

# Linking Dutch World War II Cultural Heritage Collections with Events Extracted by Machines and Crowds

Jessie Both

University of Amsterdam  
Amsterdam, The Netherlands  
jessieboth@gmail.com

Didi de Hooge

University of Amsterdam  
Amsterdam, The Netherlands  
dididehooge@gmail.com

Ramses IJff

Vrije Universiteit Amsterdam  
Amsterdam, The Netherlands  
ramses.ijff@gmail.com

Oana Inel

Vrije Universiteit Amsterdam  
Amsterdam, The Netherlands  
oana.inel@vu.nl

Victor de Boer

Vrije Universiteit Amsterdam  
Amsterdam, The Netherlands  
v.de.boer@vu.nl

Lora Aroyo

Vrije Universiteit Amsterdam  
Amsterdam, The Netherlands  
lora.aroyo@vu.nl

## 1 INTRODUCTION

The digital revolution is changing the way people access and use knowledge and information. By digitizing collections, cultural heritage institutions aim to fulfill their new role as information providers. However, the majority of the digitized collections lack sufficient metadata and provide no context. This makes the collections hard to access and to search through. Prior research [3] shows that enriching metadata of digitized collection items with events helps with providing context, structure collections, and describe relationships between collection items.

In general, collections can be significantly enriched through the addition of events. Events can place collection items into context and create structure by linking related items. This improves the accessibility of the items and helps users to understand their context. Both personal events and historical events are relevant for this cause. Personal events can provide an understanding of the day-to-day lives of people involved in historical events which can further enable large scale data research. However, detecting and extracting such events is a difficult task, as events are very ambiguous. They lack conclusive definition and their position as events is frequently dependent on personal interpretation.

In this paper, we create a pipeline that detects, extracts and represents personal and historical events to improve the accessibility of cultural heritage items. We show the added value of enriching the metadata of media objects through events and links between them in the context of DIVE+ [1], a Link Data event-centric collections

browser. This research is done on behalf of the Institute for War, Holocaust and Genocide Studies<sup>1</sup> (NIOD).

## 2 DATASET

For the current experiment we selected Wikipedia pages related to World War II (WWII), *i.e.*, "Lijsten over de Tweede Wereldoorlog" and "Tweede Wereldoorlog in Nederland". We focused on three themes: deportation, arrest and resistance. This resulted in a corpus of 2.556 pages.

**Person vocabulary:** For this research we also used a person vocabulary in order to understand which are the most relevant people in our dataset. The person vocabulary consists of data from Online Begraafplaatsen, Erelijst, and the Oranjehotel Kartotheek, filtered based on our three research themes. Our person vocabulary consists of 35,739 names.

**Personal and Historical Events:** The NIOD thesaurus is used as a source of *named historical events* in our dataset. From the metadata of the person vocabulary aforementioned we identified four *personal events*: birth, death, arrival at the Oranjehotel and departure from the Oranjehotel.

## 3 EXPERIMENTAL METHODOLOGY

This section describes our experimental methodology.

### 3.1 Automatic Entity Extraction

The first step in the pipeline is to detect and extract entities through machines. The input data was processed by two pipelines. The first pipeline uses three free NER tools: DBpediaSpotlight<sup>2</sup>, Targeted Hypernym Discovery<sup>3</sup> (THD) and SemiTags<sup>4</sup>. DBpediaSpotlight extracts DBpedia mentions in texts. The THD extractor uses lexicosyntactic patterns to find hypernyms in the text. The SemiTags extractor is an online tool that recognizes named entities and their meaning in a particular context. As this pipeline is specialized in extracting named entities, it is likely to provide historical named events. The second pipeline that is used is the pipeline of BiographyNet [2], a project that uses NLP and Semantic Web Technology to support digital humanities scholars with historical research. The

<sup>1</sup><http://www.niod.nl>

<sup>2</sup><http://www.dbpedia-spotlight.org>

<sup>3</sup><http://entityclassifier.eu/thd/>

<sup>4</sup><http://nlp.vse.cz/SemiTags/>

pipeline is a supervised machine learning system trained to extract biographical metadata such as date and place of birth and death, education, occupation, religion, and parents. This pipeline is more likely to provide personal events.

### 3.2 Crowdsourcing experiments

To improve and enrich the results from the NLP tools described in Section 3.1, we created two crowdsourcing tasks on CrowdFlower. The aim of the first task is to validate and correct entities extracted by the two NLP pipelines and annotate entities that were missed by them. As a proof of concept, for this task we select 100 sentences from the Wikipedia pages. The aim of the second task is to create links between the events and the people, places and time periods identified by the two NLP pipelines and the crowd. For each task we gathered judgments from 15 workers. The task showed one row (sentence) per page and workers were paid 2 cents for one row. The task was only distributed in the Netherlands and Belgium because of language constraints.

### 3.3 Modeling events with Simple Event Model

To model the events, actors, places and time periods extracted by the machines and the crowd we use the Simple Event Model (SEM). All data in the dataset is being converted to the RDF model of the SEM classes and relationships. We create a first graph by using the output of the second crowdsourcing experiment which is converted to RDF. We create a second graph for the people vocabulary.

The WWII thesaurus was already available in RDF, in SKOS. This format is also suitable for SEM, and therefore does not have to be converted. Only concepts which are classified as ‘gebeurtenis’ in the thesaurus are used, plus one manually added named event. To enhance the model, we conduct some alignment proceedings: for the core class place, we automatically generated "sameAs" relationships between identical places in the different graphs. For the core class actor, this can not be done automatically as an identical name does not necessarily address the same person. Therefore, these "sameAs" relationships are done manually.

## 4 RESULTS

This section describes the results for each step of the experimental methodology.

*Automatic Entity Extraction:* Table 1 shows the results of the two pipelines. The THD extractor extracted 145,617 unique entities, mostly locations months, out of which 6,675 entities are events. However, when analyzing the results from THD there is a lot of noise, e.g., "handen namen" is classified as the Belgian city "Namen". DBpediaSpotlight extracted 115,984 unique entities, with a total number of 2,518 entity types (34,785 entities were extracted without type). While performing well on recognizing locations and persons, DBpediaSpotlight also extracted 3,434 events. The SemiTags extractor extracted 99,170 unique entities, but no entity was classified as event. With the BiographyNet pipeline we extracted a number of 260,244 unique entities. The BiographyNet pipeline is especially good at recognizing events, names and precise dates.

*Crowdsourcing Results:* The 100 sentences ran through the crowdsourcing pipeline had 431 entities automatically extracted: 132

events, 74 dates, 128 locations and 97 actors. In the first crowdsourcing task the workers annotated 167 extra entities and removed 34 entities. This resulted in a total of 564 entities, consisting of 102 dates, 228 events, 125 places and 61 actors. For the second crowdsourcing task we used 71 sentences to connect events and places, 72 sentences to connect events and actors and 77 sentences to connect events and time periods. From these sentences, a total of 94 events were linked to 58 places, 61 actors and 64 time periods.

*Event Modeling:* After we converted all the data into SEM classes we obtained a knowledge graph with 209,183 triples. After generating "owl:sameAs" relations between names and locations, the knowledge graph extended to 209,245 triples. Looking at the four core classes in SEM, our enrichment contains 18,316 personal and historical event triples, 9,300 actor triples, 875 place triples and 60 time periods, as shown in Table 2.

## 5 ILLUSTRATIVE EXAMPLE

To illustrate the effectiveness of SEM for our dataset, this section describes an illustrating example through the interface of the DIVE+ demonstrator. When searching for the historical event "februaristaking" in DIVE+ all related events, objects, locations and actors are shown (see Figure 1). When looking at one of the objects we see all related entities to that specific object. For example, we can further explore the entity Communistische Partij Nederland (CPN), Figure 2, which is an important concept for the February Strike. As objects are linked to each other we make the objects easier accessible and create meaningful relations between them. With the use of the event model we tell stories around and about the object and the scholars can get a more complete overview of the event.

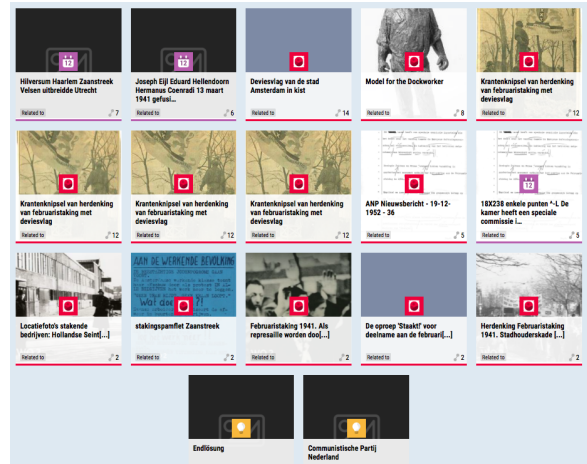


Figure 1: Media Objects, Events, Actors and Places related to 'Februaristaking'

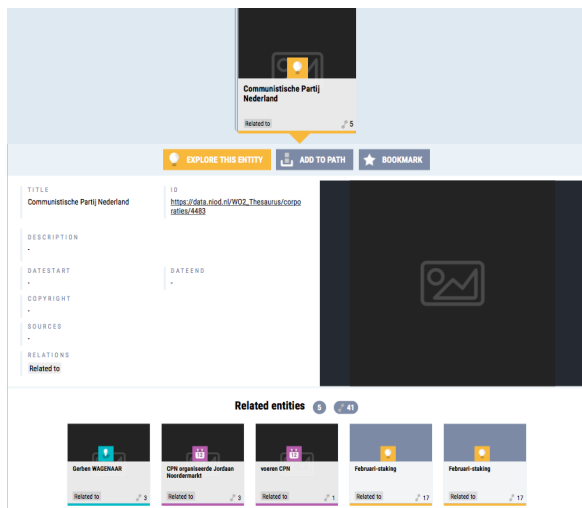
For the personal event, we focus on Eduard Popko van Groningen. Without the model, information about him is fragmented over the collections of Ereljst and Online Begraafplaatsen. All this information is now available in DIVE+ through one query. Figure 3 shows that the query leads to an actor, which is related to five events. The first four are the events of birth and death as described

**Table 1: Overview of Entities Extracted by each Tool**

Tool	#Pages Processed	#Entities Extracted	% of Unique Entities	Avg. # Entities per Page	#Entity Types	# Entities without Type
THD	2,665	145,617	73.50%	64.64	2,517	0
DBpediaSpotlight	2,665	115,984	37.69%	43.52	193	34,785
SemiTags	2,665	99,170	31.37%	37.21	4	0
BiographyNet	2,308	349,321	32.62%	151.35	11	0

**Table 2: Triples Overview per SEM Core Classes**

SEM core class	Triples
Event	18,317
Actor	9,300
Place	875
Time	60

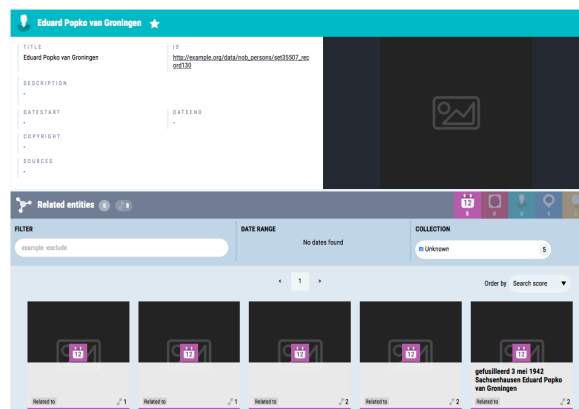


**Figure 2: Exploration of related concept 'CPN'**

in their original source. The fifth event is an event extracted from a Wikipedia page. The related entities are, in turn, also related to other entities. In the case of the extracted event, to the actor, Eduard, and a location, Sachsenhausen. New opportunities for digital humanities scholars appear through the entity 'Sachsenhausen'. The location Sachsenhausen is linked to 21 events which are all connected with other actors. This means that potential fellow prisoners of Eduard are known. Another storyline could be that the actors of the other 21 events were also shot at the 3rd of May 1942, which would indicate that a bigger event took place.

## 6 CONCLUSION AND FUTURE WORK

Digitization projects made various cultural heritage collections publicly available online. However, due to unstructured metadata and the distributed allocation of the objects in different databases, purely digitizing objects does not make them more accessible. In this research we used a hybrid machine-crowd approach to extract personal and historical events and use them for the enrichment of



**Figure 3: Events related to 'Eduard Popko van Groningen'**

the Dutch digital WWII objects metadata within the NOB portal. We used automatic event extraction, making use of different information extracting tools and manual event curation through the use of a crowdsourcing strategy. Applied our methodology a set of Wikipedia pages related to WWII. We illustrated the historical and personal events in the DIVE+ demonstrator. With our created historical and personal event model we are able to enrich the metadata of the objects in the NOB portal. In this way the available digital resources of the NOB portal can be linked and improved with storylines told about and around the objects.

Future work can improve the pipeline. For example, the entity extraction could be improved or tested with other NLP tools better suited for Dutch texts. Another way to improve entity extraction is by tweaking the crowdsourcing tasks. Referencing words, such as 'he' and 'that' were not extracted as the context around the sentence is missing. As a consequence, a lot of verbs could not be linked with an actor. We recommend doing a pilot with extracting entities from paragraphs to improve results. Lastly, we suggest enriching the SEM with adding subtypes to the core classes actor, event, place and time. For example, nationalities can be added to names of people for a richer context.

## REFERENCES

- [1] Victor De Boer et al. 2015. DIVE into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on WWW* 35 (2015), 152–158.
- [2] Antske Fokkens, Serge Ter Braake, Niels Ockeloën, Piek Vossen, Susan Legègne, and Guus Schreiber. 2014. BiographyNet: Methodological Issues when NLP supports historical research.. In *LREC*.
- [3] Walter Renteria-Agualimpia, Francisco J Lopez-Pellicer, Javier Lacasta, F Javier Zarazaga-Soria, and Pedro R Muro-Medrano. 2016. Improving the geospatial consistency of digital libraries metadata. *Journal of Information Science* 42, 4 (2016), 507–523.