DIVE+: Explorative Search for Digital Humanities in CLARIAH Media Suite

Oana Inel[2], Lora Aroyo[2], Jaap Blom[1], Victor de Boer[1,2], Werner Helmich[3], Liliana Melgar[5], Carlos Martinez Ortiz[4], and Johan Oomen[1]

[1]Netherlands Institute for Sound and Vision, Hilversum, the Netherlands
joomen@beeldengeluid.nl, jblom@beeldengeluid.nl

[2]Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
v.de.boer@vu.nl, lora.aroyo@vu.nl, oana.inel@vu.nl

[3]Frontwise, Utrecht, the Netherlands
werner@frontwise.com

[4]The Netherlands eScience Center, Amsterdam, the Netherlands
c.martinez@esciencecenter.nl

[5]University of Amsterdam, Amsterdam, the Netherlands
melgar@uva.nl

ABSTRACT

Keywords

Interdisciplinary Research, Heterogeneous Linked Data, Crowdsourcing, Digital Hermeneutics, Historical Events

1. PURPOSE

The Web has offered cultural heritage institutions a medium to make their cultural heritage collections publicly available online. Thus, there is an immerse need for them to rethink the access strategies to their collections to take a full advantage of the open Web infrastructure. In the same time, they also need to reinvent the support for research scholars and general audiences in their online explorations of these vast information spaces. In this way, cultural heritage institutions need to change their traditional task from information interpreters to that of information providers[4].

DIVE+ [1] is an event-centric linked data digital collection browser aimed to provide an integrated and interactive access to multimedia objects from various heterogeneous online collections. It enriches the structured metadata of online collections with linked open data vocabularies with focus on events, people, locations and concepts that are depicted or associated with particular collection objects. DIVE+ is result of a true inter-disciplinary collaboration between computer scientists, humanities scholars, cultural heritage professionals and interaction designers. As part of this effort, DIVE+ is also integrated in the CLARIAH[1] (Common Lab Research Infrastructure for the Arts and Humanities) research infrastructure, next to other media studies research tools, that aims at supporting the media studies researchers and scholars by providing access to digital data and tools.

CLARIAH aims to develop a digital research infrastructure to extend the research capabilities of humanities scholars. Specifically, in the context of media studies, CLARIAH develops the Media Suite[2] as a set of resources (i.e., tools and data) available to scholars interested in media studies. The Media Suite includes digital media collections such as news paper articles, radio shows, TV shows, art objects, among others and micro tools for better understanding these collections. Research in the context of CLARIAH [3] indicates that humanities scholars need four main steps to describe their research: *exploration*, *assembly*, *analysis* and *presentation*.

The exploration stage is critical for defining the research questions as it covers a throughout analysis and study of background materials for developing the research question, as well as initial information gathering. For this purpose, we integrate DIVE+ in the Media Suite as a digital hermeneutics to support the exploration stage of humanities research. Furthermore, DIVE+ is able to connect to the rest of the research stages through its functionality and its integrated and interlinked datasets: (1) assembling: find the relevant corpus with focus on finding links to additional collections (2) analysis: explore, visualize and interpret the corpus through understanding the enrichment and the links between concepts in the corpus; (3) presentation: create informative narratives through the innovative DIVE+ interface.

2. METHODS

As part of the CLARIAH Media Suite, DIVE+ has been designed as a collection of modular micro tools which can be reused separately or as a whole. The overall integration of DIVE+ and its functionality is depicted in Figure 1. DIVE+ functionality is accessed through the CLARIAH Media Suite Dashboard, which facilitates the humanities research by providing access to research technologies, functionalities, collections and enrichments across tools.

DIVE+ data collection is composed of media objects from different cultural heritage institutions which are described in Section 3. These media objects are provided in RDF format (as linked data) and can be accessed via a SPARQL endpoint. At the moment, all collections of DIVE+ are registered in a CKAN[3] registry, among other audio-visual collections, that can be accessed through the Collection API.

[1]`http://www.clariah.nl`
[2]`http://mediasuite.clariah.nl`
[3]`http://ckan.org`

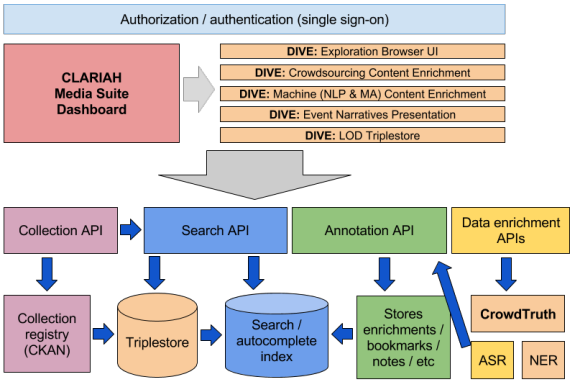

Figure 1: DIVE+ contribution and integration in the CLARIAH Media Suite infrastructure

In order to provide easy access to DIVE+ data, a SPARQLSearch micro tool has been developed. Thus, relevant DIVE+ SPARQL queries can be registered in the SPARQLSearch micro tool and then become available as a simple web-API. In this way, humanities scholars which are not familiar with SPARQL can still make use of DIVE+ data. Furthermore, the SPARQLSearch micro tool allows for the registration of other SPARQL queries, using other SPARQL endpoints if desired. SPARQL queries registered in this way should be made available via Github for the SPARQLSearch micro tool to be able to import them automatically.

DIVE+ also provides functionality for adding crowdsourced annotations to existing media objects in the DIVE+ data collection. Furthermore, the crowdsourcing component accessed through the CrowdTruth[4] [2] platform is also registered as an enrichment API in the core CLARIAH infrastructure. Thus, all collections registered in the Media Suite can access and benefit from collecting and enriching their media objects with crowd annotations. In a similar fashion other external services for data enrichment can be used, such as named entity recognition through state-of-the-art named entity recognition tools, sentiment analysis, video transcription, among others. Due to to the flexibility of the Media Suit design, humanities scholars can make use of the existing enrichment micro services or can easily create and register their own annotation tools as annotation services.

3. RESULTS

Demonstrating the digital hermeneutics approach [5], the DIVE+ browser allows for exploration of heterogeneous linked datasets containing media objects (e.g., images or videos). The metadata of these objects is enriched with entities such as events, persons, places and other concepts, depicted or associated with them. Currently, content from *four cultural heritage institutions* are made available through the DIVE+ SPARQL interface, on top of which an innovative event-centric user interface is implemented:

- *3000 Dutch news broadcasts (1920-1980)* form the Netherlands Institute for Sound and Vision (NISV)[5].
- *197,199 ANP Radio News Bulletins (1937-1955)* from the Dutch National Library (KB)[6].
- *3500 Cultural heritage objects (1950-1980)* from the Amsterdam Museum (AM)[7].
- *964 Cultural heritage objects (1950-1980)* from the Tropenmuseum (KIT) [8].

Additionally, in the DIVE+ triple store we extended the existing cultural heritage linked data cloud with an automatic alignment of the enriched metadata from the above collections with various structured vocabularies, e.g., Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA) and Amsterdam Museum Thesaurus, Persons list, Geo vocabulary, Stichting Volkenkundige Collectie Nederland (SVCN). Thus, the collections made available are interlinked in a common linked data network of events, persons, places and concepts, which provides context for browsing and exploration of the cultural heritage objects.

We enrich collection with events through crowdsourcing, through the CLARIAH Media Suite, by connecting to the CrowdTruth data enrichment API (see Figure 1). The enrichment from the different machines and crowd is consolidated to RDF and provided in the DIVE+ RDF Triple store with a SPARQL endpoint[9]. DIVE+ data uses the Simple Event Model (SEM) [6], which allows for the representation of events, actors, locations and temporal aspects. We extend SEM with additional Linked Data schemas, e.g., DC, SKOS, OpenAnnotation and FOAF to represent other types of resources linked to the media objects. Links are also established to external sources, Wikipedia and DBpedia. In the current triple store we host over 7.5 Million triples for the 210,000 Media Objects. These are annotated with 17,209 places, 95,977 actors and 199,116 event concepts.

Acknowledgements

We would like to thank The Netherlands eScience Center for funding the DIVE+ project.

4. REFERENCES

[1] V. de Boer, J. van Doornik, L. Buitinck, M. Marx, T. Veken, and K. Ribbens. Linking the kingdom: Enriched access to a historiographical text. In *Proceedings of KCAP 2013, Banff, Canada, 23-26 June 2013*, 2013.

[2] O. Inel et al. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC 2014*, pages 486–504. Springer, 2014.

[3] L. Melgar, M. Koolen, H. Huurdeman, and J. Blom. A process model of scholarly media annotation. In *CHIIR 2017, (to appear)*.

[4] K. Mueller. Museums and virtuality. In *In R. Parry, editor, Museums in a Digital Age, chapter 30, pages 295-305 Routledge*, 2007.

[5] C. van den Akker et al. Digital hermeneutics: Agora and the online understanding of cultural heritage. In *Proc. of the 3rd International Web Science Conference. ACM, New York, NY, USA, 7 pages.*, 2011.

[6] W. R. van Hage et al. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2), 2011.

[4] `http://crowdtruth.org/`
[5] `http://www.beeldengeluid.nl`
[6] `http://www.kb.nl`,`http://radiobulletins.delpher.nl/`
[7] `https://www.amsterdammuseum.nl`
[8] `http://www.opencultuurdata.nl/wiki/tropenmuseum/`
[9] ClioPatria triple store: `data.dive.beeldengeluid.nl`