Media Suite: Unlocking Audiovisual Archives for Mixed Media Scholarly Research

Ordelman, R.; Melgar , L.; Van Gorp, J.; Noordegraaf, J.

Publication date
2019

Document Version
Final published version

Published in
Selected papers from the CLARIN Annual Conference 2018

Media Suite: Unlocking Audiovisual Archives for Mixed Media Scholarly Research

Roeland Ordelman
Netherlands Institute for Sound and Vision
University of Twente
The Netherlands
rordelman@beeldengeluid.nl

Liliana Melgar
Department of Media Studies
University of Amsterdam
The Netherlands
melgar@uva.nl

Jasmijn Van Gorp
Department of Media and Culture Studies
Utrecht University
The Netherlands
j.vangorp@uu.nl

Julia Noordegraaf
Department of Media Studies
University of Amsterdam
The Netherlands
J.J.noordegraaf@uva.nl

Abstract

This paper discusses the rationale behind and approach towards the development of a research environment –the *Media Suite*– in a sustainable, dynamic, multi-institutional infrastructure that supports mixed media scholarly research with large audiovisual data collections and available multimedia context collections, serving media scholars and digital humanists in general.

1 Introduction

In some domains of scholarly research, the focus is on the creation of new data collections. In other domains, for example, in Media Studies (e.g., film and television studies) research often focuses on data collections maintained at cultural heritage institutions, such as archives, libraries, and other knowledge institutions. However, especially when audiovisual media are concerned, access to, and use of these collections is often restricted due to intellectual property rights (IPR) or privacy issues (e.g., with respect to recorded interviews). Moreover, individual institutions often do not have the technical infrastructure in place to serve basic scholarly needs with respect to search, exploration and inspection of individual items (i.e., play-out or viewing). Therefore, scholars either fall back on collections that are openly available or spend considerable amounts of time in *on-site* visits to archives for consulting data collections (Bron et al., 2016). Data collections at these institutes can be regarded as "locked", or at least hard to use for scholarly research.

To unlock these "institutional" collections and let scholars take advantage of the sheer quantity and richness of these data sets, we are developing an infrastructure for *online* scholarly exploration of collections that are distributed across various institutional content owners. Specifically, we focus on *audiovisual* data collections and related *multimedia* sources, such as radio and television broadcasts, films, oral history interviews, but also (news)paper archives, film posters and eyewitness reports. An online application, named *Media Suite*[1], serves as the interface to this underlying infrastructure, where content and metadata can be explored, browsed, compared, and where personal virtual collections can be compiled and stored in a personal workspace. In this workspace, scholars have additional tools for working with these *mixed media* collections, such as tools for automatic annotation, visualisation, analysis, and sharing.

The ultimate goal of developing the *Media Suite* and its infrastructure, is to (i) enable distant reading (Moretti, 2013), that is, identifying patterns or new research questions in and across aggregated collections, (ii) facilitate close reading: the detailed examination of individual items (e.g., videos) in a

[1] http://mediasuite.clariah.nl

Roeland Ordelman, Liliana Melgar, Carlos Martinez-Ortiz, Julia Noordegraaf and Jaap Blom 2019. Media Suite: Unlocking Archives for Mixed Media Scholarly Research. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 133–143.

collection or specific sections of these items (e.g., video segments) during search and scholarly interpretation, and (iii) make sure that the "scholarly primitives" (Unsworth, 2000; Blanke and Hedges, 2013), basic activities such as "discovering", "annotating", "comparing" and "storing", common to research across humanities disciplines, are well supported.

In pursuit of these goals, we face challenges on various levels, broadly identified as appropriate access to *data* and *tools*. Appropriate data access entails the ability to view and browse individual media objects for close reading, thus requiring solutions for accessing copyrighted materials or materials for which access is restricted due to privacy reasons. Second, appropriate data access is about "searchability", the availability of fine-grained metadata for retrieval, and about the required insight into the quality of the sources and the metadata needed for data analytics. Metadata, traditionally created manually for the data sets scholars are interested in, is typically sparse, quite diverse, and often incomplete. Apart from indexing this metadata properly, providing insight into this diversity is crucial for scholars to assess the quality of a search result, its significance to a research question, or validity of an analysis. This is traditionally referred to as *source criticism*, and currently referred to as *digital source criticism* (Hoekstra et al., 2018).

Emerging methods to generate metadata automatically, using for example automatic speech recognition (ASR) or computer vision technology, may bridge the gap between metadata sparsity and distant reading requirements of scholars, but they also bring up technological and methodological challenges. For example, questions arise on how can we efficiently generate high quality metadata for large amounts of "locked" content using automatic metadata extraction technology, and how the use of this type of metadata –that may still have classification errors or may be sensitive for biases– have to be accounted for in the interpretation of results and thus impact the methodology of scholars. Raising awareness about the operation of computational instruments for data extraction and processing and their impact on the heuristics and results of data-driven research is referred to as *digital tool criticism* (Koolen et al., 2018).

Eventually, enabling scholarly research that supports source and tool criticism should be reflected in the design of a user-interface that balances ease of use with the need to provide transparency regarding the scope and quality of the underlying data and their processing. As scholars have a wide variety of research interests, and also, have different levels of computer literacy –hence skills, or lack of skills to apply specific data processing tools themselves, for example for creating visualisations or applying content analysis tools–, the *interaction* with data and tools should be balanced accordingly: allowing for specific, specialised functions from individual scholars, without impeding the generic functions that apply to a wider community.

To solve the locked data problem and still allow for a flexible interaction with data and tools, the central approach of the *Media Suite* is to "bring the tools to the data" –as opposed to "bringing the data to the tools" that is custom in many other research areas– and to provide mechanisms that enable researchers to work with data and tools *within* the closed environment of the infrastructure, sealed with a federated authentication mechanism. In the past, substantial effort was undertaken to develop specific tooling that eventually could not be connected properly to work with the data collections they were intended for, due to access restrictions. In that sense, the *Media Suite* functions as a "virtual research environment" (VRE) that facilitates the proper functioning of the tools in the context of research and cultural heritage institutions. As a consequence, this research environment has a special liability towards the data and tools it provides in terms of transparency (source criticism), credibility (tool criticism) and flexibility.

Figure 1 shows the main elements that constitute the *Media Suite* research environment. Below we discuss shortly each of these elements.

2 Data Sources – Data Governance

The *Media Suite* currently provides access to audiovisual collections and multimedia context collections[2] from the following institutions, among others: (a) The Netherlands Institute for Sound and Vision (NISV), offering about a million hours of radio and television, film and oral history collections, including photos and digitised program guides and audience ratings), (b) Eye film institute, initially providing

[2] http://mediasuitedata.clariah.nl/dataset

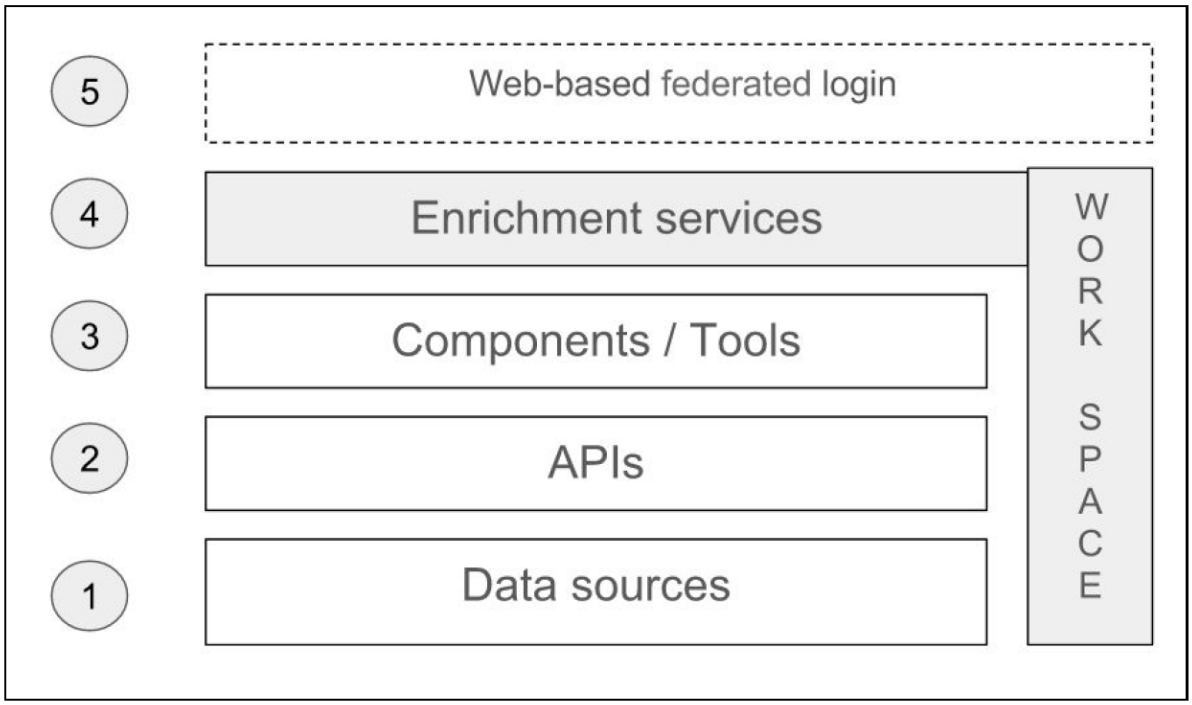

Figure 1: The building blocks of the CLARIAH Media Suite

access to the UNESCO world heritage Jean Desmet collection, including films, paper and poster collections, and (c) Oral History collections from various organisations in the Netherlands deposited at DANS. Also, although not an audiovisual collection, the large Dutch newspapers collection (more than 100 million pages) from the Dutch National Library is an important part of the *Media Suite*, as it allows scholars to make comparisons across media types.

To make these collections available in the *Media Suite*, we adhere to the general principle that collection owners provide access to collection metadata via the OAI-PMH protocol that enables the *Media Suite* to harvest the metadata and index it. It is assumed that the link to the source data (e.g., a video, scan or transcription) is incorporated in the metadata and points to a (streaming) sever hosted by the collection owner. Access restrictions (i.e., who is allowed to access what), is then organised at a broader –currently only national but ultimately international– research infrastructure level (CLARIAH[3], CLARIN[4]), via authentication and authorization protocols. In an ideal scenario, collection owners register and update their collections in a collection registry (we currently use CKAN[5]), that is "read" by the *Media Suite* for harvesting[6]. In practice however, we often have had to adapt this approach to the reality of sub-optimal situation with respect tot data governance at institutions. Institutional collection maintainers have internal data governance processes to ensure that data assets are formally managed. Data governance with respect to external processes – loosely defined as being part of an 'infrastructure' – is typically not accounted for at the institutions involved. This means that key data governance areas such as availability (e.g., metadata can be harvested), usability (e.g., source data can be viewed), integrity (e.g., protocols are in place to handle duplication and enrichment), and security (e.g., provenance information is maintained), need to be (re)organised or (re)considered, formalised and supported by the *Media Suite* and the emerging infrastructure in which it is embedded. From the practical point of view of making collections available

[3] https://www.clariah.nl/
[4] https://www.clarin.eu/
[5] https://ckan.org/
[6] http://mediasuitedata.clariah.nl/

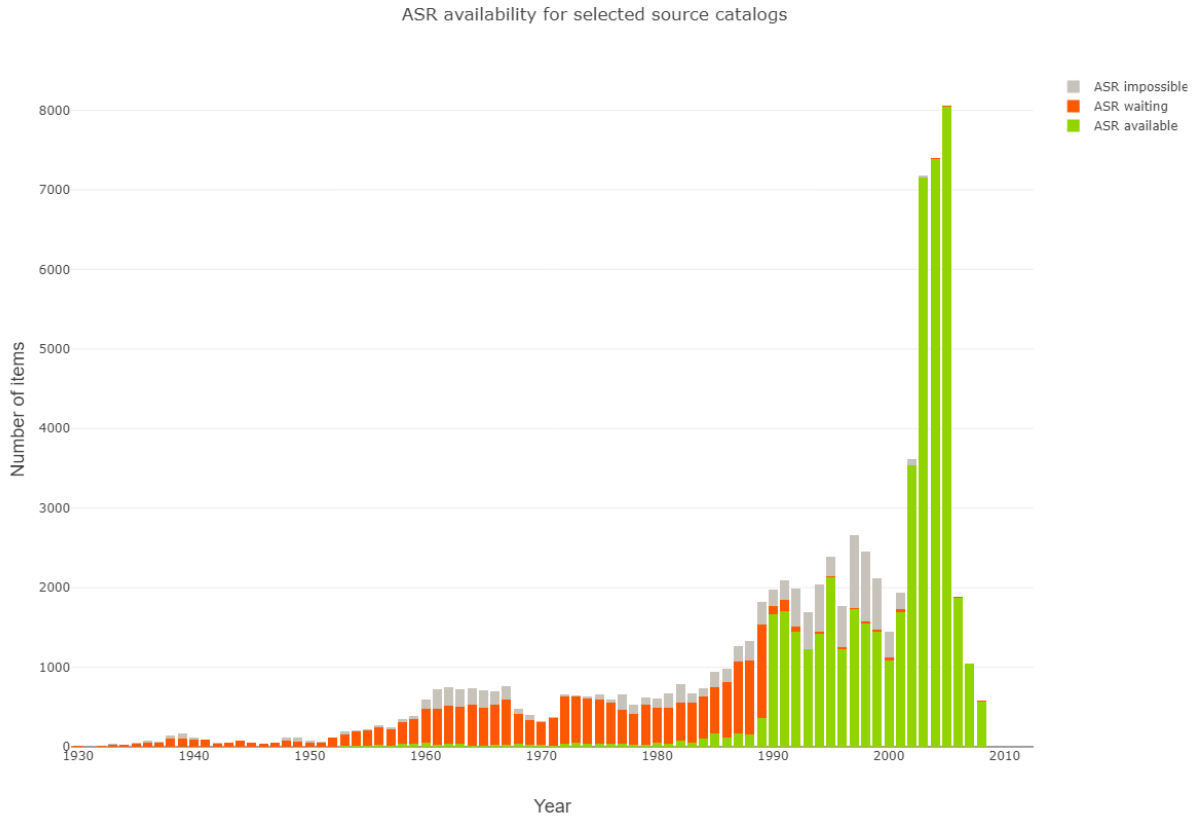

Figure 2: This chart shows the availability of automatic speech recognition (ASR) transcripts for the source catalogs per year. The green bars show the material that has ASR, the orange bars show the material that does not have ASR yet, the grey bars show the material for which ASR is currently impossible (no digital version). Screenshot date: January 2019

in the *Media Suite*, this drills down to manually mapping the various metadata models to a coherent search index and unravel technical issues for each collection individually.

The wish to use automatic metadata extraction technology to improve fine-grained, time-based access to audiovisual content is an additional complication in the provisioned distributed data chain. Here, the model is to provide services such as automatic speech recognition (Ordelman and van Hessen, 2018) for collections (”bulk” processing, up to 100 hours or more) via the infrastructure. However, making such a service available in a robust and collection-owner-friendly manner –a rather complex endeavour in itself– is only part of the work. Collection owners also need to arrange and manage internal data workflows: feeding the service with content and incorporating the output of the service (e.g, time-coded speech recognition transcripts) in the existing metadata model (including provenance information). The current status is that we have a speech recognition service available in the infrastructure tat operates faster then real-time and capable of processing approximately 1.000 hours per day. It is currently processing the NISV archive going backwards in time but also taking priority requests from scholars (e.g., to process news and actualities first). See Figure 2 that gives an impression of the progress for the NISV catalogs.

For the upcoming years, the goal is to connect more data collections from individual collection owners in the Netherlands, and increase the quantity and quality of metadata, focusing on both internal data governance processes and the use of automatic metadata extraction technology. Also, the incorporation of social media data, in particular data that are related to the ”traditional” media collections (e.g., hashtags related to television programming), is targeted. Finally, we want to make it possible for individual

scholars to upload their own data sets. Although the focus of the infrastructure is especially on opening up the ”locked” institutional collections, we noticed that scholars may also want to include in their analysis data sets coming from elsewhere or created by an individual scholar for a specific purpose, such as social media data on a specific topic, or recorded interviews.

3 Sustainable development (APIs)

The development of a digital infrastructure that is ”sustainable”, to make sure that it will remain available and maintained in the long run, including support, updates and upgrades, is central to the CLARIAH project as a whole, and specifically to the CLARIAH centres appointed by the project to support their domain-specific parts of the infrastructure [7] and to foster interoperability between these parts. For the *Media Suite*, this part of the infrastructure covers Dutch audiovisual collections augmented with available multimedia context collections. Examples of interoperability, are the connection with the CLARIAH infrastructure that focuses on textual data, containing the newspaper collections of the Dutch National Library that were mentioned before, and initial steps to link collections via Linked Open Data.

To foster sustainability we have to find a middle ground between the wishes of scholars and institutional ICT development and maintenance frameworks. Another critical requirement in this context is that the research infrastructure should also comply with other types of infrastructures that are being developed, such as in The Netherlands the infrastructure for digital heritage (Network Digital Heritage - NDE[8]) and, in a European context, the infrastructure components developed in the CLARIN and DARIAH ERICs.

Practically, this means that the infrastructure adheres to existing protocols, conventions, and standards. Moreover, to warrant interoperability and avoid proliferation of functions and processes (resulting in what is sometimes called a cauliflower architecture), a –from a research project point of view– rather strict development regime is followed, enforced by sprint plannings, focusing on a modular organisation of *Media Suite* components via application programming interfaces (APIs) that can be shared within the infrastructure. Examples of these APIs are a Collection API that provides high-level information (metadata) about the collections, such as which collections are available, data format, and volume, a Search API that allows searching the available collection indices, and the Annotation API that provides functionality for data annotation using the W3C Web Annotation data model (Sanderson et al., 2017).

4 Tools and user-friendly interaction design

The APIs discussed above are the corner stones for the development of the tools needed by scholars for doing their research. The development of these tools is to a large extent driven by requirements that were articulated in prototype applications built in earlier projects, such as video search and comparative analysis of media in *AVResearcherXL* (Van Gorp et al., 2015), search and visualization of results in *TROVe*[9], multi-collection search in *CoMeRDa* (Bron et al., 2013), exploratory search in *DIVE* (De Boer et al., 2015) and Oral History research in *Verteld Verleden* (Ordelman and de Jong, 2011). With a few exceptions and some ongoing work, the methods and functions underlying these prototypes have been extracted and re-implemented in the *Media Suite*.

The digital humanities community incorporates a wide diversity of scholars with different research questions, methods, and levels of expertise in working with information processing techniques and technologies. To address the challenges this imposes on requirements elicitation, development and evaluation of both re-implementations and new tools, the *Media Suite* team follows the principles of co-development where programmers and researchers work closely together, involving also the research community immediately via component testing, hackathons, datathons, public fora, and workshops. Because the use scenarios of scholars are diverse, it is even more important to focus on the similarities in research methods from different disciplines (de Jong et al., 2011; Melgar Estrada and Koolen, 2018), and to take

[7] https://clariah.nl/over/organisatie/centra
[8] https://www.netwerkdigitaalerfgoed.nl/en/
[9] https://www.clariah.nl/en/projects/finished/seed-money/trove

Figure 3: Collection Inspector: metadata information and completeness per field

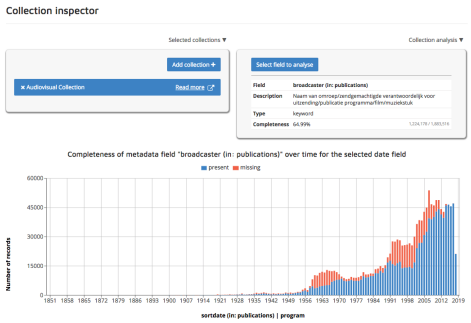

Figure 4: Collection Inspector: metadata completeness per field over time

these similarities as the baseline for tool development. Additionally, the concept of the scholarly primitives (Unsworth, 2000; Blanke and Hedges, 2013) serves as valuable guidance for identifying must-have functions in the *Media Suite*, and a model for a coherent and user-friendly design of the interface that fits in with the daily practice of scholars. Finally, as developing new tools "from scratch" for every research question would be a very inefficient (and costly!) endeavour, the analysis of tools that are "out there" has been taken up, resulting for example in a comparative study of qualitative data analysis software (Melgar et al., 2017), that provides the clues for deciding which tools we will or will not build ourselves and what type of data export functions and formats to support, of course within the boundaries of copyright and privacy restrictions.

A tool that was not directly based on previous work but actually emerged from working with "real" data has been coined as the Collection Inspector tool. As referred to in section 2, the metadata of collections from various collection owners is very heterogeneous, may not be complete, and may require some "metadata archaeology" to find out the proper meaning of fields, a meaning that may have changed in the course of an archive's history due to protocol and vendor changes. From a search perspective –the *Media Suite* allows scholars to design their own facets or filters based on available metadata fields– incompleteness and meaning of these fields is highly significant, and may lead to misinterpretations, for example with respect to recall, the search equivalent for completeness. The Collection Inspector enables scholars to assess the collection metadata, providing field descriptions, type, overall completeness, and completeness over time. Figures 3 and 4 show screen-shots of the Collection Inspector, on the left the descriptions and overall completeness data per field, on the right completeness of a single field over time. Together with the before-mentioned collection registry tool (CKAN), which contains information and visualisations that provide aggregate views on the content, scope and quality of the collections as well as their digital processing, and the options for scholars to define their own metadata filters, the collection inspector tool brings a valuable facility to the *Media Suite* for conducting digital source criticism.

Working with real data and the possibility to access (viewing/listening) the content itself was often very limited in the earlier prototypes due to the "locked data" problem. This underlined the importance of a well-thought-out design of content viewing/listening functions in relation to other functions that are associated with content-level, or in retrieval terms, document-level access, such as annotation, document level browsing, and within/cross-collection linking and recommendation. We grouped such functions in what we call the "Resource viewer" tool that currently incorporates playing video (also full-screen) and audio, annotation (see Figure 5), within-document browsing based on time-coded metadata such as speech transcripts (see Figure 6), and browsing all available metadata for the resource. However, while working with the Resource Viewer, scholars immediately suggest several opportunities to enhance "distant" reading on the document level. Note that audiovisual documents can be long and lack structure

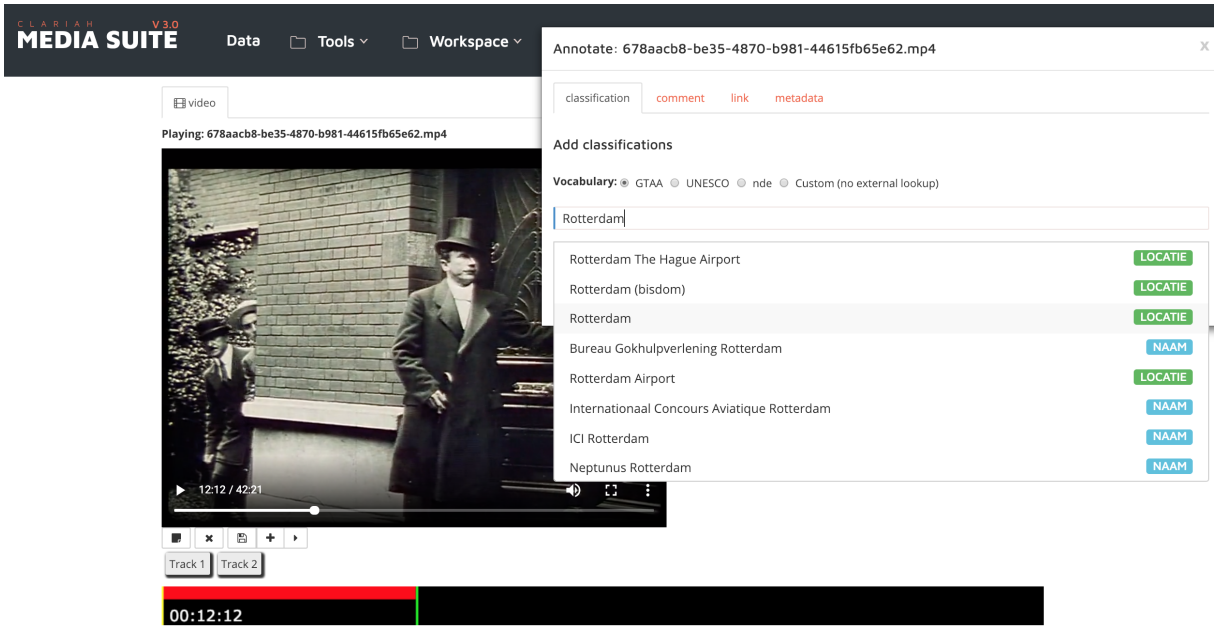

Figure 5: Annotating a video item in the Resource Viewer of the Media Suite using the NISV Audiovisual Thesaurus (GTAA) to label a segment, in this case with the location ”Rotterdam”.

such as paragraphs and headings in text. Word clouds, segmentation (e.g., based on shots) and segment-labelling (e.g., based on topic detection), or summaries based on speaker or face recognition, could alleviate this lack of structure and smooth the analysis of individual resources.

Zooming in from tools that put more emphasis on the distant reading part, as has been the focus until recently (data registry, collection inspection, search end exploration facilities as shown in Table 1) to tools that operate on levels of resource analysis and close reading, the further development of the Resource viewer requires special attention in the forthcoming period.

5 Workspace – working with personal virtual collections

In addition to copyright and privacy restrictions, access to the audiovisual content in the *Media Suite* is also limited due to its nature; consisting of pixels (video) and samples (audio) and hopefully some manually generated metadata or subtitles (text). Typically, scholars want to search audiovisual data using (key)words that may be ‘hidden’ (encoded) in the pixels or the samples. This is called the semantic gap (Smeulders et al., 2000) that needs to be “bridged” by decoding the information in the pixels and the samples to semantic representations, e.g., a verbatim transcription of the speech or labels of visual concepts in the video (a car, a face, the Eiffel Tower), that can be matched with the keywords from the scholars. These semantic representations can be generated manually or, especially when data collections are large, automatically using automatic speech recognition (ASR) or computer vision technology.

The generation of semantic representations is addressed in different ways. One the one hand, tools such as ASR are regarded as ‘must have’ components in an infrastructure focusing on fine-grained access and ’distant reading’ of large data sets. We are implementing an automatic speech recognition service that resides within the CLARIAH infrastructure and that can handle requests from the infrastructure itself (e.g., bulk processing of collections, possibly activated by a scholar with an interest in a specific data set), but also requests from individual scholars that want to process their private collections. On the other hand, supporting manual annotation is key for interpretation in scholarly contexts.

The *Media Suite* aims to support the generation of both ways of semantic representations in complementary ways via information work-flows centred around a workspace. More in general, the workspace

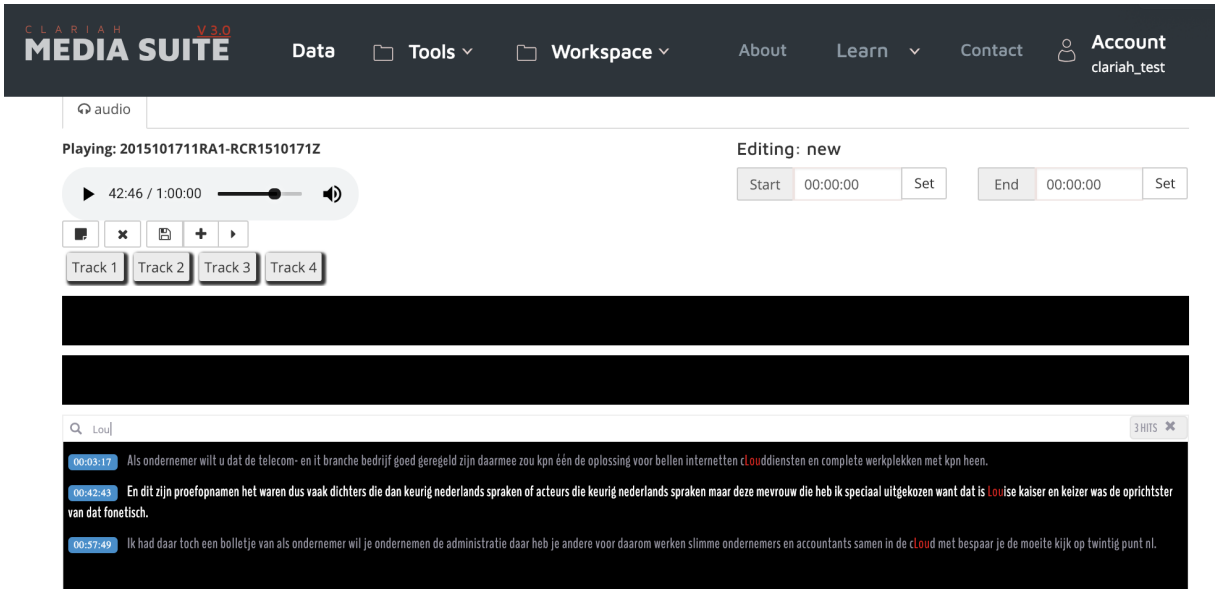

Figure 6: Browsing a radio item via speech transcripts in the Resource Viewer of the Media Suite. Via a within-document search function users are able to jump to specific parts of the radio broadcast, in this case where ”Louise Kaiser” is mentioned.

serves as the foundation for a scholar's research projects from which s/he can create and describe projects, organise collaborations (e.g., with peers or students), and keep track of all data –bookmarked selections from collections, annotations, visualisations– related to the projects. Figure 7 shows a screen-shot from the workspace, in this example on a project that is about media presentations of the 1953 flood disaster in The Netherlands. From within the workspace, a scholar can directly access the resource viewer for a stored resource, access saved session data such as queries and filtering options, and upload external data that are relevant for a project.

A special facility in the workspace is the option to generate your own visualisations using Jupyter Notebooks[10] and the (protected) APIs developed in the infrastructure. Jupyter Notebooks serve as a programming interface that allows scholars with programming skills to write their own code for creating overviews of the data, investigating a section of interest, performing advanced data analysis, and generating complex visualisations. In this way, we bring programming facilities to the data and to use third party code such as visualisation libraries and language processing toolkits (Wigham et al., 2018) to extend and complement their use of the *Media Suite*'s graphical user interface (Melgar et al., 2019).

6 Conclusion and future work

We described the *Media Suite* and its underlying infrastructure, and the challenges in building such an infrastructure that satisfies the needs of humanities scholars working with audio-visual media and contextual collections. We chose the approach of building a research environment that adheres to infrastructural requirements while at the same time being flexible, transparent, and user-friendly. In order to develop this environment in a sustainable way, that can be used and developed further after the project's lifetime, we need to carefully align the requirements of scholars with the context of the ecosystem the *Media Suite* needs to live in: an ICT infrastructure hosted and maintained by multiple institutions that in turn, adheres to a diverse set of institutional requirements with respect to, for instance, data access permissions and software development and maintenance. In order to have this infrastructure it is required that it is generic enough to cater for the general needs of every group that we have identified, while at the same time it

[10] `https://jupyter.org/`

Discovering	Overview of available collections via the collection registry (CKAN); Advanced search with options for filtering; Segment-level search on the basis of time-coded speech transcripts; Exploratory search via linked open data; A resource viewer for viewing and analysis of individual media items; Automatic metadata extraction technology
Annotating	Time/space-based multimedia annotation including segmenting, commenting, adding user metadata, links to other information sources, and use of code-book/thesaurus labels.
Comparing	Cross-media and cross-collection comparisons via saved queries
Sampling	Create personal virtual collections based on selections (bookmarks) stored in a personal workspace (see also section 5 below)
Illustrating	Generic visualisations of search results, flexible creation of ad-hoc visualisations using Jupyter Notebooks (see also section 5 below)
Representing	Understood as the need to support the ”presentation” phase of research, for example via enhanced publications with links to *Media Suite* content on the segment level (Van Den Heuvel et al., 2010). Support by the *Media Suite* is currently limited, as the infrastructure still lacks options for generating persistent identifiers on the segment-level

Table 1: Media Suite tools categorised via scholarly primitives

incorporates flexible functionality capable of addressing very specialised research questions. The *Media Suite* is currently functional and used by scholars doing actual research projects. Further development will focus on improving the current implementation of functions (e.g., development of a CLARIAH-wide annotation client[11], various interface improvements), adding collections, including new types such as social media data, increasing metadata granularity using automatic metadata extraction (e.g., speaker labelling, face recognition), and in particular, enhancing the functionality of the Resource viewer and Workspace. Also, we intend to setup a large system evaluation by a group of users outside the project to benchmark the current version of the system.

7 Acknowledgements

The research for this paper was made possible by the CLARIAH-CORE project (`www.clariah.nl`) financed by NWO. This paper is the result of a joint effort, specifically a close collaboration between all scholars, software developers, and domain specialists, listed here: `http://mediasuite.clariah.nl/documentation/faq/who-develops`.

[11]For work in progress see: `https://clariah.github.io/scholarly-web-annotation-client/`

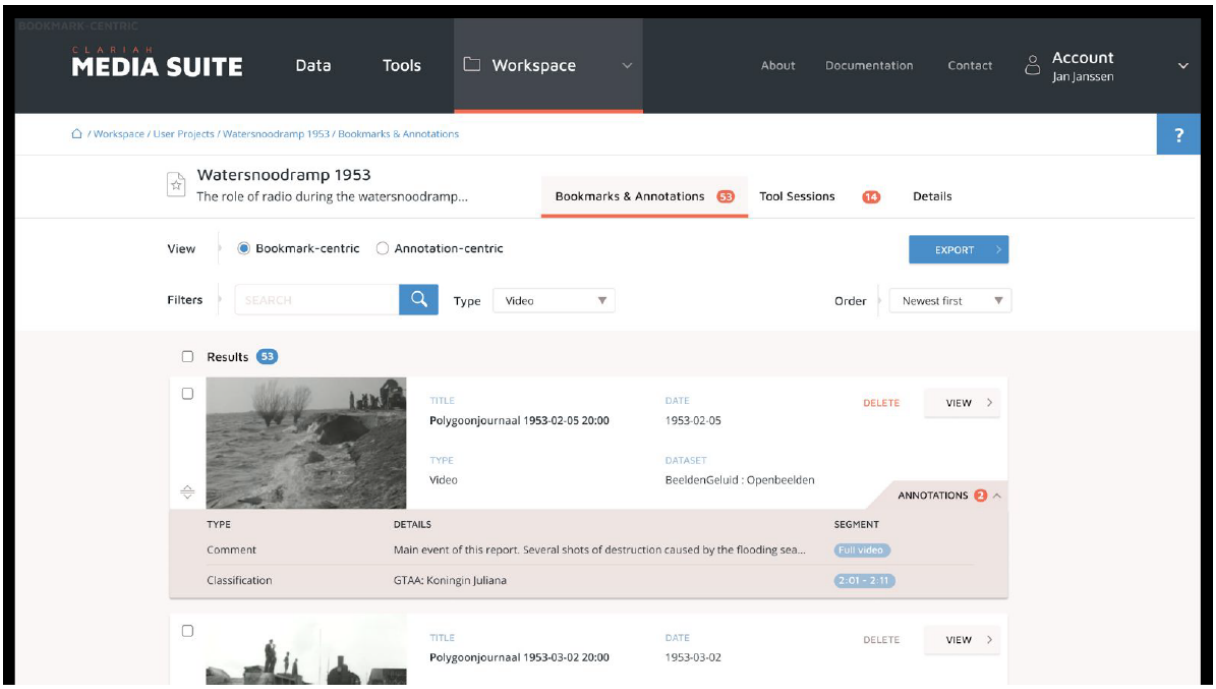

Figure 7: The CLARIAH Media Suite's Workspace

References

Tobias Blanke and Mark Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. 2013. Aggregated search interface preferences in multi-session search tasks. In *SIGIR '13: 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, July.

Marc Bron, Jasmijn Van Gorp, and Maarten de Rijke. 2016. Media studies research in the data-driven age: How research questions evolve. *Journal of the Association for Information Science and Technology*, 67(7):1535–1554.

Victor De Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco Van Staveren, Werner Helmich, and Dennis De Beurs. 2015. Dive into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:152–158.

Franciska de Jong, Roeland Ordelman, and Stef Scagliola. 2011. Audio-visual collections and the user needs of scholars in the humanities: a case for co-development. In *Proceedings of the 2nd Conference on Supporting Digital Humanities*, November.

FG Hoekstra, Marijn Koolen, and Marijke van Faassen. 2018. Data scopes: towards transparent data research in digital humanities. *Digital Humanities 2018 Puentes-Bridges*.

Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen. 2018. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*.

Liliana Melgar, Marijn Koolen, Hugo Huurdeman, and Jaap Blom. 2017. A process model of scholarly media annotation. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 305–308, New York, NY, USA. ACM.

Liliana Melgar, Marijn Koolen, Kaspar van Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martínez Ortíz, and Roeland Ordelman. 2019. The CLARIAH Media Suite: a hybrid approach to system design in the humanities. In *CHIIR 2019: ACM SIGIR Conference on Human Information Interaction and Retrieval*, Glasgow, Scotland, UK.

Liliana Melgar Estrada and Marijn Koolen. 2018. Audiovisual media annotation using qualitative data analysis software: A comparative analysis. *The Qualitative Report*, 23(13):40–60.

Franco Moretti. 2013. *Distant reading*. Verso Books, London.

Roeland J.F. Ordelman and Franciska M.G. de Jong. 2011. Distributed access to oral history collections: Fitting access technology to the needs of collection owners and researchers. In *Digital Humanities 2011: Conference Abstracts*, pages 347–349. Stanford University Library.

Roeland Ordelman and Arjan van Hessen. 2018. Speech recognition and scholarly research: Usability and sustainability. In *CLARIN 2018 Annual Conference*, pages 163–168.

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2017. Web annotation data model. *W3C Candidate Recommendation*.

Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380.

John Unsworth. 2000. Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In *Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London*, volume 13, pages 5–00.

Henk Van Den Heuvel, René Van Horik, Stef Scagliola, Eric Sanders, and Paula Witkamp. 2010. The veterantapes: Research corpus, fragment processing tool, and enhanced publications for the e-humanities. In *LREC*.

Jasmijn Van Gorp, Sonja de Leeuw, Justin van Wees, and Bouke Huurnink. 2015. Digital media archaeology: Digging into the digital tool avresearcherxl. *VIEW. Journal of European Television History and Culture/E-journal*, 4(7):38–53.

Mari Wigham, Liliana Melgar Estrada, and Roeland Ordelman. 2018. Jupyter Notebooks for generous archive interfaces. In *IEEE Big Data 2018: 3rd Computational Archival Science (CAS) Workshop*, Seattle, WA.