From DIVE to DIVEplus: Exploring Integrated Linked Media

Oana Inel[2], Lora Aroyo[2], Dennis de Beurs[4], Jaap Blom[1], Victor de Boer[1,2], Annemarie Collijn[2], Werner HelmicDIVEHelmich[4], Christiaan Meijer[5], Johan Oomen[1], and Elco van Staveren[3]

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

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

[3]Dutch National Library, the Hague, the Netherlands
elco.vanStaveren@kb.nl

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

[5]The Netherlands eScience Center, Amsterdam, the Netherlands
c.meijer@esciencecenter.nl

ABSTRACT

DIVE is a linked-data digital collection browser aimed at providing an integrated and interactive access to multimedia objects from various heterogeneous online collections. It enriches the structured metadata of online collections with open linked data vocabularies with focus on events, people, locations and concepts that are depicted or associated with those collection objects. DIVEplus extends the innovative DIVE approach in four ways: (1) adding two new collections (i.e. Amsterdam Museum and Tropenmuseum), (2) integrating links to external linked open datasets (i.e. DBpedia, AAT and ULAN), (3) designing an intuitive way to deal with event narratives; and (4) automating the crowdsourcing of event annotations of the multimedia collection objects. The overall goal of DIVEPlus is to gain insights on the scalability, robustness and reusability of the DIVE digital hermeneutics approach by testing its usability in user studies with a variety of end users (both research scholars and general audiences).

Keywords

Heterogeneous Linked Data, Crowdsourcing, Digital Hermeneutics, Historical Events, Digital History

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[3].

In this paper we present DIVEplus, which advances the way in which researchers and general audience interact with online heritage collections by allowing an integrated exploration of objects of heterogeneous cultural heritage collections. As such, DIVEplus[1] extends the digital hermeneutics approach [4] of DIVE applied for cultural heritage collections and using historical events and event narratives as context for searching, browsing and presenting collection objects [2]. It builds on the DIVE demonstrator[2], where semantics from existing cultural heritage collections and linked data vocabularies are used to link objects with events, people, locations, time periods and other concepts that are depicted or associated with those objects. The innovative interface combines Web technology and theory of interpretation to allow for browsing this network of data in an intuitive "infinite" fashion. Thus, the main focus in DIVEplus is to provide support to both digital humanities scholars and general audience with an interest in history in their online explorations.

2. METHODS

Demonstrating the digital hermeneutics approach [4], the DIVEplus 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 *three cultural heritage institutions* are made available through the DIVEplus SPARQL interface, on top of which an innovative event-centric user interface is implemented:

- *Dutch news broadcasts* form the Netherlands Institute for Sound and Vision (NISV)[3]. Within the DIVEplus we ingested a randomly selected subset of about 300 videos from the NISV collection of broadcast video published as Open Data on the Openimages platform[4] from the period 1920-1980.
- *ANP Radio News Bulletins*[5] from the Dutch National Library (KB)[6]. In the DIVEplus we ingested 2210 KB digitized typoscripts (radio news scripts, to be read during news broadcasts) from the period 1937-1984.

[1]`http://diveplus.beeldengeluid.nl`
[2]`http://dive.beeldengeluid.nl`
[3]`http://www.beeldengeluid.nl`
[4]`http://openimages.eu`
[5]`http://radiobulletins.delpher.nl/`
[6]`http://www.kb.nl`

These were selected from roughly the same period and topics as the NISV dataset to ensure that links between the collections could be established.
- *Cultural heritage objects* from the Amsterdam Museum (AM)[7]. In DIVEplus we ingested 3500 images representative of the period 1950-1980.

Additionally, in the DIVEplus 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 and Geo vocabulary. 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.

Another aspect of the DIVEplus methodology is the hybrid workflow for metadata enrichment. We bring together machines and crowds in a collaborative process of extracting relevant events and event-related entities. On the machine side, we use various Named Entity Recognition tools to extract a set of relevant concepts from our collection objects. In order to achieve a wide coverage of extracted entities we reject the notion of majority vote and focus on harnessing the disagreement between different extractors. In this way, we optimize the results in our heterogeneous collection by capturing also less popular results from each extractor and increasing the entity recall.

However, the number of recognized events and links between them in this automatic machine enrichment step is still quite low. Thus, these machine results are further ingested in crowdsourcing tasks to further curate and extend the objects' interpretation space. Here, we apply the CrowdTruth methodology [1] to refine the results of the machine NER with additional event expressions, discover new events and create links between all the detected events and their participating entities. We used the CrowdTruth platform[8] in order to perform all the crowdsourcing steps and experiments.

3. RESULTS

The results from the different NER tools and the crowdsourcing are consolidated to RDF and provided in the DIVEplus RDF Triple store with a SPARQL endpoint[9]. The basic DIVEplus data modeling rationale follows the Simple Event Model (SEM) [5], which allows for the representation of events, actors, locations and temporal descriptions. 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, including Wikipedia and DBpedia. Figure 1 shows the general DIVEplus setup, the data ingestion and enrichment, as well as the interface layers.

In the current triple store we host over 2 Million triples for the 5,000 Media Objects. These are annotated with 17,209 annotations to places, 5,044 actor-annotations and 2,992 separate event-annotations. 8,419 corresponce-triples (`skos:exactMatch`) are established between these entities, tying together the different vocabularies.

We performed user studies to test the usability of the DIVEplus approach with digital humanities professionals and students. For future, we plan to perform more user studies within CLARIAH[10]. In this way we will be able to gain insights on the scalability, robustness and reusability of the DIVEplus digital hermeneutics approach, as well as its usability for history researchers and general audience.

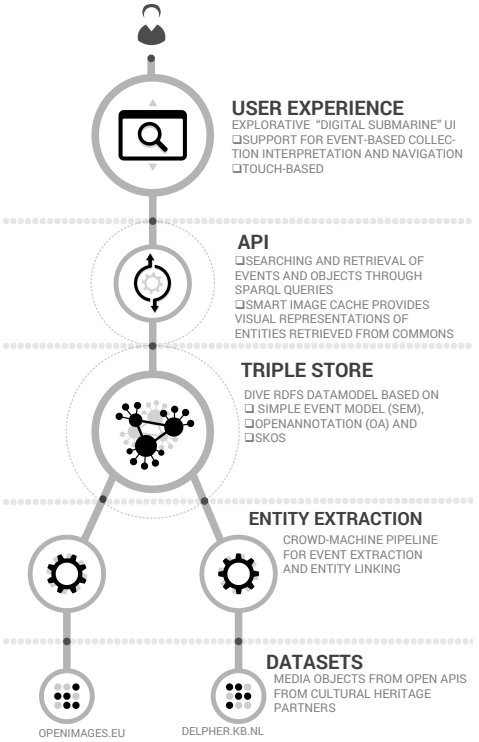

Figure 1: DIVEplus demonstrator: cultural heritage content is collected and enriched through the event generators. The resulting graph representation is stored in a triple store, which is queried by the Web interface layers.

4. REFERENCES

[1] L. Aroyo and C. Welty. The Three Sides of CrowdTruth. *Journal of Human Computation*, 1:31–34, 2014.

[2] V. de Boer et al. Dive into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on the WWW*, 35(3):152–158, 2015.

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

[4] 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.

[5] 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.

[7] `https://www.amsterdammuseum.nl`

[8] `http://crowdtruth.org/`

[9] ClioPatria triple store: `data.dive.beeldengeluid.nl`

[10] `http://www.clariah.nl/en/`