Computational Linguistics Applications for Multimedia Services

Kyeongmin Rim **Kelley Lynch** **James Pustejovsky**

Department of Computer Science
Brandeis University
Waltham MA USA
{krim,kmlynch,jamesp}@brandeis.edu

Abstract

We present Computational Linguistics Applications for Multimedia Services (CLAMS), a platform that provides access to computational content analysis tools for archival multimedia material that appear in different media, such as text, audio, image, and video. The primary goal of CLAMS is: (1) to develop an interchange format between multimodal metadata generation tools to ensure interoperability between tools; (2) to provide users with a portable, user-friendly workflow engine to chain selected tools to extract meaningful analyses; and (3) to create a public software development kit (SDK) for developers that eases deployment of analysis tools within the CLAMS platform. CLAMS is designed to help archives and libraries enrich the metadata associated with their mass-digitized multimedia collections, that would otherwise be largely unsearchable.

1 Introduction and Motivation

Since the invention of the phonograph and moving pictures, audiovisual materials have been one of the primary methods of recording modern history alongside textual records. Many historical events, important persons, social issues, and major conflicts over the last several decades have been recorded on such mass media. Researchers in both media studies and the social sciences, as well as historians have long recognized the value of audio and visual records as evidence about the past (e.g., Boykoff and Boykoff, 2007; Dalton and Charnigo, 2004; Doms and Morin, 2004). Likewise, educators have appreciated the ability of multimedia materials to make history and cultural heritage artifacts come alive in the classroom setting (e.g. Ott and Pozzi, 2011; Antonaci et al., 2013). Recently, with the advent of large digital storage, there have been many large-scale projects aimed at the mass-digitization of books (Christenson, 2011), newspapers (NDNP, 2005), oral history (Oard et al., 2002; NYPL, 2013), and public broadcasting (MDPI, 2014; AAPB, 2015). Selections of results from these projects are publicly available through web-based *digital libraries*, often accompanied by a search interface. However, users of such digital library resources can be frustrated by the difficulties associated with accessing these historical audiovisual records, not because of any lack of accessibility to the digital media themselves, but because of the lack of accessibility to the *contents* of the media (Schaffner, 2009). Audiovisual media, unlike textual records, are opaque to even the simplest text-based search capability. Finding content relevant to one's research question among thousands of hours of audiovisual records, hence, is time-consuming, involving watching or listening to hours of contents. Therefore, a key to making a digital multimedia archive useful and accessible is to generate and deploy rich metadata of collection items (Cariani et al., 2015). The availability of such descriptive, structured, textual metadata about the content of the collections and the included items radically improve the searchability and discoverability of the material (Pustejovsky et al., 2017). Yet, manually cataloging meaningful and suitably robust metadata is a general challenge across digital archives, as it will also be time-consuming and laborious, involving archivists watching and listening to items.

In this paper, we describe the CLAMS[1] platform, developed for libraries and archivists to help enrich item-level descriptive metadata by providing with automatically extracted information from time-based multimedia collections utilizing computational analysis tools for text, audio, and video (Pustejovsky, 2018). These tools for different modalities will be orchestrated via CLAMS work-

[1] http://www.clams.ai

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 91–97
Minneapolis, MN, USA, June 7, 2019.

flow engine that provides a common interchange format ensuring syntactic and semantic interoperability between these tools.

2 Prior Work

Multilingual Access to Large Spoken Archives (MALACH) (Oard et al., 2002) was one of the early studies that used computational linguistics tools to build an automatic metadata extraction system. In MALACH, oral history recording data was processed through automatic speech recognition (ASR) and natural language processing (NLP) pipelines that extracted relevant information for cataloging. In prototyping its World Service Archive (Raimond et al., 2014), the BBC developed COMMA, an metadata extraction and linked data-based interlinking system for public radio broadcasts. Its outcome is now in use by the BBC (BBC, 2015), however it is not publicly available. More recently, the EU funded Media in Context (MiCO) project (Aichroth et al., 2015). This project aimed at accomplishing a media analysis platform for multimodal media that supports customized workflows leveraging on assorted open and closed source content analysis tools. An interoperability layer, MiCO Broker, was developed based on RDF and XML structures to chain different tools. Among the latest work, Audiovisual Metadata Platform (AMP) is noteworthy as it plans to design and develop a platform that exploits chains of automated tools and human-in-the-loop to generate and manage metadata at institutional scale (Dunn et al., 2018). We actively seek collaboration with others in order to move closer to achieving a "global laboratory" for language applications.

In the computational linguistics (CL) community, UIMA (Ferrucci et al., 2009) and GATE (Cunningham et al., 2013) have been long-standing popular tool-chaining platforms for researchers and NLP developers. Particularly, UIMA provides an extremely general model of type systems and annotations that can be applied upon multimedia source data. However, there is stiff learning curve behind its high generality, combined with its tight binding with XML syntax and Java programming language. More recently, web-based workflow engines such as the LAPPS Grid (Ide et al., 2014) and WebLicht (Hinrichs et al., 2010) provide user friendly web interfaces. Particularly, these web-based platforms not only offer tool repositories of various levels of state-of-the-art NLP tools for textual data, such as CoreNLP (Manning et al., 2014), OpenNLP (OpenNLP, 2017), but also implement open source SDK for tool developers to promote adoption. These workflow engines can operate different tools which are separately developed only because of the underlying data interchange formats that impose common I/O language between those tools. For such an interchange format, The LAPPS Grid uses LAPPS Interchange Format (LIF) rooted on JSON-LD serialization (Verhagen et al., 2015), while the WebLicht uses XML-based Text Corpus Format (TCF) (Heid et al., 2010). Additionally the LAPPS Grid defines a semantic linked data vocabulary that ensures semantic interoperability (Ide et al., 2015). Having implemented in-platform interoperability has led to a multi-platform collaboration between LAPPS and CLARIN (Hinrichs et al., 2018).

3 Project Description

Figure 1 shows the overall structure of the platform in a working environment as delivered to an archive. As a platform, the primary goals of CLAMS are 1) to develop an interchange format between multimodal annotations that allows analysis tools for different modalities to work together when chained into a single workflow, and 2) to provide libraries and archivists a portable workflow engine software with a user-friendly interface to select available tools and create workflows and run them, and lastly 3) to offer various analysis tools alongside a public SDK for developers of the tools that allows easy adoption of the interchange format and streamlined deployment to the workflow engine. In the rest of this section, we will discuss how we address each of aforementioned goals.

3.1 Multimodal Interoperability

To implement the platform with interoperating analysis tools, we developed Multi-Media Interchange Format (MMIF) as the *common tongue* of CLAMS. MMIF consists of two parts – it adopts the already successful JSON-LD as syntax, and an open linked data vocabulary for the semantics of the terminology. The vocabulary is re-using the LAPPS Grid vocabulary as its linguistic terminology, while extending it further to cover audiovisual concepts such as `timeFrame`, or `boundingBox`.

Typologically, multimodal annotations in

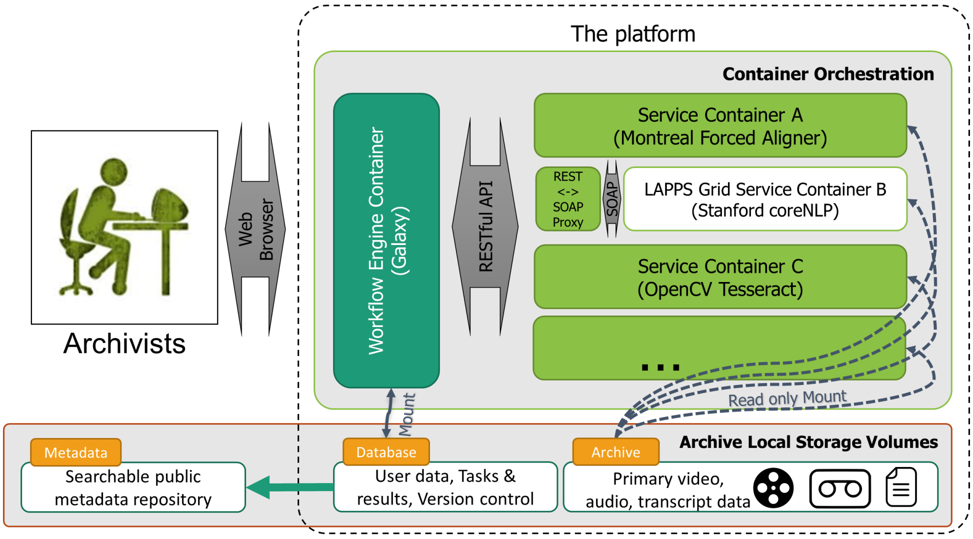

The platform
Container Orchestration
Archivists
Web Browser
Workflow Engine Container (Galaxy)
RESTful API
Service Container A (Montreal Forced Aligner)
REST <-> SOAP Proxy
SOAP
LAPPS Grid Service Container B (Stanford coreNLP)
Service Container C (OpenCV Tesseract)
...
Mount
Read only Mount
Archive Local Storage Volumes
Metadata
Searchable public metadata repository
Database
User data, Tasks & results, Version control
Archive
Primary video, audio, transcript data

Figure 1: Architectural sketch of CLAMS platform. Archives pull the containerized platform and services. The platform runs as an orchestrated set of containers that are connected to local storage to grant access to the data repository. Archivists interact with services to create, edit, and execute workflows only via the web-based front-end workflow engine.

CLAMS are first categorized by the anchor type on which the annotation is placed. That is, an annotation can be placed on 1) character offsets of a text, 2) time segments of time-based media, 3) two-dimensional (width × height) or three-dimensional (w × h × duration) bounding boxes on video frames, and 4) other annotation. For instance, a `named entity recognition (NER)` annotation can anchor on a `token` annotation that in turn anchored on character offsets. Furthermore, the characters can be from primary text data or from other annotations (such as ASR or optical character recognition (OCR)). Next, annotations are further categorized by the semantic types that are hierarchically defined in CLAMS vocabulary. For example, white noise detection and blank screen detection tools both produce subcategories of the `noisyFrame` annotation.

To address the complexity of additional annotation types and I/O constraints on tools, a layered annotation structure proved to be the best implementation choice for the interchange format based on many precedents, including LIF and TCF. Specifically, in MMIF, each tool generates a `view` object that contains all annotations as well as information about the production of the view (producer, production time, version, included annotation types, etc.). As a result, downstream tools can precisely locate any required input annotations from the input MMIF.

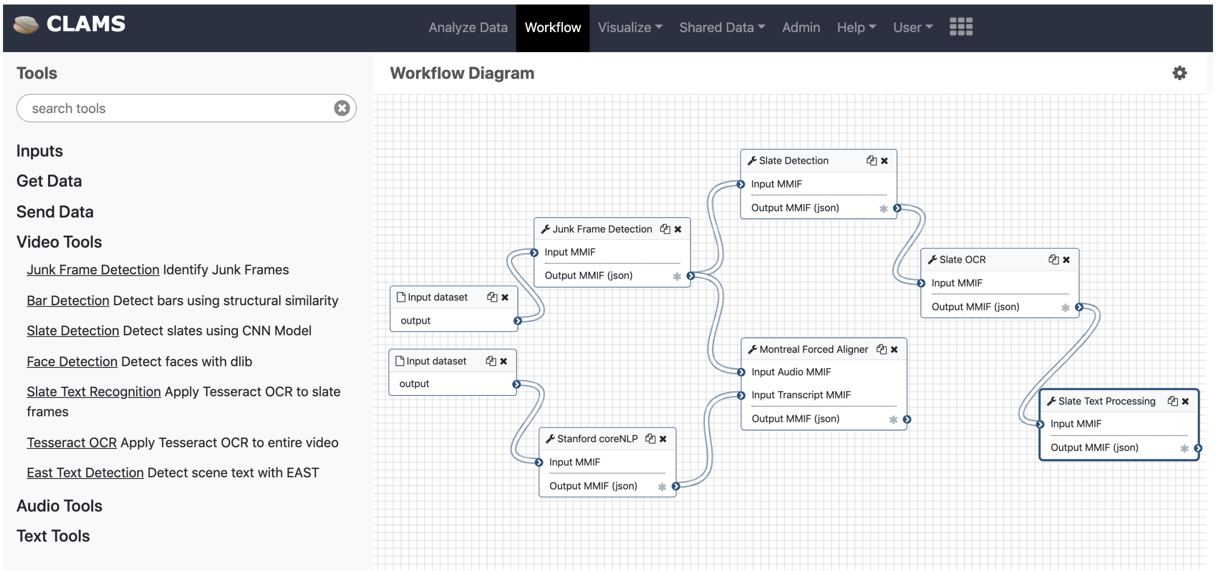
Last but not least, each tool deployed as a service on CLAMS must expose an application programming interface (API) to return its tool metadata, which contains information of the I/O constraints it poses. This tool metadata is used by the workflow engine to validate tool chains before creation and execution of workflows.

3.2 Workflow Engine

In order to facilitate the development of metadata generation workflows, we are using the Galaxy platform. The Galaxy platform was originally developed for genomic research, but has successfully been used for the deployment and integration of NLP tools (Giardine et al., 2005; Ide et al., 2016). Galaxy provides a web-based graphical user interface which will allow archivists to import data, construct complex multimodal workflows, and explore and visualize the metadata generated by applying workflows to their data.

3.3 CLAMS SDK and Services

We start with a number of fundamental analysis tools for text, image, audio, and video as CLAMS microservices. Users can easily configure a CLAMS instance with various tools based on specific needs, and then deploy it on a server where the archival data is stored. Figure 2 shows an example of a CLAMS instance configured with a set of video services. It also shows creation of a workflow of an ordered application of services to a specific set of input data.



Inputs

Get Data

Send Data

Video Tools

- Junk Frame Detection Identify Junk Frames
- Bar Detection Detect bars using structural similarity
- Slate Detection Detect slates using CNN Model
- Face Detection Detect faces with dlib
- Slate Text Recognition Apply Tesseract OCR to slate frames
- Tesseract OCR Apply Tesseract OCR to entire video
- East Text Detection Detect scene text with EAST

Audio Tools

Text Tools

Figure 2: An example workflow created using the Galaxy workflow engine

The SDK including core APIs used in the development and deployment of tools will release on an open repository under open source license.

3.3.1 Text Services

As the design of the interoperability layer, the MMIF, of CLAMS is largely inspired by and expanding that of the LAPPS Grid platform (LIF). The LAPPS Grid offers a wide range of text analysis services via its web-based SOAP API, and re-using them in CLAMS can be done by mapping these SOAP messages out to CLAMS API. These text analysis services include NER, parsing, relation extraction, and coreference resolution. Used on audio transcripts and OCR results, they will capture important entities, events, participants, and relations that can be included in the descriptive metadata.

3.3.2 Audiovisual Filtering Services

In spite of recent achievements in computer vision (CV) and ASR, such tools are still very expensive with respect to time and space to run. However, a video clip can include completely contentless blank frames or SMPTE bars as well as non-speech audio (music, natural sounds, beep, etc). Thus, blindly feeding those expensive CV and ASR tools with the entire clip can be not only a waste of computing resources, but can also result in introducing unnecessary noisy annotations. To address this problem, we added a range of less expensive *filtering* services such as blank screen detection, SMPTE bar detection, and HiPSTAS audio tagger (Clement et al., 2014).

3.3.3 ASR and Forced Alignment

The platform will include open source tools to process speech and audio from video and audio data. Audio processing will include Kaldi-based ASR which generates a transcript of the data that can then be processed with NLP tools. Additionally, CLAMS can provide forced alignment services such as the Montreal Forced Aligner, which generates time-aligned transcriptions from raw text transcripts (McAuliffe et al., 2017). These speech services in particular are very important for multimodal annotation, as they provide alignment between a time-based modality and a character-based modality.

3.3.4 Computer Vision Tools

Various types of metadata can be found in text displayed in frames of a video. Slates are video frames which display metadata such as air date, director, producer, and title. This metadata can be extracted by constructing a pipeline of computer vision and NLP tools. Text localization tools can detect the bounding boxes of text in a frame which can then be used to label a section of a video as a slate. Slate frames are then fed to a preprocessing tool and an OCR tool. The OCR tool generates unstructured text. Since the text generated through OCR is likely to contain significant errors, a subsequent tool processes this text to correct spelling errors and extract structured metadata from the corrected text.

In news programs, when a reporter or guest is introduced, it is common for their name and title to be displayed at the bottom of the frame in a chyron or “lower-third”. By applying OCR to chyrons, we can identify names of people appearing in a video. End credits contain production metadata such as cast and crew which can also be recognized by applying OCR tools.

Face detection and recognition (FDR) can be used to detect the location of faces in frames of video and to cluster detected faces so that individuals can be identified across different scenes within a video.

By integrating multiple vision and text based tools into a pipeline, it is possible to generate more robust metadata. For example, once clusters of detected faces are identified, this metadata can be combined with metadata from applying OCR to chyrons. By combining these two metadata sources, it will be possible to identify people in a video even after the chyron is no longer displayed. This metadata will be useful for researchers and archivists who are searching for all of the video segments in a dataset in which a particular person appears.

4 On-going and Future Work

We are currently collaborating with the American Archive of Public Broadcasting (AAPB) at WGBH Boston. The expertise of their archivists and librarians, as well as their perspective as target users, can provide us with insight towards selecting the analysis tools and phenomena of interest that can potentially push forward the state-of-the-art CL and CV technologies, within the vast unexplored collections of multimedia data. We actively seek collaboration with others in order to move closer to achieving an open platform for multimeida analysis.

We also believe that the platform can be used in academic settings with multimodal research datasets, such as MPII Movie Description dataset (Rohrbach et al., 2015), oral histories (StoryCorps, 2003; Telling Their Stories, 2005), and the The CHILDES Project (MacWhinney, 2014). For more technically literate users in research communities, we plan to develop a scriptable workflow engine extending the current SDK.

5 Conclusion

In this paper, we have presented CLAMS, a platform for multimodal computational analysis tools that provides interoperability between tools and a portable graphical user interface (GUI) workflow engine. Together, these tools can be used to automatically extract important information, such as timestamps (airing time, event time), people, companies, or historical events and relations, from time-based audiovisual material. We believe that archivists can use CLAMS over the digital multimedia collections they have to enrich item-level metadata of their collections and, in turn, greatly enhance the searchability and discoverability of their assets.

References

AAPB. 2015. American Archive of Public Broadcasting. http://americanarchive.org/. Accessed: 2019-02-20.

Patrick Aichroth, Christian Weigel, Thomas Kurz, Horst Stadler, Frank Drewes, Johanna Björklund, Kai Schlegel, Emanuel Berndl, Antonio Perez, Alex Bowyer, et al. 2015. Mico-media in context. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–4. IEEE.

Alessandra Antonaci, Michela Ott, and Francesca Pozzi. 2013. Virtual museums, cultural heritage education and 21st century skills. *Learning & Teaching with Media & Technology*, 185.

BBC. 2015. COMMA - BBC R & D. https://www.bbc.co.uk/rd/projects/comma. Accessed: 2019-02-20.

Maxwell T Boykoff and Jules M Boykoff. 2007. Climate change and journalistic norms: A case-study of us mass-media coverage. *Geoforum*, 38(6):1190–1204.

Karen Cariani, Sadie Roosa, Jack Brighton, and Brian Grane. 2015. Accelerating exposure of audiovisual collections: What’s next? In *Innovation, Collaboration, and Models: Proceedings of the CLIR Cataloging Hidden Special Collections and Archives Symposium*.

Heather Christenson. 2011. Hathitrust. *Library Resources & Technical Services*, 55(2):93–102.

Tanya E Clement, David Tcheng, Loretta Auvil, and Tony Borries. 2014. High performance sound technologies for access and scholarship (hipstas) in the digital humanities. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–10.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with gate’s full lifecycle open source text analytics. *PLoS computational biology*, 9(2):e1002854.

Margaret Stieg Dalton and Laurie Charnigo. 2004. Historians and their information sources. *College & Research Libraries*, 65(5):400–425.

Mark Doms and Norman J. Morin. 2004. Consumer sentiment, the economy, and the news media. Finance and Economics Discussion Series 2004-51, Board of Governors of the Federal Reserve System (US).

Jon W Dunn, Juliet L Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf. 2018. Audiovisual metadata platform (amp) planning project: Progress report and next steps. Technical report.

David Ferrucci, Adam Lally, Karin Verspoor, and Eric Nyberg. 2009. Unstructured information management architecture (UIMA) version 1.0. OASIS Standard.

Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: The d-spin text corpus format and its relationship with iso standards. In *LREC2010*, Valletta, Malta. European Language Resources Association (ELRA).

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. Weblicht: Web-based LRT services for german. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.

Erhard Hinrichs, Nancy Ide, James Pustejovsky, Jan Hajic, Marie Hinrichs, Mohammad Fazleh Elahi, Keith Suderman, Marc Verhagen, Kyeongmin Rim, Pavel Stranak, and Jozef Misutka. 2018. Bridging the LAPPS Grid and CLARIN. In *LREC2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, and Jonathan Wright. 2014. The language application grid. In *LREC2014*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nancy Ide, James Pustejovsky, Keith Suderman, Marc Verhagen, Christopher Cieri, and Eric Nyberg. 2016. The Language Application Grid and Galaxy. In *LREC 2016*, pages 51–70.

Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. 2015. The language application grid web service exchange vocabulary. In *International Workshop on Worldwide Language Service Infrastructure*, pages 18–32. Springer.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*.

MDPI. 2014. Media Digitization & Preservation Initiative. https://mdpi.iu.edu/. Accessed: 2019-02-20.

NDNP. 2005. National digital newspaper program ndnp: a partnership between the library of congress and the national endowment for the humanities. https://lccn.loc.gov/2005567119. Accessed: 2019-02-20.

NYPL. 2013. The New York Public Library's Community Oral History Project. http://oralhistory.nypl.org/. Accessed: 2019-02-20.

Douglas W Oard, Dina Demner-Fushman, Jan Hajič, Bhuvana Ramabhadran, Samuel Gustman, William J Byrne, Dagobert Soergel, Bonnie Dorr, Philip Resnik, and Michael Picheny. 2002. Cross-language access to recorded speech in the malach project. In *International Conference on Text, Speech and Dialogue*, pages 57–64. Springer.

OpenNLP. 2017. Apache OpenNLP. https://opennlp.apache.org/. Accessed: 2019-02-20.

Michela Ott and Francesca Pozzi. 2011. Towards a new era for cultural heritage education: Discussing the role of ict. *Computers in Human Behavior*, 27(4):1365–1371.

James Pustejovsky. 2018. Enhancing access to media collections and archives using computational linguistic tools. In *Proceedings of Enhancing Exploration of Audiovisual Collections with Computer-based Annotation Techniques, Workshop at AMIA*.

James Pustejovsky, Nancy Ide, Marc Verhagen, and Keith Suderman. 2017. Enhancing access to media collections and archives using computational linguistic tools. In *Proceedings of the Corpora for Digital Humanities Workshop*, pages 19–28. Association for Computational Linguistics.

Yves Raimond, Tristan Ferne, Michael Smethurst, and Gareth Adams. 2014. The bbc world service archive prototype. *Web Semantics: Science, Services and Agents on the World Wide Web*, 27:2–9.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jennifer Schaffner. 2009. The metadata is the interface: Better description for better discovery of archives and special collections, synthesized from user studies. http://www.oclc.org/programs/publications/reports/2009-06.pdf.

StoryCorps. 2003. Storycorps - stories from people of all backgrounds and beliefs. https://storycorps.org/. Accessed: 2019-04-04.

Telling Their Stories. 2005. Telling their stories oral history archives project. http://www.tellingstories.org/. Accessed: 2019-04-04.

Marc Verhagen, Keith Suderman, Di Wang, Nancy Ide, Chunqi Shi, Jonathan Wright, and James Pustejovsky. 2015. The lapps interchange format. In *International Workshop on Worldwide Language Service Infrastructure*, pages 33–47. Springer.