Jupyter Notebooks for Generous Archive Interfaces

Mari Wigham
Netherlands Institute for Sound
and Vision
1217 WE Hilversum
The Netherlands
mwigham@beeldengeluid.nl

Liliana Melgar
Faculty of Humanities
University of Amsterdam
1012 XT Amsterdam
The Netherlands
melgar@uva.nl

Roeland Ordelman
Human Media Interaction
University of Twente
7500 AE Enschede
The Netherlands
roeland.ordelman@utwente.nl

***Abstract*— To help scholars to extract meaning, knowledge and value from large volumes of archival content, such as the Dutch Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH), we need to provide more 'generous' access to the data than can be provided with generalised search and visualisation tools alone. Our approach is to use Jupyter Notebooks in combination with the existing archive APIs (Application Programming Interface). This gives access to both the archive metadata and a wide range of analysis and visualisation techniques. We have created notebooks and modules of supporting functions that enable the overview, investigation and analysis of the archive. We demonstrate the value of our approach in preliminary tests of its use in scholarly research, and give our observations of the potential value for archivists. Finally, we show that good archive knowledge is essential to create correct and meaningful visualisations and statistics.**

I. Introduction

Digitally available archive collections are a gold mine for a wide variety of users, in particular for scholars, whose disciplines are being transformed by such availability[1]. However, after the initial efforts of massive digitisation of archival heritage, access to these collections relied on offering search and retrieval support [2]. Despite the call for building "exploratory" interfaces [3], the challenge still remains to provide ways to extract meaning, knowledge and value from digital archive collections, in a way that corresponds with scholarly practices, while taking privacy and copyright issues into account [4]. The authors of [2] make the case for 'generous interfaces', and describe the experience of current search portals with a striking analogy:

"Imagine yourself outside an art gallery in a far-off city, with a collection you don't know well. You enter the building to find a small, drab lobby with an attendant at a desk. The attendant asks you to enter your query on a small slip of paper. Not knowing the collection, and not seeking anything in particular, you write down something arbitrary, and pass it over. The attendant disappears for a moment before returning with a line of artworks sitting on trolleys. These are paraded, ten at a time, through the lobby. You can submit another query at any time, calling forth more trolleys, but there seems to be no way to explore the gallery beyond this small lobby"

The "generous interfaces" proposed by Whitelaw go beyond the "query-response paradigm" [5] emphasizing their role in "revealing the scale and complexity of digital collections" [2]. They do this by supporting exploratory search via enhanced browsing features and moves. However, as Whitelaw indicates, the design and development of these generous interfaces is challenged by the level of flexible experimentation that they require before they are actually implemented. This is so because, "in order to understand the features of a collection that might be represented, we must first represent the collection: the riches and voids in each collection are only evident through a process of exploratory visualisation." [2]

Transparency is very important when designing systems that are meant to support scholars in their research process [6], since scholars need to get close to the sources and understand provenance information. Building generous interfaces in the humanities should also support meaningful browsing that respects the principles of scholarship. In that sense, interfaces should find a way to "explain the physical aspects and intellectual structure of the collection that may not be apparent and to provide enough contextual information for the user to understand the historical circumstances and organizational processes of the object's creation." [7].

In this paper, we take the "generous interface" idea a step further, and propose that a generous interface should supply dynamically specified (aggregations of) metadata, allowing for the exploration, investigation and analysis of archival collections. In addition, we show how this approach can benefit archivists and support them in assisting scholars with their use of the archive. We tackle the challenges of generous interface design by including archivists in an iterative design process.

We distinguish three important tasks that are supported by the availability of dynamically specified (aggregated) metadata. ***Collection overview*** is the exploration not of the individual items in an archive, but of the archive as a whole. ***Investigation*** is the deeper exploration of a selection of items out of the archive (e.g., a sample, or a "corpus"). ***Analysis*** facilitates the task of scrutinising these items more deeply which, in combination with scholarly annotations or enrichments, helps to discern patterns that aid in answering a particular (research) question.

To explain these tasks in terms of the above analogy,

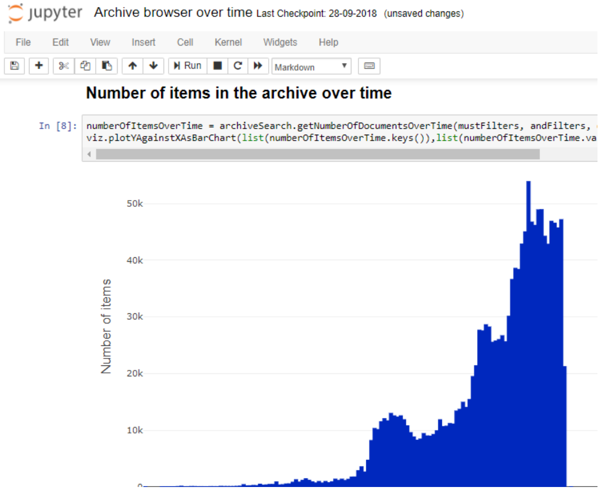
[1] As discussed in a great number of publications about the "digital humanities," for example [1]

think of a scholar visiting an archive that they have not seen before. They can start their exploration with general information about the archive, based on aggregations of the metadata over the whole collection - how many paintings does the archive contain, from which centuries do they come, for how many paintings is provenance available? This is the ***collection overview*** task. They could then wander around for inspiration and get to know the archive piece by piece - this is the ‘exploration’ or ’exploratory browsing’ task usually described in the concept of generous interfaces. A scholar will have questions about a specific selection of items, such as ‘How many oil paintings from the 18th century have information about the subject of the painting?’. They will need to consult an archivist, who may in turn need to consult their own metadata sources. This is the ***investigation*** task. Finally, a scholar will (re)formulate their research question and scrutinise a selection of items more deeply in order to discern patterns that answer their research question, for example ‘Was there a shift in the types of subjects painted during the course of the 18th century?’ - the ***analysis*** task.

It is important to note that a ‘generous interface’ is more than a ‘user interface’ - it must connect the user not to the software, but to the data. To provide good access to data requires infrastructure for storing and organising the data, tools to process the data, and finally the user interface to present the information to the user. Providing scholarly access to large archive collections is the focus of the Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH) project. Within this project, the Media Suite research environment has been developed (http://mediasuite.clariah.nl/). The Media Suite brings together diverse collections from various institutions (archives, libraries etc.). It provides the necessary infrastructure, tools and user interface for a user to access the content and metadata in the collections. The Media Suite provides extensive support for search, but search is not the only tool in the suite. The Explore tool makes it possible to browse through and jump from one individual item to related items. The Inspect tool makes it possible to examine the completeness of metadata (over time) supporting the “investigation task”. The Compare tool allows the scholar to compare search results from different collections. Finally, the Annotation tool and a personal Work Space facilitate the “analysis task”.

By offering exploration, inspection, comparison, and annotation in addition to search, the Media Suite takes important steps towards offering a more generous interface. However, the Media Suite must serve a wide range of scholars, facing a common problem for infrastructure projects, defined as the “generalization paradox” [8]. If we pose the question, ‘What general information do you want to know about the archive?’ to ten scholars, we will in all likelihood get ten different answers. Support further along in the research process, for investigation or analysis, must clearly be customised to the individual scholar, especially as different scholars will want to use different analysis methods. This need for custom support calls for a different approach, complementary to the functionality of the Media Suite.

Fig. 1. A fragment from a Jupyter Notebook



Number of items in the archive over time

```
In [8]: numberOfItemsOverTime = archiveSearch.getNumberOfDocumentsOverTime(mustFilters, andFilters, 
        viz.plotYAgainstXAsBarChart(list(numberOfItemsOverTime.keys()),list(numberOfItemsOverTime.va
```

The Media Suite has various Application Programming Interfaces (APIs) that provide direct access to collection and item metadata, making it possible in theory for scholars to perform their own analysis. However, as discussed in [9], the existence of an API is by no means a guarantee that scholars will be either willing or able to use them. Additionally, querying the API correctly requires knowledge of the archive. As we want to provide the researcher with a means of getting to know the archive, requiring a pre-existing knowledge of the archive to obtain the metadata creates a chicken-and-egg situation.

The problem is to sufficiently simplify access to the metadata to provide easy-to-use support for *collection overview*, *investigation* and *analysis*, while also offering the necessary flexibility to adapt to the individual scholar’s needs.

We propose that the solution is to present the metadata, via the APIs, in Jupyter Notebooks and to offer these notebooks as a service that complements and extends the Media Suite. Jupyter Notebooks (Figure 1) are web applications to create and share documents that contain live code, equations, visualizations and narrative text. [10] They support users with a wide range of skills, as they can be fully pre-programmed, supplied together with useful functions as building blocks for the user, or let the user program their own code. This flexibility is key to solving the generalisation paradox. The ability of Jupyter Notebooks to add descriptive text to data and visualisations, and the ability to see exactly the lines of code that have produced the data, aid transparency and supply the necessary contextual information for meaningful browsing.

In this paper, we will discuss our development of Jupyter Notebooks in combination with the Media Suite APIs to enable both scholars and archivists to explore, investigate and analyse the metadata in the Netherlands Institute for Sound and Vision (NISV) archive, part of CLARIAH. The main focus of the paper will be on supporting scholars; the task of supporting the archivist will be explored in depth at a later date. However, during the development of the notebooks we have already found evidence of their usefulness for archivists, and we will report these observations in this paper. We will explain how the notebooks support *collection overview*, *investigation* and *analysis*, give examples of the visualisations and their use, and discuss the experiences of using the notebooks for scholarly research projects at the CLARIAH summer school. We will discuss the issues encountered in the development and use of the notebooks. Finally, we will offer our conclusions and describe our plans for future work.

II. Background

A. The Netherlands Institute for Sound and Vision's archive

The Netherlands Institute for Sound and Vision (NISV) is the Dutch national institute charged with collecting, preserving and providing access to the audiovisual heritage of The Netherlands. The institute was formed in 1997 in a fusion of several existing archives. The collection contains more than a million hours of material, a large proportion of which has been digitised in a central repository. In addition to audiovisual content, photos and even descriptions of physical objects are also included. Much of the material is subject to copyright restrictions. In the Netwerk voor Digitaal Erfgoed (NDE - Network for Digital Heritage) project, the NISV is one of the five pillars, taking the lead in ensuring digital access for the audiovisual community. The NISV plays a key role in the CLARIAH project.

B. The CLARIAH infrastructure

The CLARIAH project aims to develop a distributed infrastructure for the humanities and social sciences. The NISV is a formal data centre within the project, as well as being responsible for the development of the Media Suite application for accessing the collections brought together in CLARIAH. Different requirement elicitation methods were used with one group of scholars of the CLARIAH infrastructure project, and steps for user-centered software design were followed [11], which led to the design and evaluation of the Media Suite.

Metadata is harvested from the CLARIAH collections in a variety of ways, from live metadata harvests to imported dumps, and stored in one central search index. Access to this data is provided by a number of APIs, including the Collection API that provides high-level information about the collection, and the Search API that makes it possible to search for specific items. As some collections are subject to copyright, and others (such as Oral History collections of interviews) have privacy issues, users must be authenticated before they can have access. Content is processed by automatic metadata extraction (AME) technology, such as a speech recogniser to produce speech transcripts, improving searchability. The infrastructure is designed to ensure availability, usability and integrity of the metadata, and to allow simultaneous access to a large group of users while maintaining good performance.

The Media Suite user interface uses the APIs and underlying infrastructure to provide authorised scholars with access to both the metadata and the audiovisual content itself. While the Media Suite user interface serves the needs of a wide range of scholars, and offers extensive support for working with the audiovisual content, the proposed Jupyter Notebooks offer complementary functionality by serving the specific needs of individual scholars, and offering extensive support for working with the metadata. Both are built on the same underlying CLARIAH infrastructure

III. Developing a generous interface

A. The development process

We initially created three early prototype notebooks, each containing basic functions for accessing metadata via the API and creating visualisations. The first allowed users to explore information about available metadata fields (including completeness), the second showed the availability of speech recognition transcripts, and the third extracted text from items in the user's Media Suite workspace and analysed the word frequencies, plotting these in word clouds. It was up to the user to select which metadata fields were of interest, and to determine how to use the notebooks in their research. These notebooks were tested with scholars at the CLARIAH summer school (see Evaluation section for a detailed discussion of the tests). These tests showed the need for us to take a stronger lead in showing the scholars how the notebooks could help them in their research tasks, and also for more flexibility for custom analysis. This prompted a shift in our approach. On the one hand, we developed prepared notebooks, focused on specific research tasks, that contained a standard set of visualisations for the most relevant archive metadata. On the other hand, we made it possible for the user to make their own custom selection within the archive for *investigation*, and extended the basic metadata functions into a set of 'building blocks', in order to make the notebooks more easily customisable.

We created an 'archive metadata search' Python module that translates the API calls into high-level, easy to understand functions in terms of generic archive concepts, concealing the complexity of the API from the user. At the same time, the API is still available for more experienced users to create custom queries. The functions retrieve general information about how many archive items are available (over time) and which metadata fields are present. For each metadata field, there are functions to see how complete the field is, retrieve its possible values, how frequent the values are, and how they vary over time. For numerical fields, the maximum, minimum and average can be computed (also over time); for text fields the length of the text can be calculated, and words can be retrieved for further processing. For all functions, an optional query can be specified, to calculate the statistics over a specific portion of the archive - for example to find the average length of a news programme in a given year.

We also created a 'visualisation' Python module, built on the freely available Plotly (https://plot.ly/), MatPlotLib (https://matplotlib.org/), and WordCloud (https://github.com/amueller/word_cloud) libraries, to provide easy-to-use functions for producing relevant archive visualisations, e.g. plot the most frequently used values of a metadata field, or plot the change in the number of items over time.

To create a meaningful statistic and/or visualisation requires three steps. First, we must select the correct metadata fields and statistics. Second, we must implement these in the notebook. Third, we must provide contextual information to aid interpretation. The first step is far from trivial, and may involve 'archive archaeology', as changes during the lifetime of the archive can make it difficult to find out exactly what metadata fields mean. This step requires good knowledge of the archive, and 'archival thinking'. The second step is a programming task, requiring good data science knowledge, and assisted greatly by our 'building blocks'. The third step requires both 'computational thinking' and 'archival thinking', as both affect how a visualisation is interpreted. As an example, we can take the apparently simple question of how many hours of material are in the NISV archive. Archival knowledge is necessary to select which of the available fields measuring duration is appropriate, and to know that the duration information is missing for many items, so we need to scale up the numbers - excluding items for which a duration is meaningless, such as photographs - to reach an estimate. The implementation of this is simple, but on presenting this information we again need archival knowledge, to indicate that this number includes 'clean feeds' - copies of programmes intended for reuse - and as such is not the number of hours of ***unique*** programmes. In Jupyter Notebooks we can easily add textual descriptions to explain this context.

This process relies heavily on archival knowledge. It is not possible to simply ask an archivist to 'tell everything about the archive'. There is too much knowledge, much of which is implicit. Furthermore, the 'archivist' is not just one person. For example, a metadata expert will know which metadata fields to choose, while an ingest expert will know that clean feeds are also added to the archive. This means that the process of developing visualisations is interactive and iterative. Visualisations are produced, then discussed with the experts to identify unexpected results and elicit new information to either correct the code or provide context to understand the anomaly.

To aid analysis, we installed the Natural Language Toolkit (NLTK) (https://www.nltk.org/) in the notebooks. A wide range of other analysis and visualisation functions are available, as Jupyter Notebooks support over 40 programming languages, and do not require the functions to have been created specifically for Jupyter Notebooks.

We offer the Jupyter Notebooks on a hub hosted at NISV, accessible to authorised users. This means that users do not have to install the Jupyter software themselves, but can use the notebooks in their web browser. The statistics and visualisations created in the notebook can be downloaded from the notebook.

The application of the developed notebooks and modules to the ***collection overview***, ***investigation*** and ***analysis*** tasks will be explained further in the next section.

B. Collection Overview

For a user (scholar or archivist) who wants a *collection overview*, we have developed two 'Archive browser' notebooks. One of these provides an overview of the total archive metadata, the other shows the metadata in the archive over time. For each notebook, we selected a number of important metadata fields, and prepared relevant statistics and visualisations for those fields. For example, information in the total archive overview includes the number of hours in the archive, the split between TV, radio and other types, the proportion of the archive that is digitised, the most frequently occurring genres, and much more. Information over time includes how the number of TV and radio items have changed over time, what proportion of material per year is digitised, and how the average length of a programme description has changed. In addition, the completeness of the selected metadata fields (how often the information in those fields has been filled in) is shown, also how this completeness has changed over time.

Knowledge of the archive is built into the notebook. For example, genre is determined based on the series genre, as we know that this is filled in for 99.5% of items, rather than for programme genre, which is filled in for only 18%. Aggregation per year is done using the 'sort date', the most meaningful date for each item. The visualisation of how much of the archive is digital – a seemingly simple question – is given the appropriate context to explain the underlying complexities.

As the notebook has already been filled with relevant statistics and visualisations, all the user has to do is to run it. A user can also, with relative ease, enter different metadata fields to be used in the plots. For example, in the completeness function they could replace the 'genre' metadata field with the 'broadcaster' metadata field, to see how complete the broadcaster information is.

The *collection overview* information gives a scholar a sense of what is available in the archive as a whole. It can also prompt specific questions for further investigation. For example, in our visualisation of programme duration, we found some items with negative durations, and some with extremely long durations (380 hours).

Figures 2-5 show some example visualisations from the *collection overview*. Figure 2 shows the most frequently occurring genres - it is clear that news and music performances feature greatly in the NISV archive. Figure 3 shows the proportion of the archive that is already digital, while Figure 4 shows the proportion of the material that is digital per year. Finally, Figure 5 shows the availability of speech recognition transcripts per year.

C. Investigating the archive

A scholar who wants to investigate the potential and suitability of the archive for their research question needs to know more than the general statistics. For example, if a scholar is analysing the speech transcripts of TV news programmes, then they are not interested in the availability of speech transcripts over the whole archive, but specifically for

Fig. 2. Most frequently occurring genres

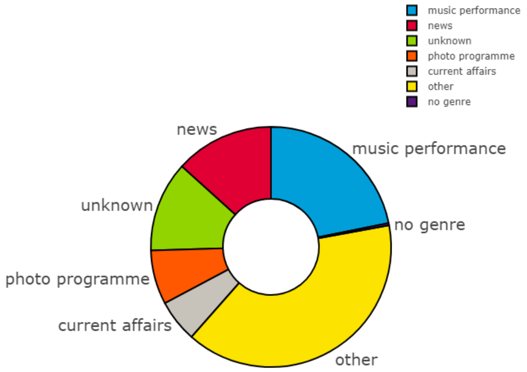


Fig. 3. Proportion of archive that is digital

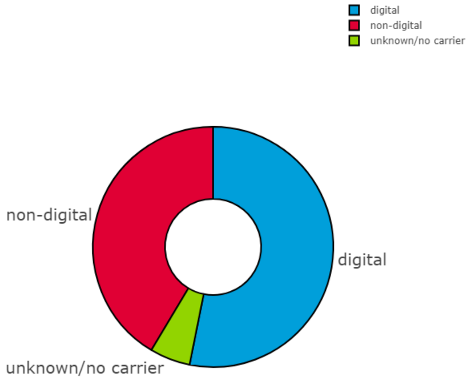


Fig. 4. Proportion of archive that is digital over time

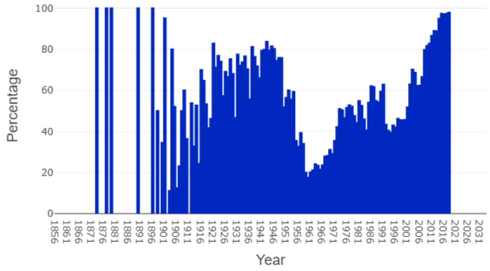


Fig. 5. Availability of speech recognition (ASR) over time

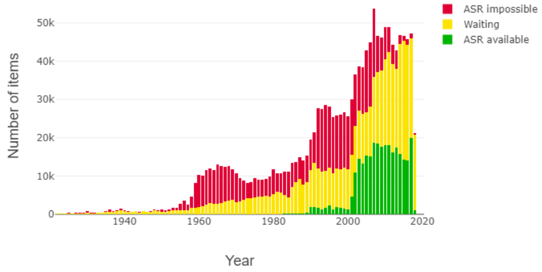


Fig. 6. Availability of speech recognition (ASR) for TV news items

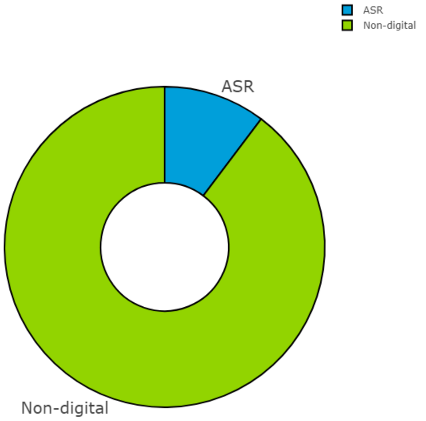
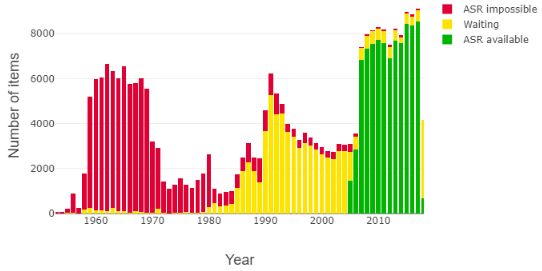


Fig. 7. Availability of speech recognition (ASR) transcripts for TV news items over time



TV news. While the Media Suite collection inspector offers information about metadata completeness for a collection in general, it cannot answer this specific question.

Investigation can be carried out using the archive browser notebooks developed for the *Collection Overview* task. To answer their question, the user adds in a search query that specifies the required subsection of the archive, for example all television programmes, all drama productions from the Eighties, or all series of the current affairs programme 'EenVandaag'. The detailed query syntax is hidden away and exposed as simple building blocks, however, some effort is still required for the user to familiarise themself with how the blocks are used to create a query. Once a query has been specified, all the available statistics and visualisations can be produced for the results of that query, simply by running the notebook. No changes to the functions themselves are necessary.

Example visualisations are Figure 6 and Figure 7. As Figure 6 shows, the scholar wanting to analyse TV news programmes seems to have a problem. Only a relatively small proportion of the programmes have speech transcripts, and those that do not, must first be digitised before they can be put through the speech recogniser. The scholar then looks at the availability over time (Figure 7, and is relieved – they are interested in the changes over the last decade, and the availability of speech recognition in that time period is excellent.

In the *collection overview* task, we also noted that the statistics and visualisations can prompt specific questions about the archive. For example, we found that some items were very long – in fact the longest item was 380 hours. Via the notebook we could discover that these items were compilations, for example all music by Elvis Presley, or an entire season of a popular soap. We also found a number of items with a negative duration. The metadata of these items showed a simple explanation – the items were broadcast late at night and finished early in the morning, so a start time of e.g. 11:30pm and a finish time of 1:30am equaled a duration of -22 hours.

The user must have some knowledge of the archive in order to create their query. Some of this knowledge is supplied by the notebook itself, such as the metadata fields available and their possible values. However, some knowledge can only be obtained by experience with the archive or contact with experts, for example to know that the 'sort date' for items is the most 'meaningful' date for that type of item, which may be the date of broadcast for a television programme, but the date of creation for a photo.

D. Analysing the archive

Once a scholar has decided the archive is suitable, and refined their research question (as part of the research cycle explained in [12]), they will select interesting items for their research. Traditionally, they would then perform "close reading" [13] - reading, watching and/or listening to the items to *analyse* them, and then aggregate their knowledge of individual items in search of a meaningful pattern. Increasingly, data science is offering an additional option for analysis by processing the metadata (and sometimes content) of the items in bulk in order to detect interesting patterns.

Even though close reading and manual scrutiny or annotation remain essential for interpretation, quite a lot of analysis, such as 'distant reading', can already be performed using the *investigation* functions previously described. For example, the scholar interested in genre can plot the top genres in the 70's and see how these changed in the course of the decade. They can also compare different selections of items, for example, they could plot the change in duration of news programmes over time and compare it with the change in duration of religious programmes over time (Figure 8 and Figure 9).

The duration plot triggers a question – why does it seem that news programmes since 2017 have become very short? Changing the query to plot the duration distribution during 2017 reveals the answer – many programmes have been given the duration '0'. The scholar can adjust their original query to filter out items with a zero duration and get a better picture of the trend over time. The visualisation helps the scholar to spot an inconsistency, and the ability to investigate the underlying data helps them find the reason. This shows how "generous interfaces" can support "data criticism" [6], and help to integrate it into the search and browsing practices of scholars. This helps to ensure that the data is properly used and interpreted, and that true trends can be distinguished from archive artefacts.

Fig. 8. Change in duration of TV news programmes over time

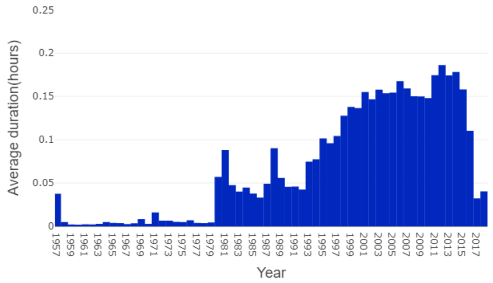
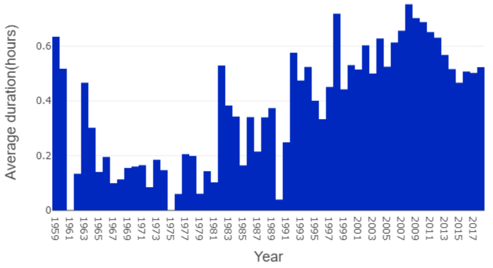


Fig. 9. Change in duration of TV religious programmes over time



In addition to analysis using the archive browser functions, the scholar can use the module functions to access the metadata, and process it with the NLTK toolkit. For example, a scholar wishing to analyse speech transcripts for frequently occurring words can retrieve the transcripts and use the NLTK to identify the most frequently occurring words or word pairs, and plot these in a word cloud. To illustrate this, Figure 10 and Figure 11 compare the most frequently used terms in news summaries in the period 1950-1959 with the period 2000-2010 (a noticeable difference is that the 50's version mentions many more different countries than the 00's).

An advanced user can program their own analysis functions in the notebook, using the Python programming language or another of the many languages supported by Jupyter Notebooks. Alternatively, if they wish to use existing software tools that are not available in the hosted notebooks, they can download the results and analyse them further in their own software, or to run their own Jupyter Notebooks server and install the analysis methods of their choice.

Fig. 10. Word cloud for news summaries 1950-1959

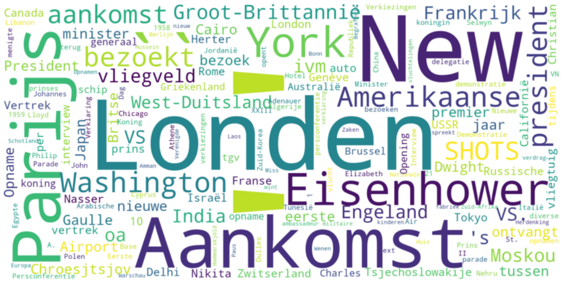


Fig. 11. Word cloud for news summaries 2000-2010

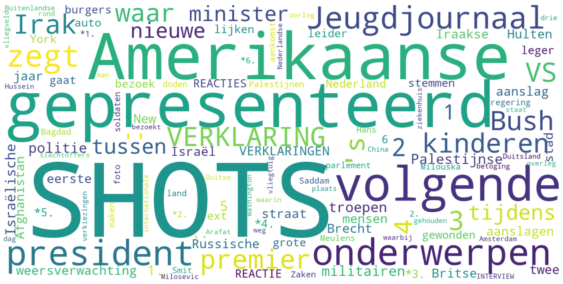
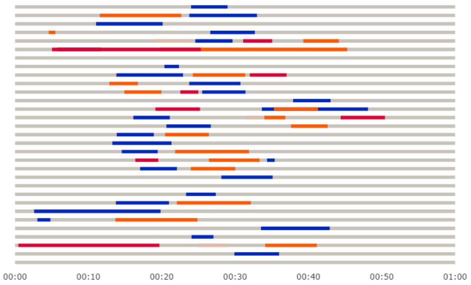


Fig. 12. Location of YouTube clips (red, orange and blue) within the timeline of the original ‘Nieuwsuur’ episodes (grey) for April 2018. Each line is an individual episode



IV. Evaluation

Informal user testing of the early prototype notebooks was carried out with scholars, to see if these notebooks could support the scholarly research process. The Media Suite (version 3) and the notebooks were tested during the Clariah Summer School in 2018, where circa forty scholars worked on eight group research projects using the Media Suite, assisted by data scientists from NISV, who also had knowledge of the NISV archive [2]. The scholars received explanation of the notebooks in a workshop, and had the opportunity to try out some examples during that session.

During the week, several of the research projects used the notebooks to access and analyse the metadata. The data scientists added additional functionality to the notebooks on the fly to support the specific research questions of the groups. For example, one project group worked on comparing YouTube clips of popular Dutch current affairs programmes to the original broadcast episodes. This required matching the clips to the programmes (based on the clip title, description and date), and then aligning the speech recognition transcripts from the archive with the YouTube transcripts to locate the clips within the programme. The speech recognition notebook helped the scholars to select current affairs programmes that had sufficient speech recognition coverage in the time frame of interest. From that point onwards, bespoke functions were written in the Jupyter Notebook by the data scientist, to perform the matching and to visualise the results (see Figure 12). The results of both the matching and visualisation functions were shared with the scholars early on, giving them insights that helped them refine their research question and also better specify their requirements to the data scientist, enabling the sort of co-development recommended in [14]. This step was repeated in a number of iterations.

The scholars involved in the summer school responded enthusiastically to both the possibilities offered to them by the Jupyter Notebooks in their research, and the involvement of data scientists in their group (see above-mentioned blog). They considered the notebooks innovative, but also challenging, and abstract. They relied a great deal on the data scientists for knowledge of the archive metadata and to use and adapt the notebooks (it should be noted here that the ready availability of the data scientists may have made the scholars less likely to use the notebooks on their own).

This sort of data science approach to analysing the archive would have been impossible using the Media Suite user interface alone. The amount of analysis completed within the duration of the summer school would not have been possible without the metadata search functions already being available in the notebook. This shows how having these ‘building blocks’ for archive metadata search speeds up the research process, even in this early prototype where the blocks were still quite basic. The matching and alignment functions developed during the summer school have since been added to the standard archive search module, ready for reuse. This shows how developments that are initially created as custom support for a specific research group, can become new building blocks that benefit other research groups.

The problems with the abstractness and complexity of the notebooks prompted us to take a stronger lead in showing scholars how they can use the notebooks in their research tasks. As a result, we developed the archive browser notebooks described in section III-B. The need for custom analysis prompted us to extend the basic functions into modules for archive metadata search and visualisation, as described in section III-A. The archive browser notebooks and modules have been functionally tested, for both *collection overview* and *investigation* tasks. These new notebooks will be tested with scholars in the future to see if they adequately address the issues experienced with the prototypes.

No versions of the notebooks have yet been tested with archivists, However, during the development of the archive browser notebooks, the information conveyed by the visualisations triggered discussions within NISV that led to concrete actions. Figure 5 prompted an initiative to make more speech recognition possible by digitising older TV material, while the discovery of negative durations mentioned in section III-B caused the metadata to be reviewed and cleaned up. These incidents demonstrate the potential role for the notebooks in giving archivists more insight into their archive, and assisting in processes such as quality control and preservation projects.

V. Discussion

Our experience in developing the Jupyter notebooks has demonstrated both their benefits, and highlighted some issues

[2] https://clariah.github.io/mediasuite-blog/blog/2018/10/01/Clariah-Media-Studies-Summer-School-report

in their use.

We selected Jupyter Notebooks as they support users with a range of skills. The prepared archive browser notebooks can be run with a simple click. However, the Jupyter notebook still makes all the code visible, which could be off-putting to a novice. The user must also be aware of the basics of how the notebook works, for example that different parts of notebooks can be run independently - this could lead to inconsistent data if not used correctly. Changing to a different metadata field requires only a very simple edit to the code - but if the user does this incorrectly, they could be faced with an error message that they may not understand. A possible solution could be to create a simple interface, using for example Dash (https://plot.ly/products/dash/) or PyWidgets (https://pypi.org/project/ipywidgets/), so that the user no longer sees the code, and can use controls to make changes - e.g. change metadata fields using dropdown lists.

The further along the research process we go, the more programming ability that is required. *Collection overview* can be done simply by clicking on 'run', *investigation* requires the ability to write queries and possibly change parameters in functions, and *analysis* (beyond comparing the archive browser results) requires the ability to use third-party functions or program your own code. The same goes for archive knowledge; for *collection overview* this is built in, but as a scholar starts to create their own queries and visualisations, they need knowledge of the archive in order to avoid misleading graphs or incorrect interpretations. A user also needs to have some general understanding of data science in order to interpret statistics and visualisations correctly. A perfect example is Figure 3, that shows the proportion of items digitised per year. It may appear that 1873 is an excellent year, with 100% of material digitised, whereas the reality is that there is only one archive item from 1873. Such pitfalls are common to statistics and visualisations, and are unrelated to the particular archive. In all three cases, one possible solution is to allow the scholar to work with data scientists and/or archivists, as demonstrated in the CLARIAH summer school. Another solution is to offer the scholar training in the use of the notebooks, or to include documentation.

As the summer school set-up already included data scientists, we have not yet been able to test how scholars manage to use the notebooks on their own. Such tests would show to what extent data science, programming and archive knowledge are already sufficiently present, and to what extent it needs to be supplied. While there has been much discussion about whether scholars should become data scientists (see for example [15]), which has led to active promotion of such skills in practice, the discussion over the need for archival knowledge has not yet found its way into practical applications that fully support data criticism [16].

Making the notebooks available also presents a technical challenge. A Jupyter notebook can run a wide range of code - that is what makes it so versatile. This opens the door to the possibility of a user entering malicious code, so measures must be taken to protect the system in such an eventuality. The hub must also be able to handle many users simultaneously, and this could become a problem if one or more users are running a computationally intensive analysis. At present, a user wanting new software packages cannot add this to the hub, they must either request this, or set up their own notebook server and run a copy of the notebook there.

Another potential security risk is the ability to download data from the Jupyter notebooks. The notebooks offer access to the metadata, not the audiovisual content. However, this metadata may still fall under copyright law, or involve privacy issues. As data and visualisations can be downloaded from the notebook, it is important to consider how we will handle copyright and privacy issues for downloaded data. This is a broader issue than for the notebooks alone.

VI. Conclusions

In this paper, we described how we combined Jupyter Notebooks with the Media Suite API, built on the CLARIAH infrastructure, to provide a more 'generous' interface to the NISV archive, using statistics and visualisations of (aggregated) metadata to assist scholars in extracting meaning, knowledge and value from archives.

We have shown that the combination of API and notebooks can support various tasks in scholarly research. The prepared archive browser notebooks give scholars an *overview* of the archive, and allow them to *investigate* a specific selection of items. The information offered in these notebooks can also be used to *analyse* and compare different selections of items. The archive metadata search module gives the scholar building blocks with which they can access the metadata for their own custom analysis, using the wide range of techniques available. The insights given by the visualisations and the ability to further explore the data behind those visualisations also help to encourage data criticism.

Scholars reacted enthusiastically to the potential of early prototypes, and these prototypes were used successfully in carrying out actual humanities research projects during the CLARIAH summer school. Issues relating to the complexity of notebooks and the need for custom analysis were addressed in the development of the current version of the notebooks. During the development process, example visualisations were discussed with archivists. The insight created by these visualisations triggered improvement actions within the archive, indicating that there is a great potential benefit of this approach for archivists.

The process of developing the notebooks demonstrated the need for knowledge in three areas - data science, programming, and the specific archive being accessed - in order for statistics and visualisations to be correctly created and interpreted. This need was confirmed by the results of testing early versions of the notebooks with scholars. This requires either that the scholar develops these skills themselves, or that they are supported in their work by knowledge from data scientists and archivists. To develop the notebooks, we worked together with archivists in an iterative process, thus solving the problems of designing generous interfaces described in the introduction. The same iterative process was used in the co-development of specific analysis functionality for the scholars in the summer school.

This approach can be applied to other archives with an API that can deliver (aggregated) metadata for the items retrieved by a search query. The module for communication with the API must be created specifically for each archive, but all other software used for statistics, visualisation and analysis in our notebooks is freely available on the internet.

In general, we conclude that our approach is a fruitful avenue for developing generous interfaces for archives.

VII. Future work

The next step is to test the improved notebooks and their supporting modules with scholars. With these insights we will adapt the notebooks to better support scholars in various stages of the research process[17]. We will also look at what form of additional data science/archival knowledge support is required. An interesting ongoing initiative in the context of the CLARIAH project is to create "metadata dictionaries" for all available collections, which add user-friendly labels and definitions. These dictionaries could assist notebook users.

We will extend the archive browser notebooks to cover all the collections contained in the Media Suite. The archive metadata search module is already capable of accessing and processing the metadata of these collections, however, as previously discussed, it will require archival knowledge to produce relevant, meaningful visualisations of this metadata. We will work with the experts of the participating institutes to incorporate their knowledge. The possibility of combining data from these multiple collections will also be explored.

On the technical side, we will look into the use of user interface packages so that users can change the metadata fields or the selected archive subset without being confronted with the code. An additional approach that we are currently exploring for users with lower "code literacy" [15] is to seamlessly link the Media Suite's graphic user interface to the notebooks, so that a user can observe the produced visualisation in the Media Suite. We will also examine possible solutions for the security issues raised by running a Jupyter hub.

VIII. Acknowledgements

The research for this paper was made possible by the CLARIAH-CORE project financed by NWO. We gratefully acknowledge the support of the entire CLARIAH WP5 team, the scholars attending the CLARIAH summer school, and the archive experts at NISV. In particular, we would like to thank Vincent Huis in 't Veld and Tim Manders for their archive knowledge, Marijn Koolen and Willem Melder for introducing the concept of Jupyter Notebooks and conducting initial experiments into their use with the archive, Kaspar Beelen and Carlos Martinez for sharing their expertise on Jupyter Notebooks, Jonathan Blok for setting up the Jupyter Hub, Jaap Blom for developing the Media Suite APIs, and Hugo Huurdeman for reviewing the paper.

References

[1] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp, *Digital Humanities*. Cambridge, MA: The MIT Press, Nov. 2012.

[2] M. Whitelaw, "Generous Interfaces for Digital Cultural Collections," *Digital Humanities Quarterly*, vol. 009, no. 1, 2015.

[3] G. Marchionini, "Exploratory search: from finding to understanding," *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.

[4] R. Ordelman, C. Martínez Ortíz, L. Melgar Estrada, M. Koolen, J. Blom, W. Melder, J. van Gorp, V. De Boer, T. Karavellas, L. Aroyo, T. Poell, N. Karrouche, E. Baaren, J. Wassenaar, O. Inel, and J. Noordegraaf, "Challenges in Enabling Mixed Media Scholarly Research with Multi-media Data in a Sustainable Infrastructure – DH2018," Mexico, 2018. [Online]. Available: https://dh2018.adho.org/en/challenges-in-enabling-mixed-media-scholarly-research-with-multi-media-data-in-a-sustainable-infrastructure/

[5] R. W. White and R. A. Roth, *Exploratory search: beyond the query-response paradigm*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. [San Rafael, Calif.]: Morgan and Claypool Publishers, 2009, no. 3. [Online]. Available: http://www.morganclaypool.com/doi/abs/10.2200/S00174ED1V01Y200901ICR003

[6] R. Hoekstra and M. Koolen, "Data Scopes: towards Transparent Data Research in Digital Humanities – DH2018," Mexico, 2018. [Online]. Available: https://dh2018.adho.org/en/data-scopes-towards-transparent-data-research-in-digital-humanities/

[7] CLIR, "The Archival Paradigm: The Genesis and Rationales of Archival Principles and Practices • CLIR," Tech. Rep. [Online]. Available: https://www.clir.org/pubs/reports/pub89/archival/

[8] J. van Zundert, "If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities," *Historical Social Research / Historische Sozialforschung*, vol. 37, no. 3 (141), pp. 165–186, 2012. [Online]. Available: http://www.jstor.org/stable/41636603

[9] J. Edmond and V. Garnett, "APIs and Researchers: The Emperor's New Clothes?" *International Journal of Digital Curation*, vol. 10, no. 1, pp. 287–297, 2015. [Online]. Available: https://doaj.org/article/c0c8444824994495adf38a23204cbda0

[10] F. Perez and B. E. Granger, "Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science," 2015. [Online]. Available: https://blog.jupyter.org/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58

[11] E. G. Toms, *Encyclopedia of Library and Information Sciences*. Taylor & Francis Group, 2009, ch. User-Centered Design of Information Systems.

[12] M. Bron, J. van Gorp, and M. de Rijke, "Media studies research in the data-driven age: How research questions evolve," *Journal of the Association for Information Science and Technology*, vol. 67, no. 7, pp. 1535–1554, 2015, jasist2015-bron-media.pdf. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.23458/abstract

[13] K. Schulz, "The mechanic muse - what is distant reading?" *The New York Times*, Jun. 2011. [Online]. Available: http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html

[14] F. de Jong, R. Ordelman, and S. Scagliola, *Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development*. Centre for Language Technology, Copenhagen, 11 2011, pp. –, eemcs-eprint-20868.

[15] V. Zundert, J. J, and R. Haentjens Dekker, "Code, scholarship, and criticism: When is code scholarship and when is it not?" *Digital Scholarship in the Humanities*, vol. 32, no. suppl_1, pp. i121–i133, 2017.

[16] F. W. Gibbs, "New Forms of History: Critiquing Data and Its Representations," *The American Historian*, Feb. 2016. [Online]. Available: http://tah.oah.org/february-2016/new-forms-of-history-critiquing-data-and-its-representations/

[17] L. Melgar, M. Koolen, H. Huurdeman, and J. Blom, "A process model of scholarly media annotation," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, ser. CHIIR '17. New York, NY, USA: ACM, 2017, pp. 305–308. [Online]. Available: http://doi.acm.org/10.1145/3020165.3022139