Open Science - Topical Project

Viktor Johansson

Oct-Dec 2022

Contents

1 Glossary

This glossary is sorted after relevance for the project of the concepts and readability of the glossary itself.

Open science: In this report open science is defined through the concepts of open source, open workflows, open data and open access.

Open source: Means anyone interacting with the project's code, text etc, can view, modify and redistribute this part of the project.

Open workflows: Means the research done in a project is reproducible and replicable. Reproducible research in turn means that anyone repeating the research process using the same tools and data would get the same results. Replicability instead asks whether someone studying the same phenomena using another set of data (collected independently) would gain the same results. I also include the degree of collaboration possible in any project in this concept.

Open data and open access: Open data means that the data used in the project is accessible for anyone in human-readable and machine-readable form. Open access means the output from the project, e.g the project report, is openly accessible for anyone to read for free.

Open science principle: Any of the "open" concepts mentioned above.

Platform: A website that hosts a specific type of project in the digital humanities or open science whelm. The project is often constructed for interactivness, and therefore the website is often interactive as well, with searchable and viewable content. Often called a "digital archive".

Digital humanities: A part of the humanities filed that uses computational and mathematical methods to complete the whole or parts of a study or platform. These projects often contain open science principles to some extent.

Cultural heritage: Historical materials of some sort that is considered relevant for a specific group of people's history, often a nation's history.

2 Introduction

Open science is a relatively new concept. For the humanities, it is seen as an emerging trend. This project therefore aimed at understanding how open science principles are used and how often in the humanities field. Below some of the conclusions are summarized.

The extent to which digital archives and platforms exist, they have so far served a purpose of cultural heritage and sustaining its materials, while the time to do extensive research on these platforms are still short and therefore the research purposes are fulfilled only to a little extent, but is expected to grow over time. A singular number of studies have been made on the platforms I have looked at, and mostly by their creators. Political engagement of digitising cultural heritage often is an important part of why these platforms are financed, either by universities themselves, governments, or public or private funds, or interest groups such as theatre conglomerates. Also, the incentives to create platforms, and the need for it as few materials are already digitized, makes it logical to first digitize a lot. The types of studies done on the platforms often involve textual analysis of e.g Shakespeare, historical analysis of a painter such as Arosenius, and or computational, quantitative methods to recreate maps or similar digitally.

Digital humanities is a concept that shares some characteristics with open science in the humanities, employing similar methods, computationally and or matematically to do most of the studies and platforms. Digital humanities projects are not always following open science principles though, and those who do rarely follow all of them. From the projects I have looked at open access is the most common principle that is followed, possibly due to that some view open access as the only neccessary critera for open science. Open data is the secondly most often followed principle in the projects I have looked at. Often the raw data used for the projects are literary texts that now are in the public domain, which makes it easy to provide open data. By only providing the publication of a text one have used and from where it was taken makes it open data for those kinds of studies. Although open data is not common for projects which use newer sources as raw data and which have some kind of copyright license, though there are exceptions (Maignant et al., 2022). Also, if statistical analysis are done on the raw text data, open data becomes more complicated and less often used. Next, open source is relatively uncommon but still exists in about half of the projects I have looked at. This often come in the form of a CC0X license (either 1, 2, 3, 4), and a github page and or a Zenodo page or similar that is more or less documented. The problem with those aiming for and incorporating open source is often that the documentation around the code is not good enough to help understanding it. This raises the bar for how much technical knowledge one needs, and the time required to use the code for another study or to use the platform. A good example of documentation that yield both transparency and makes reproducibility and replicability possible is

(Maignant et al., 2022). Least apparent is the appropriate implementation of open workflows, especially collaboratively, but also within those boundaries created by the insufficient methodologies that don't make reproducibility possible. For each project, often a single group of researchers have created it with the help of previously physical and fragmented data to create a collection in one place of data of a specific region or by a specific author over a specific amount of time. Less often a group of universities work together on a project, and only very rarely the public are encouraged to contribute or can contribute without interaction with the previous creators on their created project. Instead, open workflows are enabled via the other principles of open science, which makes it possible to reuse code, data, and descriptions of methodology to either reproduce a study, do an own study with the help of the platform, or reuse parts of the code for a platform with another purpose in the studies that have sufficient methodologies and documentation. Since most studies if they use other programs or data also use open science ones, it is often possible to use all programs used within the project as well. It is a bit ironic when a project have used open projects as base for it's own, but does not follow all and sometimes any of the open science principles, which is not that uncommon for different centres for digital humanities.

The platforms are simple in one sense, only providing the raw data, while leaving interpretation, contextualization and the responsibility of the research quality to historians and literary researchers. Examples of this are that digital archives often look like physical ones, with similar metadata, but linked to other materials more often, collected for easier usability and accessibility. I have found one critical article about this by Kuys and Scherp (2022), which instead proposes a knowledge representation that contextualizes data via an event-model that connects events to each other and showcases multiple interpretations of them from different times and authors already in the database. Since one of the purposes of many platforms are participatory teaching it is a relevant point, since incorporating whole studies together with the platform is difficult, and fewer people will probably read an article than use a platform. At the same time, it complicates the process of digitizing and will also be difficult to achieve, cost more and take more time. Also, since this kind of more ambitious platform will have a higher expectancy of quality, contributing to enable open workflows for such platforms is more important than for more archive-like platforms. Wikipedia is of course an example which tries to somewhat do this, but lacking comprehensiveness compared to archives that often want to collect everything on a certain topic, or at least as much as possible.

It is clear that richer countries still are the ones doing most digital humanities and open science projects within the humanities, due to the requirements on technology, institutions and financing to start these projects. How to close this gap and fulfill the democratic values of open science is still an open question. A framework might help to guide the transition to specific type of studies at least (Hassani et al., 2019).

3 Method

This project was aimed at understanding in what ways and how much open science is used in the humanities field. This lead up to the following research questions.

RQs: What is done in the open science of humanities field today? And what part of digital humanities research is also open science and to what extent?

The first part of the study required an unstructured exploration of what kind of projects are done in the so called "digital humanities" and in open science in the humanities. Being new to the subject, this was a necessary step to grasp what it is, and what can be studied in detail.

A bit through this process, the centre for digital humanities at GU was found. Since the platform had multiple projects, looked accessible and was following many of the open science principles, I chose to study it more. Externally of this project I interviewed two of the coworkers there to get an introduction to their work and respective fields to be able to know what working with digital humanities is about. After this I started to read up on the different projects, their license, open source code, documentation, data and access, and then analyzed these aspects for three of their projects. Most time was spent on Aroseniusarkivet, which is the most ambitious of their projects which is already finished, to the extent a digital archive can be finished.

The step in the research process which was meant as the last one was looking for studies made that are following open science principles under the topic and keyword search of Shakespeare, which is the most well-known example of an open science literature topic. These studies though did not often follow all open science principles, at least the ones found in this project. They often used open data sources such as Shakepspeare texts, did a quantitative analysis via code that were not open, or used methods that were difficult to follow and would be impossible to reproduce. Especially studies that aim to improve the arguments for what parts of different plays are made by Shakepspeare or others, and when they were first staged, had the problem of poor reproducibility. Since most studies was not as ambitious in following all open science principles, this research path was left after some time. And also due to this, I won't take time to cite all of them. Also, it should be said that the quality of the studies being openly accessible within the humanities is more varied than traditionally published ones, at least according to my experience in this project. Due to this, a lot of work was about stating that a piece of research was not good enough to include in this report, and then continue the research process.

The actual last step involved looking for other digital humanities and open science platforms. I realized these were the most ambitious and the type of projects that most often did follow one or more of the open science principles based on the previous parts of this project. It was fruitful after finding the website Open Humanities Data that works as a portal towards digital humanities platforms, often following multiple open science principles.

4 Results and Discussions

In this section results and discussions of the results are intertwined on a case by case basis. For the more closely studied projects, open source, workflows, data and access are considered.

Here is a description of some of the commonalities between the platforms studied in the three coming sections (Nordisk Familjebok, Regnum Francorum Online and Aroseniusarkivet). Researchers can use the platforms done there to do studies with the help of them (due to permissive licenses such as CC-BY most often, and downloadable data which makes it open source, open data which is downloadable and open access by the projects adaption to most of the FAIR priniciples), or continue to work with them on their own, if they host it elsewhere. The public can view the projects and use it as they best like. The projects collect previously decentralized data and data with low accesibility in one space with higher usability, and a better interface, though some projects would be better with more user tests to improve the interfaces. It also makes this data searchable, and makes relatively simple but useful statistical analysis on-site. The projects could be even more attractive with a more active work to involve external contributors and stakeholders in developing the platforms themselves, to create an open workflow in practice and to continue to develop the platforms over time at a larger pace. This could be done most importantly via more detailed metodologies for the projects and the studies made with them, and through a better documentation of the code made. Also it could be noted that most projects are in Swedish, which makes it difficult to understand for other than Swedish audiences.

4.1 Nordisk Familjebok

This platform consists of two digital encyclopedias based on the two publications of "Nordisk Familjebok", "1800-talsutgåvan" published between 1876–1899 and "Uggleutgåvan" published 1906-1926. It is a simple project overall, text to visualization, maybe even the easiest possible looking at the task without thinking about its size/scale, therefore I first looked at this project. Source code for the database is on github. The documentation on github is short, and difficult to read for someone not alredy acquainted with the topic. The metadata used for

the project is not stated, or was at least not found for this study. The documentation is short overall. This is not an experiment, therefore software, hardware used etc is not as relevant to write out, since the database is what it is, but it could still be useful, but is absent. Regarding the raw data it is written out that it consists of two versions of the encyclopedia Nordisk Familjebok, and it can be reached through github or the platform website. AI is used to recommend lexically similar words, based on a Word2Vec-model.

The project can be considered open source, open data and open access, since all code and data is accessible for anyone and possible to change and redistribute based on the CC-BY license used for all cdh GU projects. There is nothing written on coding standards, best practice or encouragement to contribute to the project, or how to do it. Overall, the possibilities for open workflows is low due to this.

4.2 Regnum Francorum Online

Regnum Francorum Online (francia.ahlfeldt.se) is an interactive historical map of the kingdom of Francia where the user can choose what different things should be viewable and not. An example can be clarifying: if the user is interested in understanding the life of a specific king, then all important happenings in his life can be followed by highlighting the wars, trades with other kings, relationships with different monasteries and similar. A consequence of this is that also the royal families power over more than one kings life can be followed, how monasteries have gained or lost privilegies etc. Also smaller happenings such as the giving out of a specific coin are sometimes included in the map. This project is mostly aimed at researchers at least from my point of view, and much more than Aroseniusarkivet, since a bit technical understanding and understanding of why certain things are included as well will make it easier to understand and use.

Most national entities that finance and are resopnsible for reserving cultural heritage often delineates their activities to cultural heritage from within the national boundaries (GU, 2022). Due to this studies that study a geographical space which nowadays belongs to more than one country may be underrepresented in the research output. The regnum francorum online project overcomes these boundaries, and open source and open workflow and open data practices allows participants from multiple nations to cooperate to create these programs that can be helpful both in the work with national cultural heritage and in larger perspectives (GU, 2022). In a digital atlas, more sources in frequence can have room, and more different types of sources may be used without distracting the whole since different materials are used by each one using the available materials. Digital maps contain maps within maps, almost endlessly, one could say metaphorically. Berington Atlas was used as the main source for the raw data of the platform first, but this restricts how much can be done of course,

while the Roman Atlas Åhldal did afterwards (spanning a different time span and region) used the national sources from different countries which have now been digitalized to a larger degree, so the national and non-national level also can help each other in that sense (GU, 2022). Roman Altas also part of the Pelagios-project, which aims to combine archeological and text sources to be able to study the whole of the Roman Empire (GU, 2022). One necessary part of realizing this project is a strucutre of all the different data sources, which in the Pelagios project is the places which all material is related to (GU, 2022). The two maps Johan created can also be combined, to see how the Roman Empire infrastrucutre looked before the Franks was in force of the same place (GU, 2022).

The platform makes it possible to show more things than in a physical map, with multiple layers that can be shown or hidden at different times. It also makes it possible to focus the map on only the things which is of interest, or to explore what is in the map by being given examples of places where a happening have been registered and put into the map. It can be reached anywhere of course, and will require some amount of pre-work depending on previous knowledge, but with the benefits above as well as reusability if the study results are to be published as a similar map, it is free, and data can be reused. Also, the data shown are hyperlinked to so that the original source can be critizised quickly as well, and all of these original sources are also on the internet, on an open access basis. It is good to be able to get access to the original source as well, while many sources that could be used are not open access and therefore not included in the map either. Since all material is open source most data points that can be searched for and read about in the map is clickable and put into a context of when, by whom etc to the extent possible. CC-BY license used as for the other projects at CDH GU.

There are about 6000 entries in the database according to the home webpage. The interface is a bit more complicated technically than Nordisk Familjebok, both due to the larger complexity of the subject of study (which necessarily makes it more complicated), but also due to design decisions. The design decisions are all made by one person and without much testing, as the project was done early (2008-2012) which meant both the technical possibilities and the number of interested people in this kind of projects were hindering efforts of testing probably, but also the prioritizations of the researcher have put user testing on a relatively low priority it seems like. Machine and human-readable code is provided and downloadable.

4.2.1 Open Source

The license make it open (CC-BY). Anyone can view, modify, redistribute. All code is downloadable. Code is on a github I found through knowing the name of the developer and searching on github. Looking at the github, if there is no new

one as well, the projects have not been contributed to during the last 5 years, so the community is not very active. The github have not any description for how to contribute or read the code or for setup etc. Hardware not specified. The map can be used with different online map interfaces, eg Leaflet, while Åhlfeldt's part of the project is mostly a database that is then connected to this map interface and that works together with it. Leaflet has a BSD 2-Clause (a permissive) license, which only requires anyone modifying or republishing code with Leaflet included to refer back to the copyright as well as in binary form, and to do the same about a statement that puts away all responsibility of the different kind of damages after usage of the library. Javascript is used for both map projects, it is a GNU (copyleft) license, while the maps are CC-BY which is permissive, which should not be possible for a GNU project (copyleft can't be translated to permissive in derivative works), even with permission from the original contributors.

4.2.2 Open Workflows

The platform and map can be used in combination with other maps. Metadata following the standard for the field is used. The main platform is not a community work, but rather the work of one person that other can download and continue working on on another website, but with little help from documentation.

4.2.3 Open Data and Open Access

All data open, links to the data in the map. The maps are accessible by anyone. The sites for the maps hosted on the developers or the institutional website.

4.3 Aroseniusarkivet

Aroseniusarkivet (Aroseniusarkivet.se) was both aimed at creating a platform, to analyze it from a perspective of how to make it for both researchers and the public, and also at analyzing how an artist's work is affected by moving to a digital platform and different viewers perception of this. Also, how the artist himself have been studied and exhibited previously and why are examples of information that the platform include and can help to study. The project lasted for three years plus another year to complete certain parts. An interesting point is that a digital archive could become the definitive version of an artist's work, since once it is digitalized a lot of work and time have been put into this and the physical copies will still be as inaccessible as previously. Conservatory is not needed for it, the costs therefore are much lower in that regard, and therefore the choice of implementation, how a photo is taken, what standards to use for color, light, resolution and production are choices that could sustain for longer

making the decisions more critical.

The parts of the project include a website or platform, two apps (one with alternative images to a children's book made by Arosenius, one with a reconstruction of Arosenius' residence), and a github.

The platform has a clarifying introduction page that introduces what is shown (photo, art, etc.), from which collections (Gothenburg University, museum or private), and how many (4700) digitized works are included in the collection, documents, etc. It explains the platforms two different purposes, both for the general public, simple and easily accessible, and for researchers with a search for metadata. It also introduces the four modes on the front page, gallery, timeline, image cloud (AI interconnects material based on various parameters of a painting such as color, lines, patterns, light), word clouds (interconnects different images through metadata, or rather for statistics on the presence of different metadata, remote reading). Search, people, categories/keywords, can combine all these together and for any of the four views. It is well described even for non-experts here, and also info about why you get the view you get when you click on an image. There is reasoning about the reliability of the material, such as uncertainty, and then about how good the color reproduction is, which is based primarily on where the photo was taken (private collections, museums, GU or auction houses, which can be found through the metadata for each photo).

All participants in the project with roles are displayed (including those from Centrum för digital humaniora at Gotherburg University, Gotenburg Univeristy Library, Nationalmuseum, Göteborg Konstmuseum and Litteraturbanken). This is good for transparency, to be able to contact specific persons, and to understand what competencies have been part of creating the platform.

4.3.1 Open Source

That anyone can view, modify and redistribute a work is the ideal of open source in theory, in short. Aroseniusarkivet follows this, but it seems like it is hard to change the project itself on the website, since the official solution is financed by multiple organisations (Riksbankens Jubileumsfond) and (Vitterhetsakademien) that probably want the product to stay up to their standard and views. The project is a bit on ice after financing have stopped after the initial development was mostly finished. The platform itself is therefore not decentralised developed, but other projects continuing on their work on other websites etc are encouraged and possible. Peer-review of the studies using the platform comes from the journals where the studies of the platform are published, while the platform itself is developed based on the developers decisions. The ones continuing the work on the database in their own projects refer both to the studies and the platform, but some may probably just use the platform to recreate a study of

another artist, and if so, a more thorough outside view of the design decisions of the platform would be valuable. At the same time it is difficult to understand how a peer review process for an open platform would look like, and if it would be better than leaving peer review to the studies coming out of the platforms. The platform is instead evaluated via stakeholders, financers and contributors, and the users use it if they like it and not otherwise. Github exists for all parts of the project, more in open workflows section. Public domain for Arosenius's works themselves unless otherwise stated (CC BY). This means you can download, re-distribute, use material commercially, etc., but you have to refer to aroseniusarkivet.se and the collection the material comes from when doing so.

4.3.2 Open Workflows

Someone within the project or someone that want to seriously contribute could probably get in contact with the previous developers, read the github etc (arosenius-api, 2022). Though there are still open questions such as on coding standards and how actively contributions on e.g git will be evaluated such as when and how pull requests will be handled if handled at all. I think to know about eg coding standards you would have to contact them as said, while the github seem to mainly serve the purpose for others to view the code, and to use it in their own studies or similar projects with other artists, while it does not encourage or make it easier than necessary to contribute to this project specifically. Parts of the design decisions are on github, and some are on the website, and some are in other publications in newspapers, books or journals that are freely available but through other links. Information about the project's hardware has not been found by me, and so discussing virtual machines or similar to create the same environment for the code is not possible. The project though has considered how to make the platform useable for multiple different screens, browsers, and quality of computers. All technical solutions are open source according to the github. There is one github for backend, one for frontend and one for "admin system for the database". The readme on github also says that you can use all parts for other similar projects but that parts of the program are hardcoded etc, so it requires that your data (e.g photos) can be described according to the same data points and on the same shape if you are going to use exactly the same program.

On another note, there is a page on the website named Forskning which links to the publications made within the project, most of them are open access. All publications coming out of the project, all of which are done internally at cdh GU are stated. This page makes it easier to overview the total output of the project, and where to learn more, making the process more effective for both researchers and the public. Searching in a scientific search engine, only one article citing Aroseniusarkivet was found, while it also has been cited in newspapers. This could be either due to little response or wrong citations, or a wrong from my side. According to the project end report the research group have went to

a lot of conferences, so the outreach itself should not be the problem. As Malm also wrote a book about Arosenius that cites Aroseniusarkiet, another source of outreach is found. Apart from university and museum websites, scientific and other publications in the form of articles and books, it is difficult to see a possible source of outreach. This is a problem especially for projects that are meant for the public, and without outreach parts of the value of being openly accessible is lost, since no-one still access it. A section on the page invites you to submit more works by Arosenius, encouraging that kind of interaction.

Regarding reproducibility and replicability the project would not be possible to replicate exactly due to a set of reasons. 1. The method for the platform is described in multiple places which makes it a bit more blurry than it could be. 2. The private collectors are anonymized due to personal integrity, which is a few hundred of the total 4700 works, so most of them are available physically, while all are digitally. 3. The method, e.g design decisions of the website are sometimes not obvious, and how choices were made regarding how to program the backend, frontend etc is not described in detail. Also, some of the studies methodology is not possible to follow to reproduce the study. This affects the research quality rather than the platform itself, and it affects research rather than the public group directly.

It should be said that the studies coming out of or from within the project besides the platform/website are all describing their method like in a typical science paper if one could say so. Eg Westin and Claésson (2017) writes about multiple different sources of information, methods etc that was used to recreate Arosenius house digitally (colors on the ground, maps, paintings, sound with the white noise of today's surrounding removed), while there are many steps in this that would not be possible for someone to follow exactly, but one would have to guess how e.g the white noise of today have been removed to create similar sound to when Arosenius lived there, or how all parts can be put together to a whole which is very difficult to understand at least for someone not acquainted with this type of study. The technical difficulty of the involved tasks probably would make stating all of them more complicated and time-consuming as well, while not doing it does not make replicability any better.

The project end report by Malm and Westin (2020) includes technical details on the interface design of the website and metadata, while not really going into technical detail. In an email conversation with one of the authors he referred to github for more technical details on the code, which as mentioned exists but is insufficiently documented for reproducibility of the study in itself. Some of the technical work of scanning material is also found in Westin (2021), e.g what cameras are used for different materials to picture and why. In summary, the methodology is not collected in one paper, but is collected on the webpage (Aroseniusarkivet, 2022) in different papers and with different degrees of clarity and robustness on different parts of the study and platform. The standard thorughout is high, and not worse than most studies, with some parts of the

study being close to replicable immediately, like Westin (2021) with the choices of scanning vs photo and resolution and metadata etc, while others are still very unclear. Some parts of the technical work could be replicated, while other parts of the study could not, such as programming decisions and the creation of the simulation of Arosenius house, which are only partly explained in (Westin & Claésson, 2017). This could be due to difficulties in explaining the methodology, poor documentation, or too little resources combined with other incentives than those that would make replicability and reproducibility more important.

4.3.3 Open Access and Open Data

The platform is free to access and download etc, as well as the data (e.g paintings). The studies that have been made using the platform are published in green open access repositories often, i.e both at the publishers website but also the institutional website of GU, including less scientific output, which is important to make the database even more useful also for those not having the possibility to access non-open access journals. The data management plan is not publicly available, if it exists, while the decisions are quite well written out in a more accessible version on the project website. A description of the metadata contained in the database is also available in text. The standard used for metadata is explicitly stated and chosen to adapt to the research field standard and the type of database, which increases usability and coordination across organizational boundaries. The long term strategy for the storage of the data of the project is the only part of the FAIR principles missing on an overview level for someone studying the project from the outside, while in detail one would also gain help from a data management plan, e.g who has the responsibility of maintaining the data, and how it is done, which would at least teach the ones reading it about how data is stored and by whom, possibly increasing the chance of further looking into what other data is stored at the same place.

The license (CC-BY) means that one may use, disseminate, download, process, etc., commercially as well, but must refer to the source while doing this and if one publishes: 1. Give proper recognition 2. Enter hyperlink, 3. Indicate whether changes have been done 4. 1-3 must be done according to best practice, in such a way that you do not think eg that the Arosenius archive supports the additional changes or the contributor of these changes. This license is often used when the author wants to allow as much use as possible of his work, but at the same time, in accordance with scientific practice, require reference of the work or article, as well as to state what is and is not the appearance of the original work. This is important for the researchers status as a common sign of research achievement. Alternative licenses could force the person making the processing to use the same license, to keep it open, i.e a copyleft license. The license used is permissive and not copyleft since the same license need not be used for derivative works.

4.4 Other centres for digital humanities

Cdh Princeton (cdh.princeton.edu) have many network studies (Victorian realist novels, Sur, Shakespeare company and library Paris) on top of the digitalize and visualize type of studies made by cdh GU. These studies connect different persons to each other in different ways. Antoher type of project takes a picture as input and creates a Mondrian painting from it. I also studied a couple of other centres for digital humanities, which did many studies using quantitative and digital methods but did not use open science principles, also often in terms of network theory, linguistics, translation, maps or lexicographical materials.

4.5 Theatre studies using open science principles to different degrees

In this chapter a few examples of platforms and studies being made using or following open science principles to different degrees will be described. All of these have in common that they wasn't developed by the Centre for digital humanities at GU.

4.5.1 Drama Critiques' Database

This platform stores and makes openly accessible a collection of data of 27 000 reviews of theatrical plays in London 2010-2020, from journalists and blogs, on the platform Zenodo, and the platform's website. It makes multiple kinds of studies possible of course, and was first meant to be used to compare the discourse of journalists compared to the blog sphere (Maignant et al., 2022). The data files are ambitiously done, named well, their are both notes on metadata, the data, and the data itself (Maignant et al., 2022). The code is open source but less well documented in github. The methodology of the study is exemplary, reproducible etc (Maignant et al., 2022). It it is clear about choices made, explicitly states where the data is from and why, considers all steps of the research process, how long time it took, what programs were tested for use and then actually used, metadata used etc (Maignant et al., 2022). All code is open as well, though the externally used ML program is not given in Zenodo, dcreasing the research quality and transparency (Maignant et al., 2022). The study categorized 1000 of the 27 000 reviews manually, to have training data for an "ML program" that categorized each reviews so called structure and then categorized the rest of them. This step in the research process was the only part that could not be found in Zenodo or github. The project also had a pedagogical ambition, similar to previous platforms (Maignant et al., 2022). Therefore, descriptions on how computational methods to analyse texts can be useful and how it is done were made. This could help in increasing the potential audience as it is easier to understand both the platform and reuse it's methods. More than 1200 hours of work was put in for the db according to the article, which shows that it is timely and costly to do these kind of platforms, another

restricting factor for the diffusion of open science practices in the humanities. The license used is CC04 for the data and the platform data, e.g the texts, again as is common in the humanities.

The study based on the platform describes it's theoretical grounding in the open science paradigm of theatre studies via Bardiot (2017), and her three categories or types of studies that can be made within this paradigm. The first one consists of projects that re-examine theatre history as a global phenomen with computational methods, e.g Mohnike (2020) and Holledge et al., (2016) studying the reasons for Ibsen's fame and one of his plays fame respectively. The second type considers theatrical texts, e.g the Shakespeare studies mentioned earlier, dating or finding the influences and authors of specific plays. The third category visualizes data and is made for interaction with this data, similar to the studies made at cdh GU.

The study itself consisted of three sub-questions, showcasing the platform's wide usability in different studies (Maignant et al., 2022). The fist one aimed to explain the eventual differences in linguistic style between blogs and journalists. The tags based on ML, and word counts of different sorts were used to analyze these differences. Here, bloggers had a much larger usage of the word I. The second experiment was based on sentiment analysis via the web program text2emotions, which categorizes texts based on different semntiments. Here, no differences were found between the groups. Thirdly, they looked at the geographical differences of where journlists and bloggers go to see theatre. Bloggers had a more widespread area of visits, naturally due to the disparate interests within this group.

Most surprisingly, the database for this platform publishes most of the articles in the open despite copyright issues, from both individual bloggers but most remarkably from large newspapers such as Times and The Guardian (Drama Critiques, 2022). How this was done is described via that they have taken their data from another publication, Theatre Record, that have collected all theatre reviews of stagings in London over the course of 30 years, but how this publication have done this is not said. Looking at that Theatre Record's website, subscription is needed to access the material, and the license for the articles is not stated. It is therefore still an open question how the license for the journalist's articles was gained. For the bloggers they have requested the right to republish, and the ones not responding are the ones still not openly accessible on the platform website (Maignant et al., 2022). Speculating how they have solved this it is possible that either Theatre Record's or themselves have got the consent from the journalists and newspapers to publish their articles as well. If so, England's theatre's culture and attitude towards open science looks very different to e.g swedish newspapers that require you to travel to Kungliga Biblioteket to read physical newspapers, it is not even possible to send them to you. It is possible that this is due to that it increases the interest in theatre overall, and that other articles are non-open access as well, for economical reasons or due to

institutional rules and decisions that require theatrical reviews to be open for everyone in England.

It is said in the article, "Finally, a part of Drama Critiques' dataset and all the programming scripts that enabled us to carry out the technical analyses are in open access on GitHub and Zenodo. Anyone can thus run the algorithms on the whole corpus again to better understand the results or adapt them to their own data" (Maignant et al., 2022). Therefore, open workflows and open science is encouraged, increasing the impact potentially. The github have code for the project but less easy to understand what is what, no documentation or description of the files in github itself, and insufficient description in the article to actually understand the github (Maignant et al., 2022). This makes the open source aspect less good, and could be improved by spending time documenting better. This though puts an increased cost on the researchers, or requires a community-driven effort, which in turn requires outreach, which is difficult to create for most research due to the resources available and the interest of the public being not on reading research most often.

4.5.2 Two Ibsen studies

Also, in some cases open data repos are of course used to do non-open science studies, which is the case in the two following ones (Mohnike, 2020), (Holledge et al., 2016). Quantitative digital methods are used, via others open data, but the books coming out of the projects are not open access or open source.

The first book have used a database containing possibly all known stagings of Ibsen plays around the world, with searchable content due to relational data being included, to analyze how Ibsen did become an influential actor on the global theatrical market, both due to his literary qualities but also institutional decisisons such as allowing unauthorized stagings and adaptations of his plays, and the likings of this from publishers, actors and translators (Mohnike, 2020). The author notes that the use of a database is also subjective, due to the choice of its structure, data, etc, while it requires extensive qualitative analysis after the initial quantitative one. This is an important aspect for anyone using the platform to note, e.g in applying the same critique of the source as towards other sources used in a scientific report. Also, the problem of changed national borders occur here as well, complicating which country each staging applies to, which may create wrong conclusions in some instances, or at least create noise in the data (Mohnike, 2020). Instead language could be used as the starting point of analysis, but then the country influence is less clearly incorporated, which may be even worse. In the review of the book, critique also points at the absence of some plays that have been documented elsewhere, and such small errors if they are not described may be due to reasons of time, or could possibly be seen as a small wrong when the content is so large, and if not it is good that the wrong have been pointed out by others and can be changed. This is also true for the

lack of some metadata (Mohnike, 2020). While these are problematic aspects, when the data is open, wrongs can as said be seen, pointed out by others and hopefully changed by someone that know how to solve it, and if it is not solvable provide a description of why the db lack some part of the data. Regarding the relatively small errors in the db, the study made is comprehensive and reliable, with better possibilities of improvements due to openness (Mohnike, 2020).

The database used is created and hosted by a university, while the program used to create the relational model in the database was taken from a project at another university (The Virtual Ibsen Centre, 2017). So, here the open access of both the db and program have yielded value, while the choice of journal and publication type makes it non-open access (Mohnike, 2020).

Looking at the db "Ibsenstage", which was done by The Virtual Ibsen Centre (2017) shows that a map is the start for the user. By clicking on one data point, we see that each staging include data on dates, theatre hosting the play, all actors and others involved in the play etc. As such all metadata included is clear. The search function is though not comprehensive as the only applicable filter is on country, time of the staging, and contributor of the material. So again a map with a large db in the background is used to create new posibilities of research. And as many other of the db's I have studied, the work put into the db is probably much larger than the amount it has been used, but over time this might change. On this note, standards for metadata is valuable to have early on to avoid the large amount of extra work later in time when there are more platforms that may need to change to a new standard, while the knowledge of how to construct the standard increases over time, and also back-compatibility is helpful here as a possible complement to having to change everything to a new standard. Also, new content will be useful to a db such as the one collecting theater reviews from 2010-2020, that wants to stay up to date as is stated as one of it's purposes.

The second study instead of looking at Ibsen's whole authorship, used the same db to study a specific play and how it has spread globablly through time, and why (Holledge et al., 2016). The reasons include social, economical and political reasons, and is based on data from the db. The data being open, ones again make evaluations of the research done more effective, while it is not open access. Having access to the raw data of stagings of Ibsen's plays probably made this study much more effective, saving parts of the work. Together with the first study, it is possible to start to see how the costs of a platform like the one for Ibsen is a fixed cost, while the value increases a lot by scale as more persons use the same platform multiple times instead of duplicating work among many.

4.5.3 Critique of the typical platform

Another article at open humanities data was looked at to know more about the platform and due to the subject of critizising other platforms similar to the ones studied by me. This one criticizes the usability of digital history research, and digital humanities in general, for not providing the knowledge representation which history is supposed to as a subject in terms of complexities arising from different time perspectives, relationships, and context of each new situation (Kuys & Scherp, 2022). The authors recognize that this is difficult, that present-day approaches does not even aim to do this, but rather is happy with providing only the most basic elements required to do anything. They therefore seem to propose more ambitious approaches, and less projects to achieve this, more time per project, everything else equals requires that there are fewer projects to compensate for the increased time in each project.

Also, they try to do this by creating an own example. It consists of an "event-model", modeling how events are connected to other events, or are part of other events, and persons are part of these events and each event and/or person is described in different ways (Kuys & Scherp, 2022). "Interpretations" are used for more uncertain data. This is different to just having persons at different places without context, or with only one interpretation, since the model should also incorporate views of different historians and involved persons on the event, to increase the sense of reality in how events and history is interpreted in real life, beyond history studies (Kuys & Scherp, 2022). Contradictions are also part of the model, with a specific knowledge representation.

5 Conclusions

There are many projects made within the whelm of open science in the humanities, some overlapping with the related concept of digital humanities, while still most of the studies in the digital humanities don't incorporate all of the open science principles, open source and open workflows is the least common, while open data and access is relatively common. The projects, platforms and studies made often consists of digitally archiving a cultural heritage of some form, today most often text form, which has previously been inaccessible physically in different places and by closed institutions. This material is then made searchable and viewable, by the use of metatags based on the field of study's standards. After this the texts can be used fot studies using previously not possible computational and statistical methods, such as with digtial maps and museums or categorizing text styles and dates. The platforms can also be used for traditional types of studies within the humanities, that only focus on close reading and don't use computational methods before that.

Open source is quite common due to the relatively little effort of publish-

ing code in itself, and since licenses like creative commons are encouraged from funding, universities and politicians. Also, coders are used to using e.g github for working with a coding project, and if so, making it open source just requires making it public. Also, clashes with licenses from other programming libraries and programs used need to be thought of, but it does not seem to have created problems for the projects I have looked at.

Open workflows is probably the least common due to the proved difficulty of replicability and reproducibility in many traditional studies in different science disciplines as well. In traditional science this have been due to little resources for finishing up a good-looking documentation and methodology, less education on this part of research, and cultural value that don't prioritize replicability and reproducibility as high as it could be. All of this might so far have followed into open science, while over time, if changes in education, prioritization and working process are designed, open workflows might be enabled. A piece that could improve this, by saving resources for specific researchers, is a collaborative community around studies and platforms. The universities themselves could make this easier by explcitly stating that they want contributions by others, by documenting at least a more used project and see what happens by investing some time into that. Also, encouraging others to change their process for this, or communicating the difficulties of this to funders might help in getting resources enough to also include a robust methodology and increase reproducibility.

Open data is relatively easy to apply for the studies studied here since most of them have been applied to data materials that already are public domain for natural reasons, while it is much more difficult for newer data. To change this, copyright issues need to be handled differently than today for researchers, if society want research also on newer content. Apart from this, the researcher need time and resources to spend time making data open, but since this takes less time, it is often done.

Open access is also common for the projects. This is mostly due to the aim of many projects to digitise historical material and data from a political and research perspective. When doing so, making this open to access for the public is the standard, like for cultural heritage traditionally as well. Also, researchers don't need to put in much more effort to make it open access rather than not, as soon as the discussion of whether or not to publish open access is done, which it often is at the research project's outset, as soon as it is financed.

Overall, the factors that decide the degree of openness in the projects studied in this report include:

1. Funding agent. Political ones have started to work for more open science, private ones not always incentivizing being open, varied. 2. The purpose of the study, e.g digitising cultural heritage is surrounded by the premise of being open, while others are not. 3. The time it takes to do the extra work of

openness for researchers in general for different principles. Open workflows a lot of extra work especially for technically complex studies, open access not much and data not as much for some studies, open source can be difficult depending on the programs used. 4. The data used, e.g copyright issue in the humanities for newer data, therefore often older data used so it is open but some studies down-prioritized that maybe shouldn't be, and studies using newer data are not possible to do following open science priniciples. 5. The culture at the university and track record of openness previously, e.g are there a centre for digital humanities at the university, and are they following all open science principles as their ideal or not. 6. The interests of researchers. Often centres for digital humanities have been created bottom-up historically, and only now are politicians and decision-makers catching up.

There are obviously more potential in the type of studies and answers that could be given by more sophisticated technoglogy of the future for digital humanities and open science within the humanities. There are such questions as the first date of each Shakespeare play in theatres, distinguishing/categorizing/clustering genres in more detailed ways, answers about what good art/literature actually is and how to objectively quantify this, at least to some degree. These are studies I have not found extensive answers to at least, but rather subproblems and estimations are necessary for computational reasons. Also, more advanced sound to text technology could enable media studies on the medias actually most used today, which are often video and or sound media.

Infrastructures and institutions for open science in the humanities are just starting to be developed, including technical solutions in poorer countries, prioritization of problems within a country, institutional belief, competencies of methodology and or technical knowledge lacking, little awareness of open science etc. Also, it may be that the interest in digital humanities and open science in the humanities may not be larger due to relatively little technical and mathematical parts of the study traditionally in the humanities, and reversly little humanities studies in technical educations.

Improvements for another study like this would include a more systematic approach to literary search. I would like to define which databases I will use, even though this is difficult before knowing the subject, at least be able to specify afterwards which databases/journals etc were used, what exact keywords, which articles were read only in abstracts, which were left out and why, to create reproducibility, replicability and increase credibility in the results such that it is possible to know what raw data the conclusions are based on, which in reality is critical to evaluate the quality of the conclusions. With this project I have contributed to my own knowledge as an overview to the field, and this knowledge makes the more thorough study and systematic literature approach I propose above possible for myself, which I think it wasn't previously.

Further studies that could be made include a more thorough and overarch-

ing investigation of the studies that have been made with the help of specific platforms, to evaluate their use today and what have made some platforms used more than others, and in that sense better. Another study could ask researchers why they have applied a certain degree of open science principles in their research, and how it could be possible to increase the degree in different ways for different researchers and projects. Lastly, a more thorough user test study of multiple platforms that aim to find the best platform designs for the public and a non-expert user would be interesting to see, especially for more complex platforms such as detailed maps with crossreferences to other open data sources.

6 List of References

Aroseniusarkivet (2022):

https://aroseniusarkivet.dh.gu.se

https://aroseniusarkivet.dh.gu.se/forskning/

https://aroseniusarkivet.dh.gu.se/projekt/deltagare/

https://aroseniusarkivet.dh.gu.se/projekt/projektbeskrivning-2/

https://github.com/CDH-DevTeam/arosenius-api

Malm. M & Westin. J. (2020). Slutrapport: Hur frammanas konstnären ur arkiven? Exemplet Arosenius Digitalisering och samordning av samlingar för nya former av tillgängliggörande och forskningsfrågor. Göteborgs Universitet. https://aroseniusarkivet.dh.gu.se/forskning/report.pdf

Nordisk Familjebok:

https://nordiskfamiljebok.dh.gu.se/

https://github.com/CDH-DevTeam/norfam-frontend

https://github.com/CDH-DevTeam/norfam-backend

Regnum Francorum Online:

http://www.francia.ahlfeldt.se/index.php

https://github.com/johaahlf/rfo

GU. (2022). Johan Åhlfeldt: Historiska och arkeologiska databaser från antiken och tidig medeltid. https://www.gu.se/digital-humaniora/aktuellt/filmade-seminarier

Drama critiques:

Drama Critiques. (2022). Drama Critiques. https://dramacritiques.com/en/home/

Maignant, M., Pellé, D., Brison, G., & Poibeau, T. (2022). Drama Critiques' Database. Journal of Open Humanities Data, 8, 21. DOI: http://doi.org/10.5334/johd.81

https://github.com/MyleneM

doi: 10.5281/zenodo.6799656

https://www.theatrerecord.com/

Two Ibsen Studies:

Mohnike, T. (2020). Jens-Morten Hanssen: Ibsen on the German Stage 1876–1918. Ibsen Studies, 20(1). pp 94-100.

Holledge, J. Bollen, J. Helland, F. & Tompkins, J. (2016). A Global Doll's House. Palgrave Macmillan London. doi: https://doi.org/10.1057/978-1-137-43899-7.

https://www.hf.uio.no/is/english/services/virtual-ibsen-centre/ibsenstage/

Other studies:

Bardiot, C. (2017). Arts de la scène et culture analytics. Revue d'historiographie du théâtre: Études théâtrales et humanités numériques, 4, 11–20.

Hassani, H. Turajli, E. & Taljanovi, K. (2019). Digital Humanities Readiness Assessment Framework: DHuRAF. arXiv. doi: https://doi.org/10.48550/arXiv.1902.06532

Kuys, G., & Scherp, A. (2022). Representing Persons and Objects in Complex Historical Events using the Event Model F. Journal of Open Humanities Data, 8 (22). DOI: http://doi.org/10.5334/johd.84