From Tools to “Recipes”: Building a Media Suite within the Dutch Digital Humanities Infrastructure CLARIAH

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

Introduction

Scholars require access to multiple, large, multimedia collections of digital resources, as well as to use a wide range of information processing tools to access and work with those collections. These requirements raise the need for developing a synchronized national and cross-national infrastructure.

Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH)[1] is a distributed research infrastructure for the Humanities, included on the National Roadmap for Large-Scale Research Facilities (2015-2018) drawn up by the Netherlands Organisation for Scientific Research (NWO). CLARIAH designs, implements and exploits the Dutch part of the European CLARIN and DARIAH infrastructures.

There are different research domains within CLARIAH: linguistics, socio-economic history, and media studies. Each work package within the CLARIAH project places at the centre of development both the technical requirements of each media type (text, structured data, audio-visual media), as well as the specific research needs of their user communities.

The CLARIAH Media Studies work package focuses on creating a research environment, the Media Suite (CLARIAH MS)[2], as part of the CLARIAH infrastructure aiming to serve the needs of media scholars by providing access to audio-visual collections and their contextual data. This paper describes the approach taken to build CLARIAH MS.

Background

CLARIAH MS incorporates a series of Digital Humanities (DH) tools and aims to make them sustainable. Prototypes are currently hosted on a new infrastructure at the The Netherlands Institute for Sound and Vision (NISV) data centre. These prototypes are: AVResearcherXL, TROVe, CoMeRDa, Oral History Today (OHT) and DIVE+. Furthermore, CLARIAH MS aims to support audio-visual archives in opening up collections in a more standardized way. Once these objectives have been accomplished, scholars will be able to search and analyse these collections via a central workspace, thus, enabling *data intensive research* in the humanities.

AVResearcherXL is an exploratory tool which enables simultaneous queries and analytic visualizations of the collections´ metadata (Van Gorp et al (2015)). TROVe was developed to

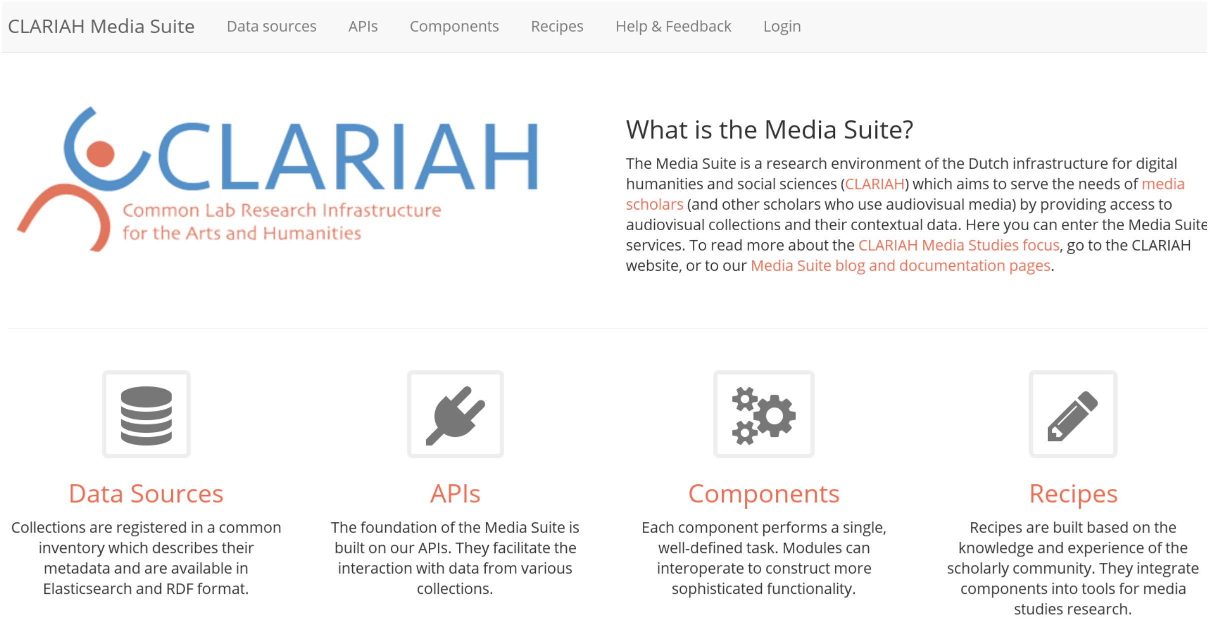
[1] http://www.clariah.nl/
[2] http://mediasuite.clariah.nl/

ease the combined access and visualization of archival collections and online social media. CoMeRDa is a web based aggregated search system for visualizing search results (Bron et al(2013)). OHT is a prototype for search and enrichment (through Automatic Speech Recognition technology) of distributed Oral History collections in The Netherlands (Ordelman and de Jong(2011)). Finally, DIVE+ is a linked-data digital cultural heritage collection browser which provides access to heritage objects from heterogeneous collections, using historical events and narratives as context for searching, browsing and presenting the objects (de Boer et al.(2015)).

These five tools support scholars in the “exploration” and “contextualization” phases of their research, a framework proposed in (Bron et al.(2015)). The original tools could not interoperate and did not operate on the same data, which limits their potential. Recreating them in a single configurable environment makes it possible to reuse functionalities across data sets and to reuse data across functionalities.

CLARIAH Media Suite

The DH community includes scholars with a wide diversity of research interests and goals; every research group in DH is working with different types of data and their research objectives have specific requirements which cannot be easily facilitated by tools using a single, generic approach. Simultaneously, there are similarities in the methods used by different scholars (de Jong et al.(2011)) that can be used for generalised tool development. There are commonalities in research questions and methods among media scholars, which we grouped into *Media aesthetics*, *Social history of media*, *Aesthetic historiography*, *Social and cultural history*, *Media representations or coverage*, *Transmedia analysis*, and *Memory studies (Melgar et al., 2017)*.



What is the Media Suite?

The Media Suite is a research environment of the Dutch infrastructure for digital humanities and social sciences (CLARIAH) which aims to serve the needs of media scholars (and other scholars who use audiovisual media) by providing access to audiovisual collections and their contextual data. Here you can enter the Media Suite services. To read more about the CLARIAH Media Studies focus, go to the CLARIAH website, or to our Media Suite blog and documentation pages.

Data Sources

Collections are registered in a common inventory which describes their metadata and are available in Elasticsearch and RDF format.

APIs

The foundation of the Media Suite is built on our APIs. They facilitate the interaction with data from various collections.

Components

Each component performs a single, well-defined task. Modules can interoperate to construct more sophisticated functionality.

Recipes

Recipes are built based on the knowledge and experience of the scholarly community. They integrate components into tools for media studies research.

Figure 1. CLARIAH MS consists of functionalities, APIs and recipes, version 1, April 2017

A generic infrastructure is required to cater for the general needs of every user group. The infrastructure needs to incorporate flexible functionality capable of addressing very specialized research questions. Media scholars expressed their desire to use the collections and tools which were previously “locked” together in the individual prototypes. CLARIAH MS has been designed in a modular way (Figure 1); each module performs a single, well-defined task. Modules can interoperate to construct more sophisticated functionality.

Metaphorically speaking: whereas previously users had access to predefined ‘meals’ - tools which could perform cross-collection search and visualize the results in the form of timelines, word clouds, snippets and/or thumbnails - we now provide users with single ingredients (individual functionalities such as searching), and ready-made recipes (combinations of several functionalities). Some ingredients may be used in different recipes, existing recipes may be complemented by adding extra ingredients.

Media Suite Architecture

CLARIAH MS consists of four layers of functionality, explained below (Figure 2):

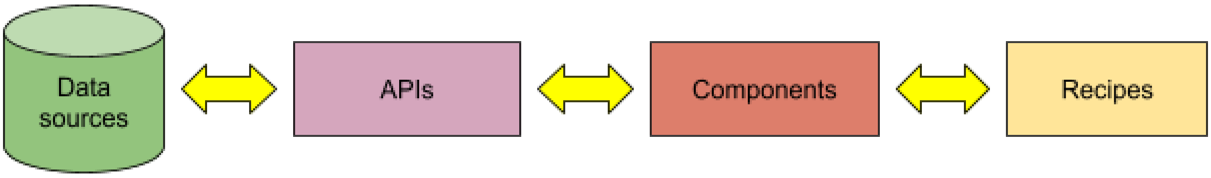


Figure 2 - Architectural design of CLARIAH MS.

Data Sources contain the collections (e.g., television broadcasts from NISV, EYE Jean Desmet collection, DANS Oral History collection). All collections are registered in a common inventory (CKAN[3]) which describes their metadata. Collections are available in Elasticsearch (full text search) and RDF format (semantic search).

APIs facilitate the interaction with data from various collections:

- Collections API - high-level collection information (metadata: data format, size, etc.)
- Search API - searching for collection items.
- Annotation API - annotating existing data using W3C Web Annotation standard (mainly for manual annotations)(Melgar et al.,(2017)).
- Data Enrichment API - collection enrichment through automatic mechanisms (e.g. name entity recognition) or by human interaction (e.g. crowdsourcing).

The APIs design allows the integration of new data of different formats and data models.

Components in CLARIAH MS are software units which perform a single functionality: each component takes data as input and produces a meaningful output using standard formats, to be connected with other components (e.g., word cloud, timeline visualizations, topic identification in newspapers, searching content in collections).

[3] http://mediasuite.clariah.nl/datasources

Recipes close the circle by integrating components to recreate the functionalities of the original tools. We focus on providing the complex functionality of the original tools in the form of four 'recipes'. Following with the metaphor above, the concept of ingredients (components) allows researchers to prepare their own personal recipes (functionalities).

Conclusion

In this paper we have explained the structure of the CLARIAH MS and how previously developed DH tools are being integrated in a sustainable infrastructure that allows flexible use of data collections and functionalities fitting the research needs of scholars. We have also sketched our strategy to enable the integration of alternative functionalities and data collections using a modular approach (ingredients and recipes). Future work includes user evaluation of the first version of the Media Suite (launched in April, 2017), and co-development involving six CLARIAH research pilot projects[4].

References

[Bron et al. (2013)] Marc Bron, Jasmijn Van Gorp, Frank F. Nack, Maarten de Rijke, Lotte B. Baltussen. *Aggregated search interfaces in multi-session tasks.* SIGIR 2013: 36th international ACM SIGIR conference on research and development in information retrieval. Dublin: ACM (2013)

[Bron et al.(2015)] Marc Bron, Jasmijn Van Gorp, and Maarten Rijke. *Media studies research in the data-driven age: How research questions evolve. Journal of the Association for Information Science and Technology* (2015), https://doi.org/10.1002/asi.23458.

[de Jong et al.(2011)] Franciska de Jong, Roeland Ordelman, and Stef Scagliola. *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 (SDH 2011), Copenhagen, Denmark, 2011. Centre for Language Technology, Copenhagen.

[Melgar et al.(2017)] Liliana Melgar Estrada, Marijn Koolen, Hugo Huurdeman, and Jaap Blom. *A process model of time-based media annotation in a scholarly context.* In ACM Conference on Human Information Interaction and Retrieval
(CHIIR), Oslo, 2017.

[Ordelman and de Jong(2011)] Roeland Ordelman and Franciska de Jong. *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, 2011. Stanford University Library. URL http://purl.utwente.nl/publications/78347. ISBN=978-0-911221-47-3.

[4] http://www.clariah.nl/projecten/research-pilots

[de Boer et al.(2015)] Victor de Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco van Staveren, Werner Helmich, Dennis de Beurs: *DIVE into the event-based browsing of linked historical media*. J. Web Sem. 35: 152-158 (2015)

[Van Gorp et al (2015)] Jasmijn Van Gorp, Sonja de Leeuw, Justin van Wees, Bouke Huurnink. *Digital Media Archaeology - Digging into the Digital Tool AVResearcherXL.* VIEW. Journal of European Television History and Culture/E-journal, 4 (7): 38-53 (2015)