Media Suite

Unlocking Archives for Mixed Media Scholarly Research

Noordegraaf, J.J.

CLARIN Annual Conference 2018

PROCEEDINGS

Edited by

Inguna Skadiņa, Maria Eskevich

8-10 October 2018
Pisa, Italy

Programme Committee

Chair:

- Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia & Tilde (LV)

Members:

- Lars Borin, Språkbanken, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Koenraad De Smedt, University of Bergen (NO)
- Griet Depoorter, Institute for the Dutch Language (NL/Vlanders)
- Jens Edlund, KTH Royal Institute of Technology (SE)
- Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute (SI)
- Francesca Frontini, University of Montpellier (FR)
- Eva Hajičová, Charles University (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Bente Maegaard, University of Copenhagen (DK)
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences (AT)
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (EL)
- Kiril Simov, IICT, Bulgarian Academy of Sciences(BG)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Ilze Auziņa, LV
- Bob Boelhouwer, NL
- Daan Broeder, NL
- Silvia Calamai, IT
- Roberts Darģis, LV
- Daniël de Kok, DE
- Riccardo Del Gratta, IT
- Christoph Draxler, DE
- Dimitrios Galanis, GR
- Maria Gavrilidou, GR
- Luís Gomes, PT
- Normunds Grūzītis, LV
- Jan Hajič, CZ
- Marie Hinrichs, DE
- Pavel Ircing, CZ
- Mateja Jemec Tomazin, SI
- Neeme Kahusk, EE
- Fahad Khan, IT
- Alexander Kŏnig, IT
- Jakub Mlynar, CZ
- Jiří Mírovský, CZ
- Marcin Oleksy, PL
- Petya Osenova, BG
- Haris Papageorgiou, GR
- Hannes Pirker, AT
- Marcin Pol, PL
- Valeria Quochi, IT
- João Rodrigues, PT
- Ewa Rudnicka, PL
- Irene Russo, IT
- João Silva, PT
- Egon W. Stemle, IT
- Pavel Stranak, CZ
- Thorsten Trippel, DE
- Vincent Vandeghinste, BE
- Jernej Vičič, SI
- Jan Wieczorek, PL
- Tanja Wissik, AT
- Daniel Zeman, CZ
- Claus Zinn, DE
- Jerneja Žganec Gros, SI

CLARIN 2018 submissions, review process and acceptance

- Call for abstracts: 17 January 2018, 28 February 2018
- Submission deadline: 30 April 2018
- 77 submissions in total were received and reviewed (three reviews per submission)
- Face-to-face PC meeting in Wroclaw: 21-22 June 2018
- Notifications to authors: 2 July 2018
- 44 accepted submissions: 21 oral presentations, 23 posters/demos

More details can be found at https://www.clarin.eu/event/2018/clarin-annual-conference-2018-pisa-italy.

Table of Contents

Parallel Session 3: Use of the CLARIN infrastructure

Parallel Session 4: Design and construction of the CLARIN infrastructure

Poster session

Media Suite: Unlocking 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

Carlos Martinez-Ortiz
Netherlands eScience Center
Amsterdam
The Netherlands
c.martinez@esciencecenter.nl

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

Abstract

This paper discusses the rationale behind the development of a research environment –the Media Suite– in a sustainable, dynamic, multi-institutional infrastructure that supports mixed media scholarly research with large multimedia data 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 astronomy for instance, new collections of astronomical observations are made publicly available on a regular basis. In other domains such as Media Studies research focuses on data collections maintained at cultural heritage institutions, archives, libraries, and 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 (play-out, viewing). Therefore, scholars either fall back on collections that are openly available or spend considerable amounts of time in *onsite* visits to archives for consulting data collections. 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 *mixed-media* sources, such as radio and television broadcasts, film, oral history interviews but also (news)paper archives, film posters and eyewitness reports. The *Media Suite* serves as the online portal to the infrastructure where first of all, content and metadata can be explored, browsed, compared, and stored in personal collections. In addition, the Media Suite provides a workspace for working with mixed media collections, providing tools for manual and automatic annotation, visualization, analysis, and sharing.

The ultimate goal is to (i) enable distant reading (Schulz, 2011), that is, identifying patterns or new research questions in all aggregated collections, (ii) facilitate close reading: the detailed examination of individual items (e.g., videos) in a collection or parts of these items (e.g., video segments) during search and scholarly interpretation, and (iii) make sure that the "scholarly primitives" (Unsworth, 2000;

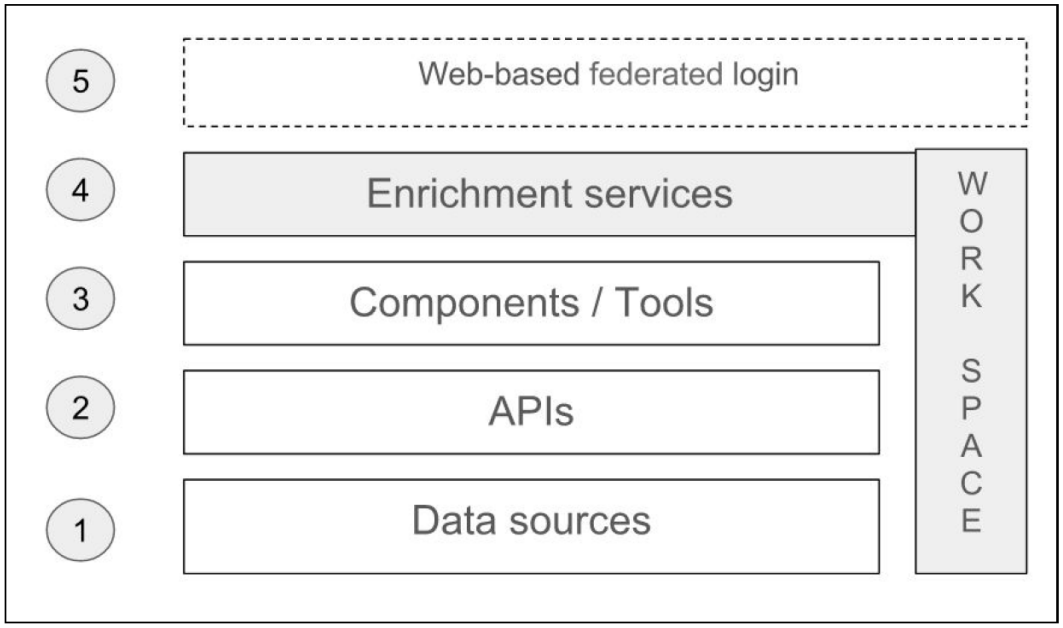

Figure 1: The building blocks of the CLARIAH Media Suite

Blanke and Hedges, 2013), basic activities common to research across humanities disciplines, are well supported.

1.1 Challenges

Questions however are: How to facilitate "close reading" when the media objects cannot be accessed because of copyright issues? How to enable "distant reading" when metadata is sparse, or diverse, and incomplete? How to cater to the needs of scholars with specific research questions and methods in the context of an infrastructure that has to be generic enough to be feasible? How to enable scholars to work with collections from different institutes using the same tools, when these collections are "locked"? How to enable scholars that are computer literate to work directly with the data or to deploy private content analysis tools such as computer vision or sentiment analysis?

The approach of the CLARIAH Media Suite to tackle these challenges is to provide mechanisms that enable researchers to work with tools and aggregated data *within* the closed environment of the infrastructure sealed with a federated authentication mechanism (SURFConext[1]) that currently only serves scholars with a university account in the Netherlands, but that soon will be expanded to the CLARIN federation. Also, the so called 'homeless users' that do not have an account with an academic institution, will eventually have the opportunity to request for a login. We refer to this approach as to "bringing the tools to the data", as opposed to "bringing the data to the tools".

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

Institutional collection maintainers have internal data governance processes to ensure that data assets are formally managed. One important aspect covered by governance processes is licensing: who has

[1] https://www.surf.nl/en/services-and-products/surfconext/index.html

permission to access the data. However, data governance with respect to external processes –loosely defined as being part of an 'infrastructure'– is typically not accounted for. 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)organized or (re)considered, formalized and supported by the Media Suite and the emerging infrastructure in which it is embedded.

3 APIs – Sustainable development

A digital infrastructure should use existing protocols, conventions, and standards. Besides obtaining data by harvesting using the OAI-PMH protocol, or using application programming interfaces (APIs), the functionalities have been organized in a modular approach, which includes (Martinez-Ortiz et al., 2017):

- Components that use API's to perform specific tasks.
- Tools that incorporate a number of components in a tool.

4 Components/Tools – User-friendly interaction design

Developing new tools "from scratch" for every research question would be a very inefficient (and costly!) endeavour. The digital infrastructure should provide tools that are suitable both for common scholarly tasks and for specific tasks required by each discipline. However, 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. We address this challenge by (i) focusing on the similarities in research methods from different disciplines (de Jong et al., 2011; Melgar Estrada and Koolen, 2018), (ii) analyzing tools that support qualitative methods (Melgar et al., 2017), and (ii) working with scholars as co-developers in the process . The resulting functionalities are built in a modular (lego) approach that supports both flexible software development of components and user-friendly interaction with assembled tools.

5 Work Space – Working with audio-visual content and private data

In addition to IPR 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 Eifel 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. We are implementing an automatic speech recognition service that resides within the CLARIAH infrastructure 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 workflows centred around a "Work Space" (see Figure 2) that has the following functionalities:

- Storing individual items from different "institutional" collections resulting in a private, virtual, multimedia, research collection.
- Storing private session data such as queries and filtering options.

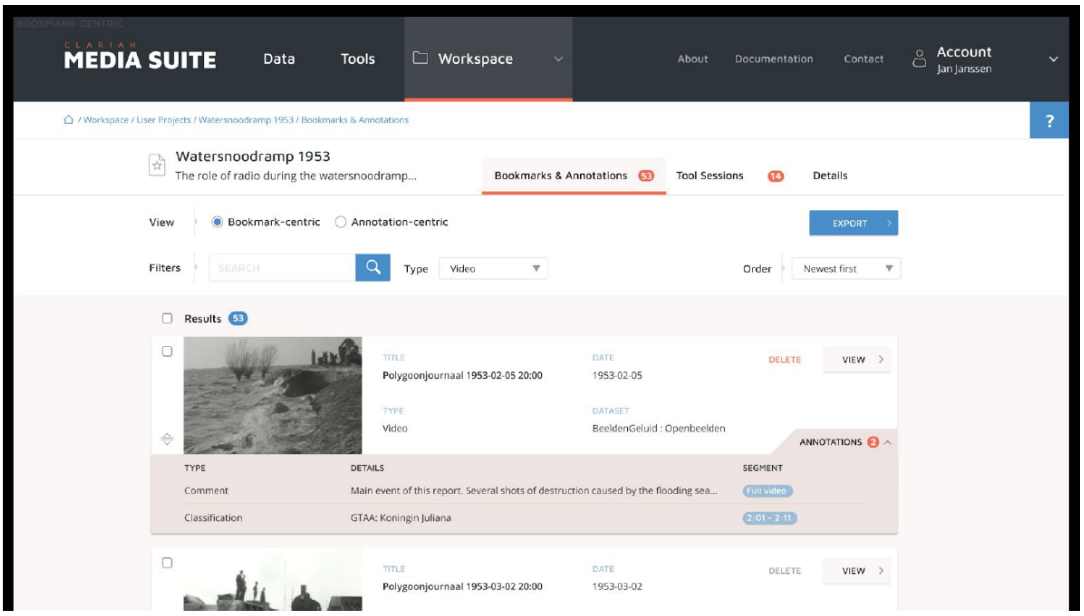

Figure 2: The CLARIAH Media Suite's Workspace

- Uploading private data and perform enrichment services to these data (e.g., speech recognition)
- Running private code on data collections in the infrastructure for creating data visualization (e.g., Jupyter Notebooks).

6 Conclusion and future work

We described the challenges found in building an infrastructure that satisfies the needs of humanities scholars working with audio-visual media and contextual collections. We choose the approach of building a research environment that adheres to infrastructural requirements while at the same time being flexible 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 incorporates flexible functionality capable of addressing very specialistic research questions. The Media Suite is currently functional and used by scholars doing actual research projects and will be developed further, e.g., by incorporating additional data sources (e.g., social media data), increasing metadata granularity (e.g., adding computer vision or emotion recognition), adding advanced annotation tools, and supporting missing data visualization (data critique) for heterogeneous datasets.

References

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

Franciska M.G. de Jong, Roeland J.F. Ordelman, and Stef Scagliola, 2011. *Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development*, pages –. Centre for Language Technology, Copenhagen, 11. eemcs-eprint-20868.

Carlos Martinez-Ortiz, Roeland Ordelman, Marijn Koolen, Julia Noordegraaf, Liliana Melgar, Lora Aroyo, Jaap Blom, Victor de Boer, Willem Melder, Jasmijn van Gorp, Eva Baaren, Kaspar Beelen, Norah Karrouche, Oana Inel, Rosita Kiewik, Themis Karavellas, and Thomas Poell. 2017. From tools to "recipes": Building a media suite within the dutch digital humanities infrastructure clariah. DHBenelux.

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

Kathryn Schulz. 2011. What is distant reading. *The New York Times*, 24.

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.