PRESERVATION AND REUSABILITY OF INSTAGRAM CONTENT - WITH EXAMPLES FROM THE PUBLIC SECTOR OF SWEDEN

Dr Rikard Friberg von Sydow
Södertörn University

Introduction

Many public sector organizations today uses social media to communicate with citizens. In Sweden, where government information is strictly regulated there have been several discussions regarding these organizations use of social media as the content is considered governmental information. In this paper I will analyze two city archives (Stockholm City Archive and The Regional Archives for Gothenburg and Region Västra Götaland) instructions regarding their use of social media in general and the application Instagram particularly. These instructions will be analyzed through a theoretical framework containing provenance, preservation and reusability. Instagram is particularly interesting because the main communication is done through pictures which can be hard to preserve, hard to record the provenance regarding, but that people often are eager to reuse. In the end two models for a practical preservation of Instagram content will be presented. Two models of preservation that take provenance and reusability in account.

Instagram content consist of a picture/video, a caption for that picture or video and comments made by other users. The pictures are added from a mobile phone (from camera or upload) and formatted into a specific size used by the application. Earlier this was a size of 612px, but now there are multiple sizes a user can use.[1] The caption, a short text connected to the photo or video, has a limit of 2200 characters. This is not an official limitation stated by Instagram, but a limitation discovered by users through trial and error method.[2] I use the term “content” here to describe what a user (with an account) adds to the social media. Content is information and I will use both the terms as they, in this case, really describe the same concept.

In Sweden all public organizations such as municipalities and government agencies have to treat their information flow carefully. Every document that is created or received by a public organization is perceived as an official document (“Allmän handling”). Such document must be released to the public on demand if the information in it is not subject to any secrecy regulations. In the legislation the definition of a document is very broad – it could be any type of information that a public organization holds. Automatically generated logfiles, maps, pictures, virtually anything that is perceived as information.[3] The information is also supposed to be preserved in an archive, if there not is a regulation from the National archives or other qualified agency that states otherwise.[4] This makes the discussion regarding preserving social media relevant. But this is of course not only a Swedish issue. In Espley et al (2014) “Collect, Preserve, Access: Applying the Governing

Principles of the National Archives UK Government Web Archive to Social Media Content" the question of public sector social media preservation is described like this in a British context under the headline "Why archive social media?";

> "The majority of government organizations now have at least one Twitter profile that they use for official purposes. YouTube has been used by government for several years to create and host video content that is embedded for streaming on the websites belonging to those government organizations. These two social media services are by no means the only ones used by government but they are by far the most heavily and consistently used. These services are also used to publish information not consistently duplicated elsewhere within the organization's web presence."[5]

One of the more interesting statements here is the uniqueness of at least parts of the information published in a social media context. Unique information that might not be preserved elsewhere and that might both have relevance today (from a legal view – what did we answer that citizen/customer online?) and in the future (how did public organizations communicate during the beginning of the 21th century?). This question is even more interesting if we take in account that we do not know how long in to the future the social media platform we use today, Instagram being one of them, will last.[6]

Instagram provides a possibility to download all posted content from your user account.[7] This has been a feature in the Instagram Application since April 2018.[8] The result is a zip-file with a series of files in JSON-format. These files contains all likes, messages, captions et cetera that has been generated through the use of the account. The zipped content also contains all pictures and videos the account has posted, and all direct messages (pictures or videos) that the account has had sent to it.[9] There are also independent scraping tools that let you download content from accounts that you don't are the owner off.[10] These types of scraping tools are off less interest in this study, but could be useful for public organizations if they collect content from other users or in cases were a public organization has lost control over an account. There is no research yet that mentions Instagram, but scraping through API, application programming interface, the programming interface that social media suppliers offers for professional users, have been recommended as an archiving tool for Twitter.[11]

Provenance, Reusability and Preservation

The theoretical foundation of this study is the concept of provenance as it is used within the discipline of Archival Science. The goal is to discuss preservation solutions that will enable us to preserve reusable content with its provenance intact. The reason to search for such solutions is to cater both to contemporary and future users of the preserved content. In this section I will describe the three concepts provenance, reusability and preservation.

Provenance is according to the Society of American Archivists: 1. The origin or source of something. - 2. Information regarding the origins, custody, and ownership of an item or collection.[12] Provenance, or *respect des fonds*, is a key concept in archival theory. There are multiple aspects of provenance but one is that what we perceive the archive as the natural accumulation of information in an organization. It is not – as in the case with a library – a collection of information, collected by the organization. It is the administrative accumulation of an organization, the mail sent to it and the protocol on which its representatives have written their decisions. To keep the provenance is to let the accumulated administrative material in an order close to its original order. Material from different organizations should not be mixed when they reach an archival institution, they should be kept sorted in relation to their original accumulator.[13] But provenance can be used in a broader sense and is a quite versatile theoretical concept. I will put focus on provenance and social media in the next part of this text.

If an organization produce content and communicate this through social media there is a possibility that it will want to use this content again. Reusability, in this study, is the possibility to use content again in any setting. It could be in a social media setting – If we want to reproduce a post on a new platform or modify it and upload it again on the original platform. But it could be another setting. Maybe we want to print content that has its origin in a social media post. To produce merchandise or for some other reason. The capability to do so depends on how the content is handled in the organization. In which format is the picture or video saved? Is it saved outside the social media, or is the content only existing through the social media account? Maximal reusability is gained if we have control over the material for our post. If we keep photos et cetera in high resolution arranged in such way that we can find the material again, and reuse it in another context.

If an organization by law must preserve its accumulated information in an archive, we need to treat this information in such way that preservation is possible. I will not focus on the technical part of digital preservation in this study. The study will instead try to focus on the administrative parts of preservation. This could be how responsibilities for different parts of an organization is formulated and how the process of preservation is organized. Preservation and reusability overlaps as criteria. Some of the challenges we face with reusability will also meet us when we try to a preserve archival matter. Preservation differs from reusability though, in that the organization does not need to exist anymore. The preservation could happen in an archival setting after the organization is ended. When preserved the archival matter does not need to be optimized for reuse although this

is of course a possibility too. Social media can be preserved in advance or preserve through the scraping or downloading of content from an existing account online.

Social media provenance and the Instagram content architecture

What is a reasonable criteria for provenance in relation to social media? The general reason for using social media is communication. A social media account and the part of the internet that you manage when you are a social media user is not an isolated product – as a web page can be. It would be worth nothing without its surroundings, the possibilities of other users to view and share your content and the possibility to comment what you post and also like or dislike it. Provenance in a social media setting must be something more than just the account and the content it produces. It must be related to the context in which the content is created. This is nothing new. Societal provenance is a concept that has been discussed before. The Canadian archival theorist Tom Nesmith has used the concept to describe archival matter from the colonial period in the history of Canada, and especially the relation between native people and colonists. The main idea in this theoretical description of archives is that the documents (often created by colonizers) cannot be viewed alone, they must be viewed with a very special context in mind, and will not tell the whole story as independent information carriers.[14]

I will now discuss some aspects of what I have chosen to call Social media provenance, a discourse of theory of provenance that targets social media. It is inspired by societal provenance in such way that I acknowledge a great relevance in the societal and technological surroundings of the preserved social media information. The societal part of this version of provenance has been described synoptically above and I will also get back to it later. Regarding the technological surroundings of the content the focus are the technological boundaries. How many characters can be used in a caption or comment? This is relevant because it is a boundary for discussions. Which functions exist in the application? Can you share other users content? Can you “Like” or in other ways react to posts? All these possibilities has to be described in the documentation to the preserved social media. If not, many clues to understanding the preserved content in the future will be lost.

In Instagram, there are parts of the content that you, as a user, can download, that is of value for the provenance. I will choose to call this an user-data download. We will go through the different files in this download to value them, and the metadata they contain, as providers of provenance.

When the ZIP-file, named after the username and the date of the download, ex: rikardfvs_20180704.zip, is unpacked you will see a number of files and three folders with photos and videos. The folder Direct, with all photos and videos sent by direct message, sorted in

subfolders by date. The folder Photos, with all photos taken, sorted in subfolders by date. The folder Videos, with all videos uploaded by the user, sorted in subfolders by date. All files are collected in subfolders – one folder per month (dated 201210 et cetera). The media-files are in the formats they were saved as in the Instagram application, with all added filters and with, in the case of photos, .jpg compression.

Besides the folders with media content there are 10 files in the mark up language JSON. JSON, JavaScript Object Notation, is a format to interchange data between different applications. It is based on the JavaScript programming language and designed to be easy to read for both humans and machines.[15] The JSON-files contains information connected to the account. Comments.json contains all the comments done by the account on its own photos. Connections.json contains information regarding users that the account has blocked, that the account has requested to follow and that the account actually followed. Included here is the date when the account blocked, requested or followed the other accounts. Contacts.json is contacts (i.e. other accounts) found by syncing the Instagram account to other social media (as Facebook). Likes.json that contains data of all likes the account has done, sorted by date and user. Which media, photo or video, that the account has liked is not stated. The file also contains information of liked comment, sorted in the same way and with only a connection to the user who has made the comment, not the comment itself. Media,json contains metadata regarding all media, photos and video. This metadata is sorted as “caption” (the caption of the media), “taken at” (the date and time the media was uploaded) and “path” which is the file path to the photo (example: “photos/201805/xxx.jpg”). Messages.json contains all messages that has been sent between the accounts and other accounts. The general information about the account – which e-mail address it was registered with, which date it was registered et cetera is saved in profiles.json. Finally saved.json, search.json and settings.json carries information about saved collections, searches that has been done by the user and settings regarding which other users that were allowed to comment on the posts made with the account.

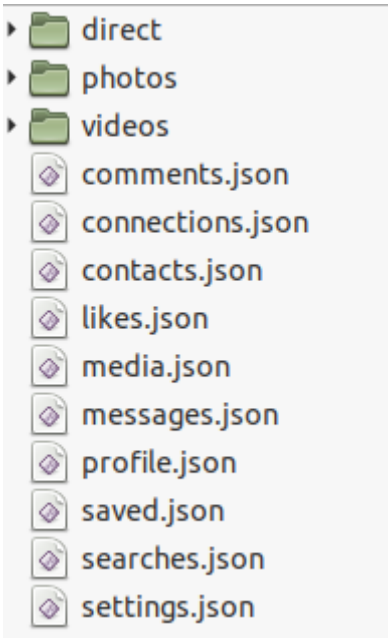

Illustration 1: Unpacked Instagram content

If the information, the content from an Instagram account, is preserved through an user-data download, we will need good descriptions of each JSON-file, to make any sense of them over time. This is also the case if we want to reuse the whole content , not just the media-files, in a closer future. As we can see, the files contains information that could be valued in determining provenance of the downloaded files “as an archive”. Especially those files containing profile information, likes, connections and comments. A problem regarding the reusability of the media files is that Instagram only provide them in a format that is affected by the application. This means that photos will be compressed, filters that were chosen during the upload will be permanent et cetera. If you want to reuse the media in another context this might be a problem. Generally the original media – the photos taken, videos filmed – would be in better quality in the format they had before they were uploaded through the application.

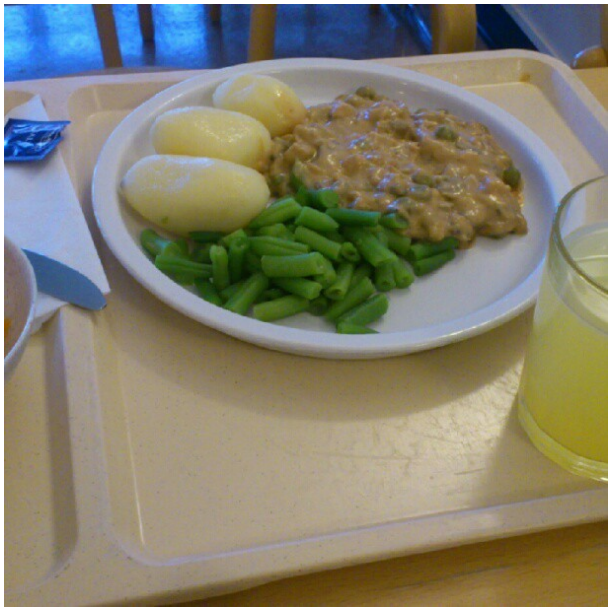

Illustration 2: A photo in .jpg compression formatted through the Instagram application with filters and 612px size. Uploaded in 2012 downloaded in 2018.

In addition to files containing information there are some other variables that will affect how we perceive the provenance. The application in itself and the social discourse that follows it– which technical functions exist and how they are used and interpreted by the users. What does a “like” mean? Is it a function you used on every post or was it a sign of deep affection? We must understand this to understand the archived social media. And if the “we” that observes the archive

exist a hundred years from now this knowledge must have been passed on to them. Through archival descriptions, interviews or other tools of documentation.

Instructions regarding official social media for the cities of Stockholm and Gothenburg

I must admit that I had some problems finding instructions for the use of official social media in Swedish public organization. Using the internet and professional contacts to try to trace such instructions. In the end I managed to find instructions from the cities of Stockholm, Sweden's capital, and Gothenburg, the second largest city of Sweden. The instructions from Stockholm is an official document published on the Stockholm City Archives webpage.[16] The instructions from Gothenburg is not published yet, but is an advanced outline of an instruction provided by an archivist at the Gothenburg regional archives.[17]

Stockholm City Archives has both a general regulation regarding its websites, but also a specific instruction regarding its different Instagram accounts. The general regulation is interesting in many ways. Here the City Archives states that incoming messages and comments on the official Stockholm City social media accounts according to Swedish law are supposed to be treated as official documents. They have to be archived if there is no other regulation that states otherwise, and they also has to be registered in special register if messages et cetera regards matters that would be under the secrecy act.[18] An example could be if a citizen provides information through comments or direct messages that would be considered under secrecy if it arrived through other channels (letter, e-mail). How should the archiving be done regarding Instagram? This is stated in a document linked from the regulation on the City Archives websites. Stockholm's official Instagram accounts should be downloaded using a stationary computer through “Print to PDF”. Each post should be downloaded using the print screen function. According to the instruction, written in 2018-05-31 there is no possibility to download all posts from Instagram, which was outdated information by then. The description about their Twitter uses the download all data-function described above, so possibly this is instructions regarding Instagram that later will be changed.[19]

The Gothenburg regional archives instruction is called “Bevarandeplan” (Preservation plan) and is more technical than the instructions from Stockholm City Archives. It is written in 2017-01-25 and thus over a year earlier than the download all data function in Instagram was implemented. It states that the Instagram accounts should be harvested once every month using Warprox, a web crawling tool that downloads the content and saves it into a WARC-file.[20] WARC was developed by the Internet Archive-/Wayback-machine-project and is used by among other the United States of Americas Library of Congress.[21] Important to note is that the instruction from Gothenburg regional archive also recommends JSON as a possible preservation format.[22] The format used by Instagram's own system for content download.

Warprox might be a decent method of preservation – even though the applications internal system for content download should be better. The internal system sorts the information in a way that is close to how it was used by the application itself, which should be more accurate both from a perspective of provenance and reuseability. I would not recommend the print screen or print to pdf version of preservation. Why? I will describe the problem with what I choose to call “Simple Snapshot preservation” in the next part of this paper.

Suggestion for two methods of preservation

In this part of the paper I will suggest two possible methods to use when preserving Instagram content. One of the methods, that I choose to call “Planned content management” is directed towards creating a reusable content. The other which I call “Content download” is more directed towards keeping the provenance of the content. But first – why is Simple Snapshot preservation not a decent method of preservation? The answer is: Because it is not suitable for either preservation and reuse. What you end up with after a Simple Snapshot preservation is pictures of content – screen dumps of in this case photos and comments. The photos are virtually impossible to reuse because the quality will be even worse than the media formatted by Instagram. Videos cannot be reused in any way – they will only show as a screen shot using this method. The text will be in picture form and thereby very hard to reuse. Regarding provenance the majority of the social media context is lost and we will need extensive explanation to convey this context to the future user. But as stated earlier I have two suggestions for preservation that I will propose. These two suggestions will be presented now.

Planned content management is my first suggestion. Planned content management do not trust the social media providers. If you choose this method you need to take care of the content yourself. When planning the posts photos et cetera are saved in good resolution, before uploaded. Captions are saved in a platform independent format. In this case we don’t take into account the social media context as part of the preservation. Focus here is reusability – not provenance. The good resolution and the platform independent text file gives us excellent possibilities to reuse the uploaded content in another setting. If someone later on, in one year or ten, wants to use the media content they have all opportunities to do so. Regarding provenance we could do some complementary work – using the Simple Snapshot version. But provenance is not the focus when we use this method.

Content download is the second suggestion. This method uses the possibilities in the Instagram application to download the media files and JSON-files described above. This result is treated as an archive. Our ability to reuse the material is bigger than through the Simple Snapshot method,

but we will still be hindered by the lesser quality of the media files affected by the compression et cetera that is implemented when a file is uploaded through Instagram. The provenance will be easier to sustain than in the Planned content management-method, but the method still holds some provenance-related problems. Even though we have a good view of the architecture of Instagram we will need some kind of program to show the content in a setting similar to the original context. We could call this type of program a “content viewer”. Still, with a content viewer, the content will carry some other problem connected to any web material. The broken links, in this case maybe best expressed in the comments from other users. Users whose own posted content will not be preserved in the archive. I will discuss this problem, inherent to preserved parts of the Internet, under conclusions below.

Of course there is a third possibility. We could use both methods and thus maximizing both the possibilities to reuse the information and still keep a larger portion of the provenance. I would call this the most preferable method. But it has a high cost both regarding work costs and storage cost. It might be bearable in a work environment were communications and archive divisions in an organization work closely together and cooperate around a unified reuse- and preserve-method.

Conclusions and further research

Always when we choose a method we must know it’s limitations. We must also know what kind of end result we want. Then we must let these two aspects intermingle until we find a solution that is both possible for the organization and gives us a preferable end result. In the end, if we strive for provenance we might lose some of the contents reusability. And if we strive for reusability we might lose some of the contents provenance. If we want both the archive will be considerably larger and there will be considerably larger amount of work accumulated in the end result.

Regarding provenance there will always be a problem to show preserved parts of the internet “as it was”. Internet phenomena, like social media applications and the world wide web are very much dependent on existing surroundings. It is very hard to preserve a small part. One effective way of showing this is through The Internet Archive/The Wayback Machine where websites from different periods are preserved.[23] Most of the links on any preserved site is dead, and linked content will be missing. They give you a hint of how things looked, but you don’t get the whole picture. Both the content and the context is mutilated. Using the Content Download method above (and applying some kind of viewer to let us see content together) we would get the same effect. We would see comments from users, but links to their accounts would be dead. The comments section would only be a collection of anonymous comments out of context. We might understand the context anyway, today, but it will be much harder in say 20, 50 or 100 years.

We can be sure about one thing regarding archived Instagram and other social media content. We will need a viewing application for archived social media. In the end nobody will be happy with some media-files and an assorted collection of JSON-files. We will need applications that can turn such content into something that is possible to view for a future researcher. As of today at least no official application to present downloaded content exist. Photos and videos can of course be viewed in ordinary media viewers, but the JSON-files containing comments et cetera are not as easy to view. There are some rudimentary viewers for these kind of files online, but they will only sort the content of the file in a way that give us a better view of the content – not present the content as it was presented in the application. In a experimental setting you can use these viewers to get an overview of a material (examples showed in illustrations below). In the future, at archive institutions that store many different organizations digital archives we will need applications that researchers can use to view archived social media content from different providers in an user friendly environment.


```
{"photos": [{"caption": "", "taken_at": "2018-05-22T07:43:08", "path": "photos/
201805/31c8bce6690b0bd106f82540c4f72864.jpg"}, {"caption": "Leon har fått medalj på Föris.",
"taken_at": "2018-03-05T02:18:24", "path": "photos/
201803/06e30f440ad9e91ec6a09372dcb4a967.jpg"}, {"caption": "Installerar Linux Mint på en
Thinkpad X220i (Thinkpad X201 med Linux Mint i bakgrunden).", "taken_at": "2018-03-01T12:56:03",
"location": "Uppsala, Sweden", "path": "photos/201803/440bbd355905a3170cbc3ceca14c0246.jpg"},
{"caption": "Det har varit långa förhandlingar men nu ser det ut som vi lyckats välja påve för
Tjörn. Habemus papam!", "taken_at": "2018-02-23T00:44:38", "path": "photos/
201802/8b209cebd1fdf5a7fbaaae17184561d0.jpg"}, {"caption": "Kallt ute! Tur att en har en eldig
fru!", "taken_at": "2018-02-22T01:13:35", "path": "photos/
201802/0f9192592947ea79ea166de59dab744f.jpg"}, {"caption": "Han ska sova nu men vägrar och vill
att vi ska prata om fiskar.", "taken_at": "2018-02-21T13:45:49", "path": "photos/
201802/4325bc8132c986efb04d0f99d340a568.jpg"}, {"caption": "Då var vi framme. Där bakom finns
ingången till isjättarnas rike.", "taken_at": "2018-02-21T06:25:05", "path": "photos/201802/
b3dfd70e4d3d8731f46caab6e4f2ceb1.jpg"}, {"caption": "4 år", "taken_at": "2018-02-21T06:23:44",
"location": "Rörastrand", "path": "photos/201802/8f590674b5cd50c53cc2bc5865f02919.jpg"},
{"caption": "Vi fick den här i en vegan box swap. Hibiskus och chili-té. Ska prova nu.",
```


Illustration 3: Part of the media.json-file from my own Instagram account, opened in a text editor

Illustration 4: The same file as in Illustration 3, now viewed through http://jsonviewer.stack.hu/

Finally, a suggestion for further research, that will be possible in at least a couple of years. Interesting then would be case studies of archived Instagram material kept at organizations or archival institutions. This research could examine both technical and administrative problems during archiving of social media material. But to do this we will need a variety of cases where actual organizations have archived their Instagram content, and to my knowledge this is not yet possible. But as soon as there are examples it will be an interesting topic to investigate.

1 “What is the size...” https://colorlib.com/wp/size-of-the-instagram-picture/ , viewed 2018-07-04
2 “What are your limits...” https://www.jennstrends.com/limits-on-instagram/ , viewed 2018-07-04
3 “Public Access to Information and secrecy act” https://www.regeringen.se/informationsmaterial/2009/09/public-access-to-information-and-secrecy-act/ , viewed 2018-07-04
4 Arkivlag (1990:782) https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/arkivlag-1990782_sfs-1990-782 , viewed 2018-07-04
5 Espley, Suzy et al “Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content”, *Alexandria*, vol 25, no 1-2, 2014
6 “How long will...” https://www.forbes.com/sites/jaysondemers/2016/02/09/how-long-will-todays-social-media-platforms-last/#410da5b51b72 , viewed 2018-07-19
7 “How do I access...” https://help.instagram.com/181231772500920?helpref=search&sr=1&query=download%20all%20my%20data , viewed 2018-07-04
8 “Instagram is rolling out feature...” Business Insider. http://www.businessinsider.com/instagram-data-download-feature-gdpr-privacy-photos-searches-2018-4?r=US&IR=T , viewed 2018-07-08
9 Descriptions of a download done via the Instagram application in 2018-07-04
10 “Instagram Scraper” https://github.com/rarcega/instagram-scraper , viewed 2018-07-05
11 Littman, Michael et al “API-based social media collecting as a form of web archiving” International Journal of Digital Libraries. 2018:19:21-38.
12 “Provenance” https://www2.archivists.org/glossary/terms/p/provenance , viewed 2018-07-07
13 Duchein, Michel “Theoretical Principles and Practical Problems of respect de fonds in Archival Science” Archivaria 16. 1983. p. 1ff
14 Nesmith, Tom “The concept of societal provenance and records of nineteenth-century Aboriginal–European relations in Western Canada: implications for archival theory and practice. Archival Science 2006 6:351–360
15 https://www.json.org/ , viewed 2017-07-07
16 “Att arkivera webb och sociala medier i Stockholms stad” https://stadsarkivet.stockholm.se/contentassets/5bb1b8749be64c5eb7e259cafe87ca11/att-arkivera-webb-och-sociala-medier-i-stockholms-stad-1_0.pdf , viewed 2018-07-08
17 “Bevarandeplan för webbarkivering”, 2017-01-25, in the authors possession.
18 “Att arkivera webb...” , p. 13
19 “Arkivering Sociala medier” http://www.stockholm.se/Fristaende-webbplatser/Fackforvaltningssajter/Stadsledningskontoret/Handledning-sociala-medier/Juridik-i-sociala-medier/Arkivering/ viewed 2018-07-08
20 “Bevarandeplan...” , p. 8
21 “WARC Web Archive File Format”, https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml , viewed 2018-07-09
22 “Bevarandeplan...” , p. 6
23 “The Wayback Machine”, https://archive.org/web/ , viewed 2018-07-10