. Streaming borders and distinctions are under negotiation and in flux: (1) between legal and illegal, (2) between live and on-demand, (3) between user-generated and professional content, (4) between singlepurpose and multi-purpose services, and (5) between niche and general audiences. Niche appeal, for example, can be determined by language or geography, as well as an audience's preferences for content genre (Spilker & Colbjørnsen, 2020, p. 1221). For example, Netflix distributes three seasons of the Danish language post-apocalyptic youth TV-series *The Rain* as a Netflix Originals Series (Potalivo & Kirkskov, 2020). The fact that *The Rain* even ran for three seasons suggests an appeal, however brief, to a general audience outside Denmark. As such, they provide a conceptual framework that can be used to identify the characteristics of different streaming solutions and the differences between them (Spilker & Colbjørnsen, 2020, p. 1221). I find it useful to introduce a sixth time-sensitive dimension, a meta-dimension perhaps, which places streaming between the present-day and heritage. This dimension encompasses the other five dimensions in order to approximate how they will affect the process of collecting streaming content for long-term preservation. This holistic and unfolding approach to the streaming concept also provides a good basis for a definition of streamification.

Streamification

Streamification frames the process of adopting streaming solutions and streaming discourses that follow from the re-branding, reformatting, re-materialising, and similar changes to the representation, production, distribution, and use of cultural content (Aegidius, 2020). The role and ontology of software is encapsulated in the broader concept of *softwarisation* (Manovich, 2013). I adopt a critical media studies approach when I conceptualise the increased influence of streaming in the media landscape as streamification. This allows us to hint at a macro-level point of view regarding access models, along with a meso-level analysis of the collection of content from streaming media. The collection of digital cultural heritage happens alongside more or less regulated transcodes of content, and rewrites of metadata that reformats content, i.e., when users stream-rip and distributors prepare compressed transcodes to fit various bandwidths of streaming media. These examples detail the streamification of cultural content at the file-level (Aegidius, 2020).

Streaming services such as Netflix and Spotify are not, and should not be seen as, static cultural objects (Lobato, 2019), nor do they necessarily provide or use consistent distribution and production models from market to market. At the frontend, graphical user interfaces (GUI) vary across national borders, devices in use, user preferences, and the tracked history of use.

It is possible to analyse similar processes via the influence of singular companies, which is what Andersson Schwarz (2014) does when he traces the concept of *spotification*. Spotify has massively streamified digital music in a legal and easy to use service, as opposed to file-sharing. Andersson Schwarz (2014) argues that Spotify has done this by following the universal business model of enclosure, which mirrors moves toward a more institutionally sanctioned, regulated, and commercialised internet. Growing parts of the physical infrastructure are owned or controlled by service- and/or content providers; an increasing share of traffic consists of streamed audiovisual content such as video and online games (Andersson Schwarz, 2014, p. 115). This development suggests that spotification is a subset of streamification. Fleischer (2021) discusses universal spotification based on analyses of software and services that borrow the Spotify model, however, the Spotify model is far from the mainstay, and when discussing video streaming we could just as easily refer to the Netflix model to outline graphical user interfaces, infrastructures, and global markets of streaming. Fleischer (2021) does point out important differences between the two companies as he traces metaphorical overlaps between them that changed during the 2010s. The unifying reliance on a business model based on access is central to the process of streamification, and includes a continuous negotiation regarding what a company-named model actually stands for.

Method

Acknowledging and reflecting on my position and role is part of the method, because it situates the case and the arguments that this article provides about how streamification challenges the collection of streaming media content.

With a background in digital media research, I took on the project-based time-limited task of analysing the collection of digital cultural heritage at the Royal Danish Library. It was an internal position, meaning I was akin to an academic consultant in the Department of Digital Cultural Heritage. I received a monthly salary, and access to a desk in a shared office space. For the purposes of analysis, I had access to any internal information system deemed necessary for my position and the project specification. Confluence by Atlassian, which is an online collaborative platform used by the department for documentation (wikis), project management, and version control. My mandate was to interview personnel involved in archiving Danish Digital cultural heritage through collection, preservation, and making it available. This position provided me with an insider's view that I negotiated using an embedded ethnographer's method (Lindlof & Taylor, 2011). Analysing the collection of Danish digital cultural heritage made by the Royal Danish Library was a first step in elaborating a new digital collection strategy for the coming years. The strategy implements the current accession policy (Royal Danish Library, 2018).

The empirical insights garnered from analysing the existing practices led me to identify the testing of collections of streaming content via API from DRTV as a prominent example with which to discuss how best to collect content from streaming services. The Danish legal deposit legislation mandates that content published online must be collected by the library, while the publishers are mandated to provide access to the content (Kulturministeriet, 2005). This means that the access model that characterises the current availability of online content provides an acute challenge for the library. Mandel (2019) argues that such processes require new partnerships and new modes of working.

Collection through an API is one of the viable responses to the challenge. Another solution that is part of the library's response to the challenge is collection via contracts with intermediary aggregators or infrastructural actors. The Royal Danish Library has made a collection contract with the infrastructural company Cibicom.[2] Instead of taping all Danish TV programming[3] via the broadcast digital signal of TV, the Library receives all programming as a feed directly from datacentres that distribute them nationwide. These solutions mark a shift in collection practices from traditional "downstream" gathering towards "upstream" capture processes.

Testing the collection through the DR API

This case illustrates the challenges of collecting born-digital material that is only available online via streaming service. All curators at the library who are responsible for digital content under the access model agree that they urgently need to address the issue of VOD material. DRTV has existed since 2010. Even though DR has not had incentives to make content exclusively for DRTV, the same cannot be said of the other commercial streaming services that offer Danish content. The curators are aware that a great deal of Danish VOD content has not yet been collected if it has not been broadcast on traditional television that is covered by the aggregator contract with Cibicom. This corresponds with current analyses of the development of streaming as part of internet television that typically does not replace broadcast. Instead, it adds complexity to the existing distributive network (Lobato, 2019, p. 5).

I base this case on written documentation of the tests involving how to transfer content from DR via API that led to a recommendation as to how the library should proceed. I subsequently interviewed the persons involved and read documentation of internal discussions among the staff about how the API solution for collection could best integrate into existing preservation and access workflows and infrastructure. Note that I use the past tense to signal that we cannot assume that the content types and storage methods described in the case will remain consistent.

The content in question was until recently published via two individual flow-TV channels named *DR3* and *Ultra*. Due to budget cutbacks, the two youth-oriented channels were migrated to an online presence, i.e. dr.dk/ultra. The transition took place on January 1st 2020. Alongside this transition of flow-channels to digital only, DR also created *DR2+* as an online supplement to the flow-TV channel DR2 (Christensen, 2020).

[2] Cibicom manages the Danish national radio (AM/FM/DAB+)/TV (DTT) broadcast network, the nationwide LoRaWAN IoT network, and datacentres. https://cibicom.com/about-us/about-cibicom

[3] in Danish, sendeflade.

For the test, DR provided access to their on-demand catalogue service (OCS) API, which allowed collection of the relevant files, content files and metadata. DR also supplied a written guide for third party use of their OCS API.[4] The API provided access to two classes of metadata:

1. *The content class* provided links to video files, subtitle files, and image files. The content was chained in a structure of series > season > episodes.
2. *The publications class* provided further metadata, primarily about where and when the content had been published. This made it possible to ascertain how the content has featured in DR's content universes or publication channels, including flow channels.

The various metadata (JSON) and the image files (JPEG) used for thumbnails were stored locally on DR servers. The video files (MP4 h264) and the subtitle files (VTT) were stored on the servers of Akamai, a commercial content distribution network (CDN). This meant the obligatory generation of a special authentication token, along with specific endpoints provided by DR.

In order to collect the video files and subtitle files, it was necessary to write a script for Akamai token generation using one of several conventional programming languages (Akamai, n.d.). The script was uncomplicated and easy to create using the guides provided by DR and Akamai, even though the curators found the guide provided by the latter lacking in information. Here I quote the first Python script used in the test to let the reader assess its extent and details. It includes curator comments for sharing between the personnel involved in the test:

```
import requests
import urllib.parse
from akamai.edgeauth import EdgeAuth, EdgeAuthError

# Example URL from a DR OCS API post
url=<insert test-url here>
# Everything apart from the filename is lowercase, so I split
it and fix that:
components=url.rsplit(‘/’,1)
url=str.format(‘{}/{}’,components[0].lower(),components[1])
u=urllib.parse.urlparse(url)
```

[4] The guide is marked confidential. I only quote from the library's API test script and internal documentation.

```
#Setup token
ET_ENCRYPTION_KEY = <insert key here>
DEFAULT_WINDOW_SECONDS = 86400 # seconds
et = EdgeAuth(**{‘key’: ET_ENCRYPTION_KEY,
        ‘window_seconds’: DEFAULT_WINDOW_SECONDS})
# The token is generated from the path until the “username”,
before the “token”, with a wildcard to work with every
request. I am not sure if url may change between posts, but
I haven’t seen it.
token = et.generate_acl_token(u.path.split(‘token’)[0]+”*”)

#Get a response (this takes time)
response = requests.get(url, cookies={“hdnea”: token})

#Save the file
with open(‘video.mp4’,’wb’) as f:
    f.write(response.content)
```

During the test, the script was updated to perform daily checks of the DR metadata, which would enable the collection of any changes to the content, and new content. During the period of the test, January 1st-May 6th 2020, the API provided access to 557 files totalling 843GB in size from the channel references *DR3*, *Ultra*, and *DR2+*.

The DR API solution proves advantageous on two accounts. Firstly, the well-structured metadata meant that it was easy to read, in a technical sense. The test showed that it was necessary to collect all the above metadata from the two classes for the collection to be representative of the content and further identification of overlapping collections in the existing DR archive at the library. Traditional broadcast TV has formed the existing archival structures and metadata at the library, however, i.e. time of airing, which does not feature in the VOD metadata. This could mean a need for metadata restructuring (see challenges below). Secondly, the collected metadata and content files correspond better to the data structures of the bit preservation at the library (KB.dk, n.d.-a). Based on a test of content collection from the three digital only channels, the library curators assume they can collect the files in the full fidelity used for playback throughout DRTV instead of recordings of the daily DR flow-TV programming. This would eliminate the need for segmenting recordings of a full range of programs. It also greatly diminishes transmission compression artefacts such as glitches and signal fallouts.

The test also provided a list of actions needed for the collection of VOD to reach the first stages of preservation:

- Analysis of origin data structure and establishment of a database model to handle metadata.

- Creating a script to check for and collect changed and new content from the DR OCS API daily.
- Creating a script to generate Akamai tokens to collect and validate files.
- Creating a data structure that pairs metadata with the video files.

It is important to note that the Royal Danish Library is not obliged to collect metadata or metrics about views or interactions with the DRTV content, unless this type of metadata is considered published, i.e. in the form of reports, or on websites, in which case legal deposit is mandatory.

The frontend GUIs of DRTV adapted to various devices and screen dimensions are not collected through the API. This means that only parts of what users will see and interact with are collected; the initial thumbnail image of the content, the video file, and the corresponding subtitles. The GUI can be collected using screen capture or web recording (see discussion of collection methods in the section below).

The challenges

A range of challenges arose from the DRTV API test in addition to the above list of actions that only work to secure a workflow, a database model, thumbnail images, video files, and corresponding subtitles.

Publishers and distributors of video content or any type of content might provide their own API and have parts of their content managed by intermediary CDNs, such as Akamai in the case of DRTV. Collection via one or more APIs takes place on their terms; they set the conditions for collecting the content and metadata (a challenge that is common for all types of data collected via API but incidentally related to streaming services, which makes it a challenge incurred by streamification). Moreover, as part of a highly competitive market, the APIs are subject to frequent change.

It is important to determine what constitutes the stream in any given act of collection, whether it is for preservation or research. The requirements for simultaneous or subsequent metadata collection must reflect what constitutes the stream, as an optimal analytical entity, and what aspects of it to collect. This challenge can be illustrated using streams of tweets from Twitter.com. Tweets are only an aspect of the content on Twitter. Linaa (2017) argues that Twitter and Facebook content consists of updates, reactions, and metadata. These are all part of the data that progressively downloads to the cache on a user's device during use. The Twitter software needs these to show the interactions that link and propagate the tweets, the information about whether, how, and where the tweets are read and shared, however transient these interactions might be. This also indicates the dynamics involved in the on-demand vs. live streaming dimension (Spilker & Colbjørnsen, 2020). We must consider whether the stream is a progressive download of equally progressively recorded live content, as opposed to a progressive download of produced and recorded content.

Mapping the content and its metadata will challenge the collection of streaming content in a broad sense, especially when colleting from various services. In each case, the collector must analyse how the content and metadata is structured. When interviewed, the library curator noted that the metadata for VOD is not yet standardised. The same is true of music metadata and the industry struggles to rectify the matter (Morris, 2012; DDEX, 2020).

Following from the challenge of mapping the content and metadata, there is a need to sort technical terminologies. In this case, this conflicts with the daily distinction between digitised and born-digital material at the library. When is content "content"? Is it a publication, in the terminology of any given CDN and their APIs? Are the data structures similar between the various APIs? These and similar questions require time consuming technical translation work at the library in order for the collection, preservation, and access measures to produce searchable and playable content.

The internal analysis showed that streamed programs were sometimes re-run on flow-TV despite a frontend label stating the content was streaming-only. This was evident in the case of DR2+, which is not a new channel but rather a supplement to the existing flow-TV channel DR2. The curators learned that eventually DR2+ programs would be broadcast on flow-TV. This was only the case with 2 out of 15 content series produced for DR3.

The collection of digital-only Danish content from international streaming services, such as Netflix, HBO or Disney+, poses a challenge. The streamification of cultural content is acutely a transnational concern. The library cannot rely on the legal deposit of older streamed content eventually published as physical copies. *Polar* is a Netflix Film starring a famous Danish actor in the lead role (Åkerlund, 2019). The Netflix film is only available via the Netflix streaming service. For the time being, streaming services control national cultural heritage in international streams. We can speculate about a further degree of streamification should the library or national archive choose to rely on an external streaming service to partly or fully collect, preserve, and make available its digital cultural heritage. Such a scenario would mirror what Burkart and Leijonhufvud (2019) have characterised as the spotification of public service media based on the dematerialisation of the Swedish public broadcasting corporation's music archives into Spotify's streaming service.

In an analysis of streaming networks, Colbjørnsen (2020) identifies the actors that benefit from the established standards and protocols, and how rules of inclusion are negotiated. Citing network theory, including Castells (2013) among others, Colbjørnsen (2020) notes how the dominant players have control of the databases, which enables them to constitute networks and facilitate cooperation within and between networks. International streaming services will not necessarily respond to an individual nation's requests for access as mandated in national legislation. See below for a discussion of research collaboration with international services such as Netflix or Spotify.

Discussion

Collection through API is different from other collection methods, which each have varying strengths and weaknesses. Laursen et al. (2017) present and discuss four different ways of collecting digital cultural content: *still image screen capture, web recording, API-collection*, and *web harvesting*. They use Facebook as a case study. Their discussion is instructive, if we consider the content of Facebook, specifically the news feed, as a stream of cultural content. They note that we can generalise from the challenges met when collecting from Facebook. Building on the tool box-oriented approach provided by Laursen et al. (2017), I discuss collection methods for video streaming content. They specifically point out that methods for harvesting content published on the web cannot collect on-demand content (Laursen et al., 2017, p. 40). My internal analysis of the present day collection practices at the Royal Danish Library confirms that this is still not possible with their existing web-crawler solutions. The API-collection was thus tested as described above.

I restructured and nuanced their distinctions with the aim of suggesting ways to collect what we miss when collecting from streaming media via API. Firstly, I find it useful to conflate still image screen-capture and web recording. Secondly, I include stream-ripping. Thirdly, API collection is a formalised and automatised way of *asking for* the content. At other times, we can ask the producers directly for a digital copy of their content or engage in collaborative research. Lastly, I refer back to the case of collecting content from DRTV through API, and argue that researchers should always consider whether other actors collect what we require for research purposes.

Depending on the research question at hand, it is possible to pair these methods for the optimal collection of streamed content, its metadata, and its context. All the collection methods discussed below, including asking for the content, and, especially, collaboration, warrant a thorough consideration of possible limitations and biases in the resulting research.

Screen capture is the process of producing an image or video of parts or all of the content displayed on a computer's screen. It is good for capturing the frontend GUIs and the promotional context of the born-digital content that we cannot collect via API. In this way we do not have to reconstitute the born-digital by re-pairing content and metadata, in effect making it *reborn-digital* content (Brügger, 2016).

Today stream-ripping connotes illegal practices following the high profile take-downs of YouTube downloader software a few years ago (Van der Sar, 2017). We must remember that Apple, in the early 2000s introduced iMacs with iTunes and CD-RW drives, and later promoted cd-ripping as a necessary feature leading to the launch of the iPod. Perhaps we should change the concept to stream reformatting by way of web recording. I do not intend to promote stream-ripping as a rogue method. I caution that if we use stream-ripping software or include stream-ripped material as research data we must consider its consistency and quality. If we build collection software based on (perhaps pirate) code repositories of stream-ripping software we must be extra careful when vetting for research purposes. Metadata might be missing or compression rates might vary throughout a dataset as a result of the inconsistent transcodes performed during stream-ripping that may or may not opt for quick and dirty file conversions reminiscent of the "YouTube to mp3"-era (Aegidius, 2020, p. 8).

APIs can be seen as asking for content using a request procedure that is formalised in programming. APIs can be either open/public or private/closed. The Library of Congress provides open API-access to bulk data from their archive of historic newspapers. The Royal Danish Library provides equally open access to its digital collections (api.kb.dk/data/). The *Spotify for Developers* service (https://developer.spotify.com/) is an example of a private API that is nonetheless very inviting. However, as the name suggests, and this true of many APIs, it primarily addresses developers of third-party software. We cannot open the cache and collect the music files, as it is content that is bound by stream-only license agreements with the music industry. We can only set parameters for obligatory caching. However, the Spotify API does feature rich options for collecting data *about* music and content, i.e. playlists, which Spotify and its subsidiaries actually do produce. Spotify terms and conditions do not specifically forbid the use of API for research purposes, however, the Spotify terms and conditions for developers forbid the use of API that could facilitate "stream-ripping" or other functionalities that make it easier for users to capture or otherwise make permanent copies of Spotify content (Spotify, 2020). Eriksson et al. (2019) faced legal repercussions after having accessed the Spotify frontend with bots that did not stream-rip or download music files. I mention this merely as a note of caution, not to deter others from probing (I discuss collaboration with streaming services below). It should be possible to construct an application that collects data and metrics for research purposes, i.e., the software client connects to the Spotify server and the end users of the client would be yourself and other researchers.

As the case of collecting DRTV through API shows, it is important to test the possible parameter of access and the content provided. This includes determining what types of *content metadata* we are looking to collect alongside the streaming content. Moreover, we should consider what types of *processual metadata* we produce in the process of collating the streaming content. Vestergaard (2017) suggests we do several smaller pilot tests to determine the types of data that the API outputs, and whether two identical requests collect the same content. Snodgrass and Winnie (2019) argue that an API provider, and users of a particular API, should always work to consider the forms of programmatic accessibility, applicability, interoperability, transferability and durability that are provided by APIs. This includes API-collection. They note that APIs are important for the way in which they can give some sense of how data is being circulated, and made accessible and inaccessible. By studying and paying attention to the makeup of their structure and parameters we can also begin to detect the priorities, and thus politics, of the streaming services (Snodgrass & Winnie, 2019, conclusion section para. 3).

Asking the content creators or distributors directly for access to content is similar to a classic researcher's approach to gatekeepers (cf. Kozinets, 2015; Lindlof & Taylor, 2011). This method is viable in the era of streaming media even if we do not always have a legal deposit mandate as an advantage. Creators can often create copies or share access easily through the very same platforms and access points that regulate access in the first place. Equally, independent video creators could provide access, and act as guides, to the parts of the backend of the streaming infrastructure to which they have access. The number of content producers needed for a specific research project can quickly complicate this option.

The streaming services are technology companies, and they employ highly educated engineers who often have research degrees. In-house researchers might not readily engage in collaboration. Warning letters sent to researchers are a testament to this fact. Eriksson et al. (2019, p. 1) incorporated the e-mail they received from Spotify's legal counsel in their initial methodological considerations. However, we will not know if we do not try. Netflix and Spotify have internal research facilities. Netflix collaborates with universities, industry partners and standardisation bodies on the standardisation of video encoding and video quality as long as it ensures that they innovate and meet future streaming needs (Netflix, n.d.). Spotify allows its researchers to collaborate with researchers from public universities, as evident from the collective list of publications at Spotify Research (n.d.-a). Spotify has also made datasets available as part of challenges for anybody to participate in. It is important to note that these datasets are not necessarily representative of content on the Spotify streaming service, and must not be interpreted as such in any research or analysis performed on the dataset. They do offer an opportunity to conduct research on different types of content, as in the case of the Podcast dataset (Spotify, n.d.-b).

Web harvesting, aggregator contracts, and soon API-collection are methods used in the collection of Danish digital cultural heritage. The Royal Danish Library, upholding its mandate from the legal deposit legislation, has a position to collect, preserve and make Danish digital content available to the public to the extent allowed by copyright legislation. *Mediestream* is the Royal Danish Library's own streaming service. It provides access to content dating from the flow-TV era. As indicated by the case presented in this article, we will see the increased collection and subsequent availability of streaming VOD content from Danish TV. The DR API shows the work done to increment the availability of VOD content for research purposes, and eventually public access, pending the finalisation of license contracts between the library and CopyDan Arkiv. Danish students and researchers have access to radio, television and video-based commercials shown on television and in cinemas. The library does not currently provide API access to archived video content, however, when certain criteria are met the otherwise streamed content can be downloaded for research purposes (Mediestream, n.d.). This means Mediestream is more than a streaming service, and that it embodies the sixth dimension that places streaming between the present-day and heritage as an exigent factor. This ultimately means that digital cultural heritage was streamified (cf. Burkart & Leijonhufvud, 2019).

Conclusion

The stream is born-digital, however, collection through an API can result in many media files and various types of metadata files. Sorting, validating, and preserving this data and subsequently providing access to it, means that there is a possibility of the material being reborn-digital.

The library is facing a shift in the work load connected to the legal deposit of physical copies from the publishers to the library, which now has to collect even more material published online. The increased amounts of published digital content make it necessary to select what is collected. Since streaming is mainstream in Denmark, the question of how to collect streamed content based on access models has become an acute challenge for the curators at the library.

In technical terms, streaming is a form of media content distribution called progressive downloading that allows playback before the entire media file downloads to the client computer. Following the five dimensions proposed by Spilker and Colbjørnsen (2020), I find it useful to consider a sixth (meta-)dimension of streaming between present-day and heritage. This lets us characterise traits that will affect how we collect material that is transient and never fixed, and modes of access that shift and evolve.

Following the commodification of access, the move from owning media files to access is what essentially separates the era of download purchases and file sharing from the era of streaming.

I have elaborated on the notion of streamification using a critical media studies approach that centres on streams and access models. I related streamification to the concept of spotification, and in doing so discussed approaches centred on the influence of specific streaming companies, and software in general, on the collection of streamed content and its metadata.

The case examined the testing of content collection through the DR API. This method has two advantages. Firstly, the metadata is well-structured. Secondly, the metadata and content corresponds better to in-house data structures than previous collection methods.

The case provided grounds for a discussion of how streamification challenges the collection of digital cultural heritage. A number of challenges are involved in collection from streaming services. The curation process includes discussions of the streaming concept: how to map content and metadata, what is on demand, what is live streaming, and how these categories relate to the legal deposit legislation. Danish digital cultural heritage is spread out across various international streaming services, which challenges the collection process.

The next step for the Royal Danish Library seems to be testing for optimal combinations of collection methods in order to secure representative metadata and content. Building on the traditional methods presented by Laursen et al. (2017), I find it necessary to caution against the use of stream-ripping as a collection method. Collaboration with content creators or distributors, private as well as public, seems to provide options for gaining access and researching streaming.

This study is situated in a Danish context, however, researchers and archivists worldwide grapple with these issues. Streamification is a worldwide process that very much influences the collection of digital cultural content. Researchers, curators, publishers, distributors, and legislators need to find a level playing field that can mitigate the challenges of streamification for the collection of digital cultural heritage for the common good.

References

Aegidius, A. L. (2020). The music streaming metaphor and its underlying tangle of transcodes. *Popular Communication*, 19:1, 42-56, DOI: 10.1080/15405702.2020.1744609

Akamai. (n.d.). *Akamai / EdgeAuth-Token-Python*. Akamai. Retrieved Jan. 4th 2021 from https://github.com/akamai/EdgeAuth-Token-Python

Andersson Schwarz, J. (2014). *Online File Sharing: Innovations in Media Consumption*. New York: Routledge. DOI: https://doi.org/10.4324/9780203744420

Brügger, N. (2016). Digital humanities in the 21st century: Digital material as a driving force. *Digital Humanities Quarterly, 10*(3), retrieved Jan. 4th 2021 from http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html

Burkart, P., & Leijonhufvud, S. (2019). The Spotification of public service media. *The Information Society*, 35:4, 173-183, DOI: 10.1080/01972243.2019.1613706

Burroughs, B. (2019). A cultural lineage of streaming. *Internet Histories*, *3*:2, 147-161. DOI: https://doi.org/10.1080/24701475.2019.1576425

Castells, M. (2013). *Communication Power* (2nd ed.). Oxford: Oxford University Press.

Choy, S. C., Crofts, N., Fisher, R., Choh, N. L., Nickel, S., Oury, C., & Slaska, K. (2016). *The UNESCO/PERSIST Guidelines for the selection of digital heritage for long-term preservation*. Retrieved May 7th from https://unescopersist.org/2016/03/10/launch-of-the-persist-digital-heritage-selection-guidelines/

Christensen, K. M. (2020). *Nu kan du streame DR2+: Her er, hvad du kan forvente*. DR. Retrieved Dec. 29th 2020 from https://www.dr.dk/nyheder/kultur/film/nu-kan-du-streame-dr2-her-er-hvad-du-kan-forvente

Clifford, L. (2017). Stewardship in the "Age of Algorithms". *First Monday, 22*(12). Retrieved Dec. 29th 2020 from https://doi.org/10.5210/fm.v22i12.8097

Colbjørnsen, T. (2020). The streaming network: Conceptualizing distribution economy, technology, and power in streaming media services. *Convergence*, retrieved Jan. 4th 2021 from https://doi.org/10.1177/1354856520966911

DDEX. (2020). Retrieved May 7 2021 from https://ddex.net/

DR. (2019). *Medieudviklingen 2019* [Report]. DR. Retrieved May 7th 2021 from https://www.dr.dk/static/documents/2020/02/18/medieudviklingen2019_f1f4fafc.pdf

Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). *Spotify Teardown: Inside the Black Box of Streaming Music*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/10932.001.0001

Fleischer, R. (2021). *Universal Spotification? The shifting meanings of "Spotify" as a model for the media industries*. Popular Communication, 19:1, 14-25, DOI: 10.1080/15405702.2020.1744607

Have, I., & Pedersen, B. S. (2020). The audiobook circuit in digital publishing: Voicing the silent revolution. *New Media & Society*, *22*(3), 409-428. https://doi.org/10.1177/1461444819863407

Herbert, D., Lotz, A. D., & Marshall, L. (2019). Approaching media industries comparatively: A case study of streaming. *International Journal of Cultural Studies*, *22*(3), 349-366. https://doi.org/10.1177/1367877918813245

Holm, R. A. (2014). *DR er klar med betaversion af DRTV*. Retrieved Jan. 4th 2021 from https://www.dr.dk/om-dr/nyheder/dr-er-klar-med-betaversion-af-dr-tv

KB.dk. (n.d.-a). *Digital preservation strategy*. Royal Danish Library. Retrieved Jan. 4th 2021 from https://www.kb.dk/politikker-og-strategier/det-kgl-biblioteks-strategi-digital-bevaring

KB.dk. (n.d.-b). *Strategy*. Royal Danish Library. Retrieved Jan. 4th 2021 from https://www.kb.dk/en/about-us/tasks-and-goals/strategy

Kozinets, R. V. (2015). *Netnography: Redefined* (2nd ed.). London: SAGE.

Kulturministeriet (2005, Jun. 13th). *Bekendtgørelse om pligtaflevering af offentliggjort materiale*. Retrieved May 7th 2021 from https://www.retsinformation.dk/eli/lta/2005/636

Laursen, D., Brügger, N., & Sandvik, K. (2017). Metoder til indsamling af internetmateriale og deres effekt på senere analyser - med Facebook som eksempel. In K. Drotner & S. M. Iversen (Eds.), *Digitale metoder: at skabe, analysere og dele data* (pp. 31-49). København: Samfundslitteratur.

Lindlof, T. R., & Taylor, B. C. (2011). *Qualitative Communication Research Methods* (3rd ed.). London: SAGE.

Linaa, J. (2017). Udfordringer og muligheder ved at undersøge og sammenligne Facebook og Twitter. In K. Drotner & S. M. Iversen (Eds.), *Digitale metoder: at skabe, analysere og dele data* (pp. 109-125). København: Samfundslitteratur.

Lobato, R. (2019). *Netflix Nations: The Geography of Digital Distribution*. New York, NY: New York University Press.

Lobato, R., & Thomas, J. (2015). *The Informal Media Economy*. Cambridge, UK: Polity.

Mandel, C. A. (2019). *Can We Do More? An Examination of Potential Roles, Contributors, Incentives, and Frameworks to Sustain Large-Scale Digital Preservation*. Council on Library and Information Resources. Retrieved May 6th 2021 from https://www.clir.org/can-we-do-more/

Manovich, L. (2013). *Software Takes Command: Extending the Language of New Media*. New York, NY: Bloomsbury.

Mediestream. (n.d.). *Access for researchers*. The Royal Danish Library. Retrieved Dec. 28th 2020 from http://www2.statsbiblioteket.dk/mediestream/info/14

Morris, J. W. (2012). Making Music Behave: Metadata and the Digital Music Commodity. *New Media & Society, 14*(5), 850-866. https://doi.org/10.1177/1461444811430645

Netflix. (n.d.). *Encoding & quality: The best bang for your bytes*. Netflix Research. Retrieved Jan. 3rd 2021 from https://research.netflix.com/research-area/video-encoding-and-quality

Potalivo, C. (Producer), & Kirkeskov, J. (Director). (2020). *The Rain* 3. [Tv-series]. Denmark: MisoFilm & Netflix.

Royal Danish Library. (2018). *Accessionspolitik*. Retrieved May 7th 2021 from https://www.kb.dk/om-os/accessionspolitik

Snodgrass, E., & Winnie, S. (2019). API practices and paradigms: Exploring the protocological parameters of APIs as key facilitators of sociotechnical forms of exchange. *First Monday, 24*(2). Retrieved May 7th 2021 from https://doi.org/10.5210/fm.v24i2.9553

Spilker, H. S., & Colbjørnsen, T. (2020). The dimensions of streaming: Toward a typology of an evolving concept. *Media, Culture & Society, 42*(7-8), 1210-1225. https://doi.org/10.1177/0163443720904587

Spotify. (2020). *Spotify developer terms of service*. Spotify. Retrieved Dec. 28th. 2020 from https://developer.spotify.com/terms/#vi

Spotify. (n.d.-a). *Publications*. Spotify Research. Retrieved Jan. 3rd 2021 from https://research.atspotify.com/publication/

Spotify. (n.d.-b). *Spotify million playlist dataset challenge*. Retrieved Jan. 3rd 2021 from https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge#rules

Sterne, J. (2012). *MP3: The Meaning of a Format*. Durham, NC: Duke University Press.

Van der Sar, E. (2017, Jan. 4th). YouTube-MP3 settles with RIAA, site will shut down. *Torrentfreak.com*. Retrieved Jan. 4th 2021 from https://torrentfreak.com/youtube-mp3-settles-with-riaa-site-will-shut-down-170904/

Vestergaard, V. (2017). Datagenerering med API fra sociale netværksplatforme. In K. Drotner & S. M. Iversen (Eds.), *Digitale metoder: at skabe, analysere og dele data* (pp. 51-67). København: Samfundslitteratur.

Williams, R. (1975). *Television: Technology and Cultural Form*. New York, NY: Schocken Books.

Åkerlund, J. (Director). (2019). *Polar.* [Online movie]. The United States: Netflix.

Andreas Lenander Ægidius
Project Lead and Researcher (PhD)
Department for Digital Cultural Heritage
Royal Danish Library
andreas.aegidius@gmail.com