Enriching Media Collections for Event-based Exploration

Victor de Boer[14], Liliana Melgar[24], Oana Inel[1], Carlos Martinez Ortiz[3], Lora Aroyo[1], and Johan Oomen[4]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands {v.de.boer, oana.inel, lora.aroyo}@vu.nl
[2] Universiteit van Amsterdam, Amsterdam, the Netherlands, melgar@uva.nl
[3] eScience Center, Amsterdam, the Netherlands, c.martinez@esciencecenter.nl
[4] Netherlands Institute for Sound and Vision, Hilversum, the Netherlands, joomen@beeldengeluid.nl

Abstract. Scholars currently have access to large heterogeneous media collections on the Web, which they use as sources for their research. Exploration of such collections is an important part in their research, where scholars make sense of these heterogeneous datasets. Knowledge graphs which relate media objects, people and places with *historical events* can provide a valuable structure for more meaningful and serendipitous browsing. Based on extensive requirements analysis done with historians and media scholars, we present a methodology to publish, represent, enrich, and link heritage collections so that they can be explored by domain expert users. We present four methods to derive events from media object descriptions. We also present a case study where four datasets with mixed media types are made accessible to scholars and describe the building blocks for event-based *proto-narratives* in the knowledge graph.

1 Introduction

With the recent increase in availability of digital data relevant to humanities researchers, the term *digital humanities* is used to indicate the increased role that digital archives and computational tools play in scholarship. Different tools that cater to the different information needs of these scholars are often associated with specific phases in their research [25,7,24]. Such tools should support exploration of collections in early research phases, but also filtering of material during assembling or building the corpus, and during contextualization and analysis to identify links or commonalities between dataset items. During the process of information extraction and interpretation from such datasets, users identify connections between entities based on their specific domain knowledge. This creates a "narrative chain of events" that explains certain phenomena [27].

In most cases, the information needed to create such narratives are found in multiple collections that are heterogeneous in form and scope. To this end, many data curation efforts have turned to the principles of the Semantic Web and specifically the practice of Linked Data [5] as they provide key technologies

to allow for data integration and re-usability. Linked Data is increasingly used to publish both archival (meta)data and datasets that are the result of curatorial and research activities. Linking such heterogeneous datasets allows for new types of analyses across collections, including event narrative creation.

However, a deeper integration of access to these heterogeneous linked datasets and the real praxis of humanities scholars still remains an open challenge for the field of digital humanities. Datasets often lack the structured metadata needed to identify the necessary links in order to investigate potential building blocks for event narratives. In previous work [2], we introduced the notion of 'proto-narratives', building blocks found in interconnected datasets that can serve as starting points for more elaborate narratives. Such proto-narratives are often based on events. In order to support scholars in making sense of these heterogeneous data sources during their research by creating narratives, we observe that: (1) at the data level, these collections need to be integrated in a common semantic data model and enriched with event information, and (2) at the functionality level, scholars should be able to create their own links between entities and events by using navigation paths and annotations during their browsing and search activities. In this paper we focus on the first aspect, by investigating how we can use linked data principles to publish, enrich and connect heterogeneous datasets. Specifically, we describe:

- A simple and generic data model for connecting heterogeneous media datasets.
- A variety of methods to enrich media collections in such a way that they can be explored by cultural heritage scholars. These focus on identifying *events* and connecting them through shared persons, places and concepts.
- A case study in the context of the DIVE+[1] project where four heterogeneous datasets are enriched and interlinked using the proposed strategies.

2 Existing Data Enrichment Techniques and Modelling

Metadata schemas and vocabularies are key technologies for improving the access to cultural heritage collections and the cross-walks among their objects [4]. In this section we present an overview of existing data enrichment techniques that are further adapted and extended in our method (Section 4).

Machine Enrichment includes multi-modal information extraction techniques from free text, but also image or video content retrieval technologies. Specifically, Named Entity Recognition (NER) tools (for a review see [16]) typically extract persons and organizations (actors), places, and to some extent temporal definitions from the texts. Such tools can be used to extract entities from textual transcriptions of media objects (for example OCR'ed text for textual objects or subtitles for videos) or from descriptive metadata fields. Image analysis methods such as [28] or video analysis tools can be used to identify entities of different types in the visual content of media objects. Speech or speaker recognition can furthermore be used to identify entities in audio content.

[1] http://diveplus.beeldengeluid.nl

While NER tools extract named entities such as actors and places with high accuracy, their performance in detecting events is still poor [16,20]. However, Natural Language Processing (NLP) techniques proved to be successful in event extraction tasks for specific domains such as the Biomedical domain [22]. More recently, advances have been made in the cultural heritage domain [26] as well. For the Dutch language, the FROG NLP suite [6] includes functionalities to identify Named Events in Dutch language texts. Such Named Events are denoted with proper names, in contrast to other pipelines that extract parts of sentences based on object-verb occurrences.

Human Computation includes crowdsourcing, but also smaller-scale annotation efforts by experts, games-with-a-purpose or nichesourcing [13]. The common factor here is that for specific enrichment tasks, human annotators outperform machines. This is the case for hard-to-extract entities, but it also depends on the quality of the source material. For example, if OCR'ed text is of low quality, off-the-shelf NER tools have a hard time accurately identifying entities [17], while humans are able to deal much better with the textual errors. Furthermore, since events are difficult to extract by machines, various crowdsourcing approaches have been defined [9,23].

Hybrid Methods combine both machine enrichment and human computation for effective and efficient optimization. Current research [21] showed that the performance of NER tools can be improved by allowing crowd workers to validate and correct their output. Furthermore, hybrid methods have been used for solving complex tasks of linking events with their participating entities [9].

Reusing event metadata Though not an enrichment technique, a source for events can be existing structured metadata. Some metadata standards do have explicit modelling of events. For example, the LIDO metadata schema [8] allows for representing events as metadata for museum objects. The CIDOC-CRM metadata schema [14] or the Europeana Data Model [15] allow for event-centric modelling of cultural heritage content.

3 The Need for Event-based Exploration of Collections

During the "Agora" project, historians and computer scientists laid the basis for the concept of *digital hermeneutics*, which couples exploration of cultural heritage collections via browsing tools with the interpretation needs of historians. Essential to *digital hermeneutics* is the concept of *event*, which is the building block of the interpretation process. Events add context to information objects in collections since they consist of the related entities extracted from these media objects (e.g., persons, locations, concepts). When two or more events are related to each other, they start to compose the so-called *proto-narrative*, in which relations can be observed between actors (e.g., F. de Casembroot, in a biographical proto-narrative), types (e.g., battles, in a conceptual proto-narrative), or places (e.g., Shimonoseki, in a topological proto-narrative) [2]. In this way, events are able to connect data and describe the relation between historical events and digital resources or objects [29].

The core of digital hermeneutics is formed by two components: *object-event relationships* and *event-event relationships*. By making explicit relationships between the objects and events, and between the events themselves we can facilitate users in their access and interpretation processes (i.e., in creating narratives) based on objects in online cultural heritage collections.

Several user studies with historians and other humanities scholars have shown the need for events in supporting meaningful browsing of cultural heritage collections, for instance, [1]. Recent studies in the context of the CLARIAH project[2] have also shown that media scholars and other scholars using mixed-media collections require to annotate their sources, which consists in great part of entity identification and linking [24]. Media scholars in these projects have also shown the need to identify media events and their different types (e.g., disruptive media events) [19]. While for concepts, places and persons structured data is often available, for events this is less the case. For this reason, we introduce our method for media data enrichment and event extraction in the next section.

4 A Method for Media Data Enrichment

In this section we present our generic method for connecting heterogeneous media collections, for enabling the explorative search as described above. Each subsection describes a specific step. In Section 5, we detail these steps for our specific case study in the DIVE+ demonstrator.

4.1 Collections and Vocabularies

We assume as input a number of collections consisting of Media Objects, curated by the providing institution. These can be text, audio, image, video, or multimedia. Furthermore, we assume that descriptive metadata is available in RDF form or that it can be converted to this format so that we have at least syntactic interoperability in this common data format. The metadata can be textual (descriptions, titles, among others) or other literal values (e.g., numbers) or it can contain values from controlled vocabularies. These can be either in-house vocabularies or can refer to external vocabularies. In these cases, we assume that these vocabularies are available and can be imported in the common framework.

4.2 Mapping to Generic Schema

The collection RDF metadata will be defined according to an (RDFS) metadata schema, listing the properties and classes used in that metadata. In order to link the schema-level information of the heterogeneous collections, we establish sub-property relations between the individual collections' properties and that of a generic schema and we do the same for the classes. Because of these relations, queries at the level of the generic schema can use information declared at the level of the individual collections, as outlined in [12].

[2] https://www.clariah.nl/en/

This generic schema describes basic properties and classes that are to be used for the type of hermeneutical exploration we described above. It should at least contain the classes **Media Object**, **Person**, **Place**, **Concept** and importantly **Event** and properties relating these to each other. Furthermore, it should contain descriptive metadata properties for textual metadata about the media objects (title, description, dates, among others).

We base our generic model on the Simple Event Model (SEM) [18]. This model allows for the representation of events, actors, locations and temporal descriptions. One of its features is that it is a very basic event-centric model, but more complex relations between, for example, events and persons (such as the role that a person plays in an event) can also be expressed. We select this model over other event-models such as the aforementioned CIDOC-CRM or LODE because of its relative small size, flexibility and low ontological commitment, allowing for easy mapping of data with various heterogeneous data models. Furthermore, SEM includes explicit mappings to these models, making it easy to interpret the data integrated at the SEM level using CIDOC or LODE tooling. For a detailed discussion comparing SEM to other event models, we refer the reader to [18]. We furthermore use SKOS[3] to represent concepts from vocabularies and DCTerms[4] for the descriptive metadata of the objects. In order to visualize media objects in a tool, we define two relations to web-accessible resources: one for image thumbnails, one for large-sized image or video. We define additional generic relations (`isRelatedTo`) between these persons, places, concepts and objects and extend it with specific relations to relate media objects to entities depicted in the objects (`depicts` and `depictedBy`). Finally, to allow for metadata about user annotations, we use the Web Annotation datamodel[5]. Figure 1 shows the classes in the datamodel as well as properties that hold between them.

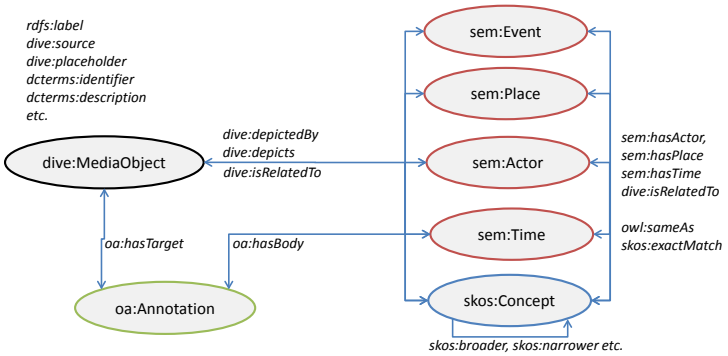

Fig. 1. Graphical representation of the generic data model.

[3] https://www.w3.org/2004/02/skos/
[4] http://dublincore.org/documents/dcmi-terms/
[5] https://www.w3.org/TR/annotation-model/

4.3 Constructing a Knowledge Graph

Schema-level mappings as described above are not sufficient to establish an explorable graph. For this, we need a) instances of person, place, concept and event classes for media objects and b) links between these instances.

Enrichment: Persons, Places, Concepts. Most collections come with structured metadata which includes persons, places and concepts related to collection items. In these cases, the values for these metadata fields are either literal values or RDF Resources from controlled vocabularies. In the first case, to allow for establishing links, these literal values can be converted to RDF resources which have as a label the original literal value. The resource type is mapped to one of the data model classes. This 'promotion' to resource results in a simple vocabulary. In the latter case, the vocabulary is simply loaded together with the object data. In cases where one or more entity types are missing from the original metadata (for example, no Persons, no Places), extra enrichment is needed. Here, we can build on existing strategies as listed in Section 2.

Enrichment: Events. As events play a key role for connecting media objects into larger storylines or proto-narratives, extraction of such events is key. For some objects, we can assume one or more events are represented in the content or in the metadata. One way these can occur is through properties denoting events, including object creation timestamps or even start- and end-times. For example in the MADS metadata schema[6], the property *creationDateStart* indicates such a creation event. These implicit events can also be 'promoted' to explicit events. A second type of interpreted event is when we can infer from the type of media object that an event is present. For example, in cases of news media objects, we can infer that a news event is described and therefore create an event resource associated with the media object itself. In Section 5, we describe how this is done for radio news bulletins. Finally, for cases where events are already indicated in the metadata, we can directly map them to SEM constructs, resulting in the required event entities. For media objects that lack such identifiable events, we apply human computation and NLP techniques as detailed in Section 5.

Hybrid pipeline. In our method, we also combine different strategies into a hybrid pipeline where machines and crowds collaborate in order to extract events and event related entities and find links between them [9]. First, we use various NER tools to extract the set of relevant concepts from object description. We reject the notion of majority vote and focus on harnessing the disagreement between different extractors to achieve a more diverse set of entities. In short, we preserve all the entities extracted by every NER tool, disregarding the number of tools that extracted them. Each media object description is then used in a crowdsourcing task where the crowd is asked to highlight events mentioned in the text. The NER and crowd output are then aggregated into a second crowdsourcing task that aims to create links between the detected events and their participating entities. We use the CrowdTruth platform[7] and methodology [3] in order to perform all the crowdsourcing steps and experiments.

[6] http://lov.okfn.org/dataset/lov/vocabs/mads
[7] http://crowdtruth.org/

Establishing Links. With structured metadata, including persons, places, concepts and events established, there still remains the task of interlinking these entities across the different collections and vocabularies. In some cases, existing alignments between controlled vocabularies exist, where in other cases -especially for the controlled vocabularies constructed in the process of enrichment- these need to be established. Different methods and tools for such alignment exist and include fully automatic tools. We mainly employ a transparent, interactive tool, CultuurLink[8], which allows collection managers and other experts to combine various string matching algorithms to build alignment strategies [10].

5 Case Study: Data Enrichments in the Dive+ Project

In this section we present a case study concerning the DIVE+ platform where we employ and validate the method described above. DIVE+[9] builds on research [2,9] supporting *digital hermeneutics* for historians and media scholars, through connected heterogeneous datasets and vocabularies. DIVE+ uses historical events and event narratives as context for searching, browsing and presenting cultural heritage collection objects. The interface (Fig. 2) combines Web technology and theory of interpretation to allow for browsing this knowledge graph.

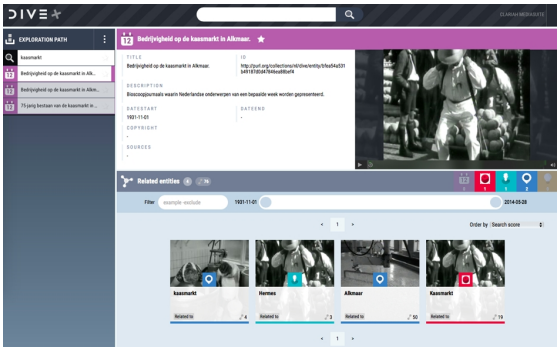

Fig. 2. Screenshot showing the current UI of the DIVE+ browser where an event (top) as well as related entities (bottom) are shown.

5.1 Four Datasets

We first list the datasets and their accompanying vocabularies. These openly licensed collections are curated by Dutch cultural heritage institutions.

OpenImages.eu broadcasts (OI). This collection is archived by the Netherlands Institute for Sound and Vision[10]. It consists of 3,220 videos published on the Openimages platform[11] from the period 1920-1980. Most of these videos are

[8] http://cultuurlink.beeldengeluid.nl
[9] http://diveproject.beeldengeluid.nl
[10] http://www.beeldengeluid.nl
[11] http://openimages.eu

news items originally shown in movie theaters. For the structured metadata, the thesaurus Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA)[12] is used containing 160,000 concepts, places and persons, but no events.

ANP Radio News Bulletins (NB). These are scans of typoscripts which were read aloud on broadcast radio in the period 1937-1984 and are now archived by the Dutch National Library[13]. In DIVE+ we ingested 197,199 digitized typoscripts with their OCR'ed content. The original metadata only contains dates and administrative metadata, but no persons, places, concepts or events. Where the other datasets were available as RDF, for NB we did a straightforward conversion to RDF from the XML output of the OAI-PMH API[14].

Amsterdam Museum collection (AM). This concerns an RDF version of the collection metadata of Amsterdam Museum [11]. It contains 73,447 cultural heritage objects[15]. The collection is described using the Amsterdam Museum Thesaurus, containing some 28,000 concepts, persons and places. It also includes 148 Events. The thesaurus was previously partially aligned with GTAA.

Tropenmuseum collection (TM). This consists of an RDF version of 78,270 cultural heritage objects from the collection of the Tropenmuseum[16] related to ethnological research. The collection is described using the SVCN thesaurus[17], which contains 3,896 places and 13,269 content subjects and was also previously aligned with GTAA. Again, in this thesaurus there are no events identified.

5.2 Mapping Schema

We established schema mapping files for each of the four datasets by inspecting the properties and classes used in those datasets and writing sub-property or sub-class links to the common DIVE+ schema as RDF triples. Very few of such links are required: for the three datasets OI, AM, TM, we respectively had to add 3, 12 and 18 sub-property or sub-class triples. For NB no extra triples were needed as these were converted in the scope of the project[18].

5.3 Enrichment and Linking

We here describe for each dataset the types of enrichment performed.

OI enrichment. For OI news videos, some structured metadata in the form of subject relations to GTAA terms existed. However, these do not include events

[12] http://gtaa.beeldengeluid.nl
[13] http://radiobulletins.delpher.nl/
[14] This conversion code is available at https://github.com/biktorrr/dive/
[15] https://www.amsterdammuseum.nl
[16] https://tropenmuseum.nl/en
[17] http://svcn.nl
[18] http://data.dive.beeldengeluid.nl/browse/list_triples?graph=http%3A//purl.org/collections/nl/am/am_additions.ttl shows the 12 triples added for Amsterdam Museum. These include mappings of object-image relations, object-entity relations as well as object classes.

and persons. For these entities, we deployed the hybrid pipeline as described in Section 4.3. NER and Event extraction tools for Dutch text are used including the xTas[19] and Opener[20] toolkits. In a second stage, crowdsourcing through the CrowdTruth platform is employed to have human-recognized entities and to refine the results from Natural Language Processing. For the extraction of the News Bulletins, we also use the results of the NER employed by the KB, which is based on the Stanford parser, optimized for Dutch texts. This results in a total of 11,474 individual annotations provided by the human annotators. These include 1,916 unique Events, 1,249 Actors, 1,412 Places and 162 Time annotations. We aligned the newly generated related entities to places, persons and concepts in the GTAA. For this, we used the CultuurLink alignment tool to develop a simple pipeline based on basic string matching. In total, we aligned 452 Places, 171 Actors and 117 Concepts with GTAA.

NB enrichment. To enrich the Radio News bulletins metadata, we used the original publishers' own enrichment services as described in [30], which include NER. At data-conversion time, for each record, we retrieve the content as well as the Named Entities (Person, Place, Organization, Unknown) which were then mapped to DIVE classes. This resulted in 54,571 Places, 197,200 Actors and 6,736 SKOS Concepts. Events are not returned by the NER module, however, here we can use our strategy of interpreting events. As every news bulletin object is actually a description of a newsworthy event, we can deduce that there is at least one event described by the media object. We therefore generate one Event object for each Media Object. The label of this Event object is derived from the OCR'ed description of the Bulletin using simple heuristics[21]. This generated event is then also related to the Actors, Places and Concepts that have been identified in the NER process. The Entities are aligned with GTAA to establish correspondences between the NB and OI datasets. Again, we used the CultuurLink tool to establish a simple string-matching based pipeline. This results in 3,223 Places and 3,130 Actor matches.

AM enrichment. For Amsterdam Museum, we already have identified Persons, Places and Events in the collection metadata. We therefore suffice by aligning the thesaurus terms to GTAA. We here reuse alignments between the AM thesaurus and GTAA that were established in [11]. In total this alignment contains 1,500 Place matches 5,301 Actor matches and 64 Concept matches.

TM enrichment. In the Tropenmuseum data and SVCN thesaurus, we find no Events. However, upon inspection, some of the item descriptions contain mentions of events. To identify these, we use the FROG Event extraction to extract named events from those descriptions. This results in 115 relations between Media Objects and Events. In total we find 16 of such Named Events (for example "Day of the Dead" or "Second South-New Guinea Expedition"). These events, as well as the places and concepts in SVCN are aligned with GTAA, again using the CultuurLink tool, resulting in 2,573 alignments.

[19] http://xtas.net

[20] http://www.opener-project.eu/

[21] http://tinyurl.com/diveplusexample2 shows an example event in the DIVE+ UI

5.4 Results

These enrichment strategies result in a large knowledge graph. Table 1 summarizes the results of the enrichment process for all datasets as well as the total. Table 2 shows the total number of links between different entity types and events in the whole dataset. These correspond to the building blocks for the exploratory browsing and to the establishment of proto-narratives. The data is stored in a public RDF Triple store[22] and is available at a public GIT repository[23].

Table 1. Number of objects resulting from data conversion and enrichment

	Enrichment method	Media Objects	Actors	Places	Events	Other
OI	Hybrid pipeline	3,204	1,249	1,412	1,916	185,846
NB	Interpreted + NER	197,200	194,890	54,571	197,200	6,736
AM	Original thesaurus	73,447	66,966	5,973	148	28,047
TM	Original thesaurus + NER	78,226	27,829	3,896	16	13,269
Total		352,077	290,934	65,852	199,264	233,898

Table 2. Links statistics in the total knowledge graph

Subject-Object	property supertype	count
Media Object-Event	`dive:depictedBy` or `dive:isRelatedTo`	199,233
Event-Actor	`sem:hasActor`	265,677
Event-Place	`sem:hasPlace`	220,726
Event-Concept	`dive:isRelatedTo`	230

6 Discussion and Future Work

In this paper, we showed the need for integrating heterogeneous media collections from a user perspective. To be able to deliver a usable exploratory functionality to our end-users, original collections need to be not only converted, but also enriched with structured metadata. Where existing methods often focus on persons, places, and concepts, we emphasize extraction of events and provided a number of methods to do this. By linking events with objects, persons and places, an interconnected knowledge graph is constructed, which has the required characteristics. Different errors or conflicting extractions can occur both based on automatic extraction, human intervention and in the hybrid method. A discussion on appropriate quality measures and the use of these measures to determine link quality is out of the scope for this paper. However, current work on this is to build on current work on harnessing disagreement [3].

The methodology described in this paper can be used for any heterogeneous cultural heritage linked data collection. We show how we validated this methodology in a specific use case, where we present enrichment statistics as well as explorable paths between entity types. Next steps include evaluation of these paths.

[22] The triple store can be accessed at http://data.dive.beeldengeluid.nl/
[23] https://github.com/biktorrr/diveplusdata/

For this, we are currently integrating annotation and improvement functionalities in the DIVE+ UI based on continuous user studies with scholars. This will allow for in-browsing crowdsourcing of annotations and corrections, and will provide a way to continuously update and upgrade our knowledge graph. The ability to develop storylines from this interconnected knowledge graph during exploratory search will be evaluated and mediated through the DIVE+ UI. Therefore, this also becomes another enrichment method as part of our methodology.

Acknowledgements This work was partially supported by CLARIAH (http://clariah.nl/) and by the Netherlands eScience Center (http://esciencecenter.nl/) DIVE+ project. We furthermore thank Victor Kramer, Jaap Blom and Werner Helmich.

References

1. van den Akker, C., van Nuland, A., van der Meij, L., van Erp, M., Legne, S., Aroyo, L., Schreiber, G.: From Information Delivery to Interpretation Support: Evaluating Cultural Heritage Access on the Web. In: Proceedings of the 5th Annual ACM Web Science Conference. pp. 431–440. WebSci '13, ACM, New York, NY, USA (2013)
2. Akker, C.v.d., Legêne, S., Erp, M.v., Aroyo, L., Segers, R., Meij, L.v.D., Ossenbruggen Van, J., Schreiber, G., Wielinga, B., Oomen, J., et al.: Digital hermeneutics: Agora and the online understanding of cultural heritage. In: Proceedings of the 3rd International Web Science Conference. p. 10. ACM (2011)
3. Aroyo, L., Welty, C.: The Three Sides of CrowdTruth. Journal of Human Computation 1, 31–34 (2014)
4. Baca, M.: Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. Cataloging & classification quarterly 36(3-4), 47–55 (2003)
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
6. Bosch, A.v.d., Busser, B., Canisius, S., Daelemans, W.: An efficient memory-based morphosyntactic tagger and parser for dutch. LOT Occasional 7, 191–206 (2007)
7. Bron, M., van Gorp, J., de Rijke, M.: Media studies research in the data-driven age: How research questions evolve. Journal of the Association for Information Science and Technology 67(7), 1535–1554 (2015)
8. Coburn, E., Light, R., McKenna, G., Stein, R., Vitzthum, A.: LIDO-lightweight information describing objects version 1.0. ICOM International Committee of Museums (2010)
9. de Boer, V., Oomen, J., Inel, O., Aroyo, L., van Staveren, E., Helmich, W., de Beurs, D.: DIVE into the event-based browsing of linked historical media. Web Semantics: Science, Services and Agents on the WWW 35, 152–158 (2015)
10. de Boer, V., Priem, M., Hildebrand, M., Verplancke, N., de Vries, A., Oomen, J.: Exploring Audiovisual Archives Through Aligned Thesauri, pp. 211–222 (2016)
11. de Boer, V., Wielemaker, J., van Gent, J., Oosterbroek, M., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Amsterdam museum linked open data. Semantic Web 4(3), 237–243 (2013)
12. de Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: the amsterdam museum case study. In: ESWC. pp. 733–747 (2012)

13. Dijkshoorn, C., Leyssen, M.H., Nottamkandath, A., Oosterman, J., Traub, M.C., Aroyo, L., Bozzon, A., Fokkink, W., Houben, G.J., Hovelmann, H., et al.: Personalized nichesourcing: Acquisition of qualitative annotations from niche communities. In: UMAP Workshops (2013)
14. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine 24(3), 75 (2003)
15. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA general conference and assembly. pp. 10–15 (2010)
16. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Extended Semantic Web Conference. pp. 351–366. Springer (2013)
17. Grover, C., Givon, S., Tobin, R., Ball, J.: Named entity recognition for digitised historical texts. In: LREC (2008)
18. van Hage, W.R., Malais, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the Simple Event Model (SEM). Web Semantics: Science, Services and Agents on the World Wide Web 9(2), 128–136 (2011)
19. Hagedoorn, B., Sauer, S.: Getting the Bigger Picture: Exploratory Search and Narrative Creation for Media Research into Disruptive Events. Utrecht (2017)
20. Hooland, S.v., De Wilde, M., Verborgh, R., Steiner, T., Walle, R.V.d.: Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities 30(2), 262–279 (2013)
21. Inel, O., Aroyo, L.: Harnessing diversity in crowds and machines for better ner performance. In: The Semantic Web. ESWC 2017. pp. 289–304 (2017)
22. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of bionlp'09 shared task on event extraction. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. pp. 1–9. ACL (2009)
23. Lee, K., Artzi, Y., Choi, Y., Zettlemoyer, L.: Event detection and factuality assessment with non-expert supervision. In: EMNLP. pp. 1643–1648 (2015)
24. Melgar Estrada, L., Koolen, M., Huurdeman, H., Blom, J.: A process model of time-based media annotation in a scholarly context. In: ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR). Oslo (2017)
25. Palmer, C.L., Teffeau, L.C., Pirmann, C.M.: Scholarly information practices in the online environment: themes from the literature and implications for library service development. Tech. rep., OCLC Research, Dublin, Ohio (2009)
26. Richards, J.D., Tudhope, D., Vlachidis, A.: Text mining in archaeology: extracting information from archaeological reports. Mathematics and Archaeology p. 240 (2015)
27. Sauer, S., de Rijke, M.: Seeking serendipity: A living lab approach to understanding creative retrieval in broadcast media production. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 989–992. SIGIR '16, ACM, New York, NY, USA (2016)
28. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., et al.: Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator. Web Semantics: Science, Services and Agents on the World Wide Web 6(4), 243–249 (2008)
29. Shaw, R., Troncy, R., Hardman, L.: Lode: Linking open descriptions of events. ASWC 9, 153–167 (2009)
30. van Veen, T., Lonij, J., Faber, W.J.: Linking Named Entities in Dutch Historical Newspapers, pp. 205–210 (2016)