




OAVA: the open audio-visual archives aggregator

Polychronis Charitidis¹ · Sotirios Moschos¹ · Chrysostomos Bakouras¹ · Stavros Doropoulos¹ · Giorgos Makris¹ · Nikolas Mauropoulos¹ · Ilias Nitsos² · Sofia Zapounidou³ · Afrodite Malliari² 

Received: 15 September 2022 / Revised: 17 October 2023 / Accepted: 26 October 2023
© The Author(s) 2023

Abstract

The purpose of the current article is to provide an overview of an open-access audiovisual aggregation and search service platform developed for Greek audiovisual content during the OAVA (Open Access AudioVisual Archive) project. The platform allows the search of audiovisual resources utilizing metadata descriptions, as well as full-text search utilizing content generated from automatic speech recognition (ASR) processes through deep learning models. A dataset containing reliable Greek audiovisual content providers and their resources (1710 in total) is created. Both providers and resources are reviewed according to specific criteria already established and used for content aggregation purposes, to ensure the quality of the content and to avoid copyright infringements. Well-known aggregation services and well-established schemas for audiovisual resources have been studied and considered regarding both aggregated content and metadata. Most Greek audiovisual content providers do not use established metadata schemas when publishing their content, nor technical cooperation with them is guaranteed. Thus, a model is developed for reconciliation and aggregation. To utilize audiovisual resources the OAVA platform makes use of the latest state-of-the-art ASR approaches. OAVA platform supports Greek and English speech-to-text models. Specifically for Greek, to mitigate the scarcity of available datasets, a large-scale ASR dataset is annotated to train and evaluate deep learning architectures. The result of the above-mentioned efforts, namely selection of content, metadata, development of appropriate ASR techniques, and aggregation and enrichment of content and metadata, is the OAVA platform. This unified search mechanism for Greek audiovisual content will serve teaching, research, and cultural activities. OAVA platform is available at: <https://openvideoarchives.gr/>.

Keywords Audiovisual material · Speech-to-text technologies · Cultural heritage · Open access · Content aggregators

Polychronis Charitidis, Sotirios Moschos, Chrysostomos Bakouras, Stavros Doropoulos, Giorgos Makris, Nikolas Mauropoulos, Ilias Nitsos, Sofia Zapounidou, Afrodite Malliari have contributed equally to this work.

✉ Afrodite Malliari
malliari@ihu.gr

Polychronis Charitidis
pcharitidis@datascouting.com

Sotirios Moschos
smoschos@datascouting.com

Chrysostomos Bakouras
chbakouras@datascouting.com

Stavros Doropoulos
doro@datascouting.com

Giorgos Makris
gmakris@datascouting.com

Nikolas Mauropoulos
nmauropoulos@datascouting.com

1 Introduction

Locating sources of information on the Internet is a common task performed for a variety of reasons ranging from research and education, to business and leisure. There is an obvious need for continuous improvement of search services regardless of the type of content. Audiovisual material is often left

Ilias Nitsos
initsos@ihu.gr

Sofia Zapounidou
szapounidou@nlgr.gr

¹ Datascouting, 30 Vakchou Street, 54629 Thessaloniki, Greece

² Department of Library, Archival and Information Studies, International Hellenic University, P.O. Box 141 Sindos, 57400 Thessaloniki, Greece

³ Cataloging Department, National Library of Greece, Leof. Andrea Siggrou 364 Kallithea, 17674 Athens, Greece

unused as it is more difficult to retrieve and index compared to written content. Especially in the case of Greece, audiovisual content is even harder to find because a National Registry for audiovisual providers has not been developed yet.

This paper focuses on the OAVA (Open Access AudioVisual Archive) platform, an open-access audiovisual aggregation and search service for Greek audiovisual content. The platform allows the search of audiovisual resources utilizing metadata descriptions, as well as full-text search utilizing content generated from automatic speech recognition processes through deep learning models. The dataset contained in OAVA is multidisciplinary with a variety of Greek audiovisual resources, e.g., videos and narratives of historical and everyday life, scientific, academic and cultural events, etc. All resources are free of copyright restrictions, with informative content (not literary or artistic), as this project aims to provide access to aggregated metadata as well as to the actual content.

To utilize audiovisual resources the OAVA platform makes use of the latest state-of-the-art automatic speech recognition technologies. OAVA platform supports Greek and English speech-to-text models. Specifically for Greek, to mitigate the scarcity of available datasets, a large-scale ASR dataset was annotated in order to train and evaluate deep learning architectures. The ASR models are supported by many auxiliary tasks like boosting the quality of the transcriptions with language models, adding punctuation and capitalization, speaker diarization, and gender identification.

In the sections that follow, related work is reviewed and discussed, the selection process used to create the dataset of reliable Greek audiovisual providers and content is briefly described, the OAVA schema and details on the platform's architecture and technologies are presented and explained. The paper concludes with a discussion on the OAVA project, its goals and ambitions, and current limitations.

2 Related work

Review of aggregation services and national registries

For the development of the OAVA platform, other well-known aggregation services and in particular their collections, the underlying models and technologies have been considered (see Table 1).

Europeana is an aggregation service that collects information from over 4,000 different institutions through a network of aggregating partners. It is funded by the European Union and provides digital access to more than 53 million digital items of European cultural heritage material [17]. It was first launched on 20 November 2008 giving access to 4.5 million digital items from over 1,000 collections at the time [61]. At

its current development stage, Europeana is built using Vue.js and Nuxt [20] and offers several APIs that third-party developers may use: REST API, Search API, Record API, Entity API, Annotations API, IIIF APIs, SPARQL [21]. Europeana uses the Europeana Data Model (EDM). This means that publishers need to describe data using EDM before delivery [22].

Similar to Europeana, the Digital Public Library of America serves as the national registry for digital collections in the United States making over 46 million of digital items available to everyone in a one-stop search portal [11]. It receives content from a network of Content Hubs and Service Hubs that offer collections, or are responsible for compiling records from collections in a state or geographical area [14, 30]. DPLA uses MAP (Metadata Application Profile), a model based on Europeana's EDM model and publishers are responsible for complying with the metadata guidelines [13, 30].

Trove is the aggregation service offered by the National Library of Australia [34, 35]. It focuses on free digital content about Australia or of interest to the Australian community, that is stored in the collections of organizations such as: Australian Libraries, Archives, Museums, Art Galleries, Universities, etc. [53]. The schema used by Trove is a basic schema known as Simple Dublin Core combined with custom fields as well as additional elements from other schemas such as DCMI Terms to meet specific Trove needs [53]. Trove's aggregator receives metadata through the National Library of Australia Harvester that regularly visits websites and repositories of partner institutions. The NLA Harvester first came online in 2008 and it is developed in-house by the National Library of Australia using Java Technologies. The Trove discovery interface has been based on Vue.js technology since 2020 when it was rebuilt from scratch. More details about the technologies used to enable Trove features can be found on the Trove Technical Ecosystem webpage [51].

DigitalNZ is New Zealand's national registry that contains more than 30 million digital items from more than 200 organizations on facts about the country [8]. Its content partners include libraries, museums, galleries, government departments, the media and community groups [8]. The DigitalNZ harvesting system uses field names and schemas that are loosely based on Dublin Core (DC) [9]. The core software that implements the aggregation, search and sharing of metadata records is Supplejack, a Ruby on Rails platform [10, 26] developed by DigitalNZ at the National Library of New Zealand. The same platform has also been considered by Canadian Libraries to develop a Cross-Canadian Digital Library Platform [2].

The National Digital Library of India currently offers access to more than 87 million digital items [50] including textual documents, audio and video files [49]. The metadata schemas used in the case of NDLI include Dublin Core

Table 1 Aggregation services

Name	Description
Europeana	Funded by the European Union, it provides access to European cultural heritage material
DPLA	The Digital Public Library of America is USA's national registry for digital content
Trove	Australia's national registry for content that is of interest to the Australian community
DigitalNZ	New Zealand's national registry with content items about New Zealand
NDLI	The National Digital Library of India is the country's national registry
PARTHENOS	An aggregation service for researchers using Digital Humanities Infrastructures

for generic material, Learning Resource Metadata Initiative (LRMI) for educational content and Shodhganga for theses [49]. NDLI harvests content from national sources and aims to act as a one-stop shop mainly focusing on educational resources [3, 49].

PARTHENOS (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies) was a project initiated by the European Commission and its main objective was to build an integrated environment for researchers using Digital Humanities Infrastructures (DHIs). DHIs are e-infrastructures that facilitate research by providing access to data and enabling the use of ICT tools [24]. The Content Cloud Framework of PARTHENOS uses a common data model to aggregate resource metadata (i.e., metadata about research data, services and tools) in order to make them available to humans and machines via: a Solr index for search and browsing, OAI-PMH for bulk download and Virtuoso server for SPARQL queries [24]. The aggregator is implemented using the D-NET Software Toolkit, a framework designed for constructing custom aggregative infrastructures in a cost-effective way [24, 46].

Carare has been another project for the aggregation of archeological and architectural resources. In the context of this project, the Carare schema [23] was developed to harmonize providers' metadata and to aggregate them for Europeana. Given the project's scope and the nature of the resources being aggregated, the development of the Carare metadata schema considered the CIDOC-CRM model [4] for cultural heritage and the EDM [22].

Review of automatic speech recognition

In recent years, following the advent of deep neural networks, automatic speech recognition has gained a lot of attention in a variety of academic and commercial works. More specifically, in the related research literature for the English language, there is an abundance of publicly available datasets and architectures that outperform all traditional machine learning automatic speech recognition approaches [6]. These datasets consist of several thousand hours of transcribed English speech and they are used to train state-of-the-art deep learning architectures for automatic speech recognition. The

availability of such pre-trained models made it possible for the field to progress even for languages that are not as popular as the English language.

For the task of automatic speech recognition, the most prominent architecture categories consist of the Connectionist Temporal Classification (CTC) models [29] and RNN-Transducers (RNN-T) [28]. The former models are based on CTC objective function which computes the alignment between the input speech signal and the output sequence of the words and aims to maximize the total probability of the correct alignments. CTC models use a simple encoder to map the speech signal to target labels. The latter models extend CTC modeling by incorporating an acoustic model with its encoder, a language model with its prediction network, and a decoding process with its joint network.

CTC models include preliminary implementations, such as Jasper [43] and Quartznet [41], and more recent ones, like Citrinet [44]. The Jasper model is a deep time delay neural network (TDNN) comprising of 1D-convolutional layers with residual connections between them, where each sub-block contains a 1-D convolution, batch normalization, ReLU, and dropout. QuartzNet is a version of Jasper model with separable convolutions and larger filters, where each sub-block contains a 1-D separable convolution, batch normalization, ReLU, and dropout. Citrinet is a deep residual neural model which uses 1D time-channel separable convolutions combined with sub-word encoding and squeeze-and-excitation. For the RNN-T models, experiments were conducted with ContextNet [32] and Conformer [31]. ContextNet features a fully convolutional encoder that incorporates global context information into convolution layers by adding squeeze-and-excitation modules. At last, Conformer combines convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way.

3 Finding resources: selection process

The obvious goal of content aggregation services is to facilitate individuals and/ or software services, to locate information resources derived from content-contributing col-

lections. The OAVA platform utilized a range of selection criteria to evaluate audiovisual collections, considering factors such as copyright status, content quality, and technical characteristics. Information regarding these selection criteria and their application can be found in [45]. The final dataset contains audiovisual resources, free of copyright limitations, with informative content that varies from videos and narrations regarding historical and everyday life events to scientific, academic, and cultural events. Literary or artistic content is not being considered as a source due to potential copyright restrictions.

The resources were gathered from October 2020 to March 2021 mainly by searching the web and browsing content, as there is no national searchable registry of audiovisual material in Greece. It is worth noting that the National Centre of Audiovisual Media and Communication (EKOME) currently prioritizes the attraction of investments through the production of films, television series, documentaries, animation, and digital games in Greece, despite its aim to register all institutions holding audiovisual archives [48]. A set of selected keywords was used for searching, and several websites of organizations expected to provide audiovisual content due to their activities and functions were visited and examined in detail [45]. A study of mapping collections of audiovisual resources in Europe [38] was taken into account in order to identify types/ categories of possible providers of Greek audiovisual content such as museums, libraries, archives, and private collectors.

The above search resulted in 500 providers of Greek-language audiovisual resources with their content considered as candidate material for the OAVA platform. Those providers and their resources were reviewed. The review process contained two stages: the first was to be evaluated using the crAap test (Currency, Relevance, Authority, Accuracy, and Purpose). The second stage was to evaluate further the resources of the eligible providers (497 out of 500) with specific selection criteria. Criteria based on studies [12, 52, 57, 59, 62] were grouped under context, content, form/use, process or technical, and metadata. The second stage resulted in 233 trusted providers containing: libraries (8%), archives (3%), museums (9%), universities (28%), governmental (30%) and non-governmental organizations (6%), private institutions (13%), media organizations (1%), and specific course projects (2%). The collections involved encompassed various topics such as open courses, education and training, academic and scientific events (e.g., webinars, lectures), cultural events, interviews (e.g., oral histories, press conferences), and campaigns promoting the missions of NGOs. The whole procedure of searching and selecting the resources is presented in detail in “Mapping audiovisual content providers and resources in Greece” [45].

A dataset containing reliable Greek audiovisual content providers and their resources (1710 in total) was created. The dataset can be found on Zenodo.¹

4 Aggregation schemas and the OAVA schema

Since aggregation involves homogenization of many different descriptive practices on models, application profiles, metadata, etc., it makes sense for early aggregation services to have started aggregating content using core metadata elements, such as the Europeana Semantic Elements used in the context of the Europeana Digital Library [19]. In order to enrich metadata using semantics, the next step was to move to the Europeana Data Model (EDM) [18] or to the DPLA MAP of the Digital Public Library of America (DPLA). As already mentioned, DPLA has designed its Metadata Application Profile (DPLA MAP) building on EDM [13]. It is worth mentioning that both EDM and DPLA MAP reference properties from well-known schemas such as Dublin Core, RDF Schema, SKOS, etc. Trove on the other hand uses the Simple Dublin Core schema combined with custom fields and additional elements from other schemas to meet specific Trove needs [54]. In the case of DigitalNZ, the harvesting system uses field names and schemas that are loosely based on Dublin Core (DC) [9].

For the development of the OAVA schema, the aforementioned practices were considered. Yet, there are two core differences. First of all mentioned aggregation services gather content in various media, with text and image being the prevalent media types. The core classes in each model are designed to be generic in order to accommodate information that is relevant to all types of resources. For instance, the Provided Cultural Heritage Object in EDM and the Source Resource in DPLA MAP are examples of such generic classes. Secondly, the aggregation process involves close collaboration with content providers, which entails the sharing of information regarding their schemas, application profiles, and vocabularies. This collaboration is crucial in terms of both the resource description approaches adopted by the providers and the mapping of local schemas to the aggregator’s schema or model. A case in point is the collaboration between content providers and Europeana, whereby mapping tools like MINT [55] are used to map local schemas to the Europeana Data Model [56].

In contrast to Europeana, the OAVA project is not an official organization, and collaboration with stakeholders cannot be guaranteed. Additionally, Greek audiovisual content providers often adopt local practices that involve unstructured formats [45], which can hinder the use of mapping

¹ <https://doi.org/10.5281/zenodo.5112283>.

v.1.1

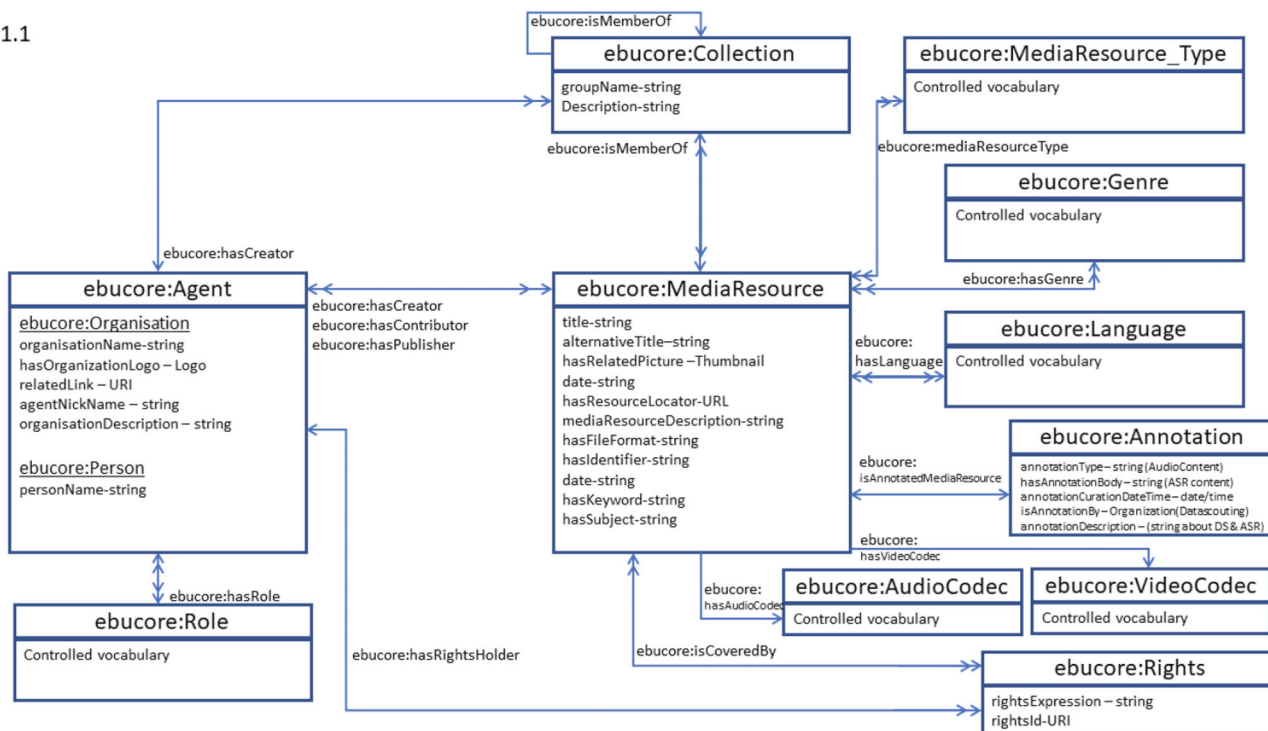


Fig. 1 The OAVA schema

tools such as MINT. Moreover, audiovisual content, which is the only media type to be aggregated by the OAVA aggregation service, has some special descriptive needs. Due to these factors, metadata aggregation was carried out using heuristic methods, and a specialized model known as the EBUCore model and schema was selected as suitable for the project’s needs. EBUCore has been developed by the European Broadcasting Union for describing audiovisual and streaming content. According to the specifications, “EBUCore is the Dublin Core for media” [16]. The primary class in the EBUCore model is the Media Resource class, which includes specialized properties designed to capture information relevant to audiovisual resources. In addition, EBUCore features other classes, such as Collection, Genre, and Annotation, which, when instantiated, can provide additional contextual information about a particular Media Resource instance.

For the needs of the OAVA service, a subset of EBUCore classes and properties were selected to aggregate metadata about the audiovisual resources of interest. The concluding OAVA schema is presented in Fig. 1. Rectangles in the diagram represent classes, and each class contains its properties. The relationships between classes are illustrated with arrows, where a single arrow represents a cardinality of 1, while double arrows represent a cardinality of many. Figure 1 presents all types of cardinality constraints (1-1, 1-M, M-N) in a clear and concise manner.

The primary class of the OAVA schema is MediaResource with which an audiovisual resource is described. Each MediaResource may be part of one or more Collections. The Agent class is used for the description of persons or corporate bodies that relate to the MediaResource content in terms of creation, contribution, etc. Additional information regarding the MediaResource is given regarding its type (MediaResource_Type), its language (Language), its genre (Genre), the compression of its audio (AudioCodec) or video (VideoCodec) content. The Annotation class is used to hold the ASR content and details about its creation, e.g., date.

It must be noted that the OAVA platform stores the aggregated metadata in a VuFind (reference) instance. VuFind also serves as the main search mechanism in OAVA.

5 Harmonization process

The schema used in VuFind is a flat list of more than 100 elements that serve record-based descriptions. For the purpose of the OAVA project, an analysis was conducted on the VuFind schema elements to establish any correspondences to the elements of the OAVA schema. In some cases, no such correspondences were found. To address this, the VuFind schema was enriched by incorporating the necessary OAVA schema elements, which is a subset of the EBUCore Schema. This resulted in the proper description of aggregated audiovi-

sual resources, with a total of 38 elements utilized in VuFind. Specifically, 22 VuFind schema elements corresponded to OAVA schema properties, while 16 properties were derived from the EBUCore Schema.

Aggregation of audiovisual content usually employs heuristic methods as most content providers use local elements that are not available in structured formats, as noted by [45]. Consequently, metadata is displayed as plain text within HTML content elements on their webpages. To effectively map metadata from these content providers' websites to the OAVA schema, an analysis was conducted on the HTML content elements of each website, which were then mapped to the previously described VuFind elements. This process was time-consuming and requires constant updates due to changes in providers' websites. The technical description of the crawling and scraping processes can be found in the following section of this paper.

Metadata enrichment is an integral part of the aggregation and harmonization process. The enrichment is carried out for two main reasons: firstly, to ensure conformity with the OAVA model by providing necessary information, and secondly, to enhance the user experience on the OAVA platform. For instance, enrichment related to the OAVA schema may involve the automated assignment of type (e.g., video or audio) and language of the resource, while enrichments related to user experience may involve the creation of a thumbnail for the resource and the storage of the provider's logo. Notably, the most significant enrichment is derived from the full text generated by the Automatic Speech Recognition (ASR) mechanism, which is described in detail in the next section of this paper. However, it is important to note that the ASR-produced text is solely used for full-text searching and is not utilized as a source for additional metadata enrichment.

It is worth noting that the descriptions within the VuFind mechanism are flat. Each record in VuFind provides a description of a single MediaResource instance, along with information on other EBUCore classes that are linked to the specific MediaResource instance. As a result, OAVA semantics are derived from these flat descriptions by instantiating the appropriate classes and properties.

6 OAVA platform

A general overview of the OAVA platform architecture is presented below. Basic components and processing stages of the OAVA platform are discussed next, and detailed information is included for the Automatic Speech Recognition approach, especially in the case of Greek content.

The OAVA platform has been built around a distributed microservice architecture (see Fig. 2). All platform services communicate through a message broker, RabbitMQ, which

not only acts as a load-balancing mechanism but also allows individual services to scale horizontally.

At the core of the OAVA platform lies the orchestrator, a system responsible for monitoring sources of audiovisual content. The orchestrator, periodically, instructs the workers to check every registered source for content that has not yet been admitted to the platform through the crawling mechanism. Crawling produces a set of links for each of which a scraping procedure is started. The product of a scraping procedure is the extracted metadata (title, description, authors, publishers, video links, etc) which are then analyzed, stored and indexed by the orchestrator. A side-benefit of the extraction process is that all information ends up being represented by a universal schema which makes handling a lot more manageable. Any multimedia links are subsequently handled by the storage manager service which is responsible for finding the optimal approach to retrieve the remote content and storing it in the file system. The paths to the files downloaded by the storage manager are then forwarded to the ASR service. Any audiovisual content remains stored until the ASR service has extracted the text. All text produced from the audiovisual content is sent to the orchestrator while the storage manager is being instructed to delete the file. It is ensured that no original content is permanently saved or served, even in cases where the ASR service would fail to provide output, proper actions are being taken to erase the problematic file.

An in-depth insight follows that presents both processes and key parameters in the OAVA platform.

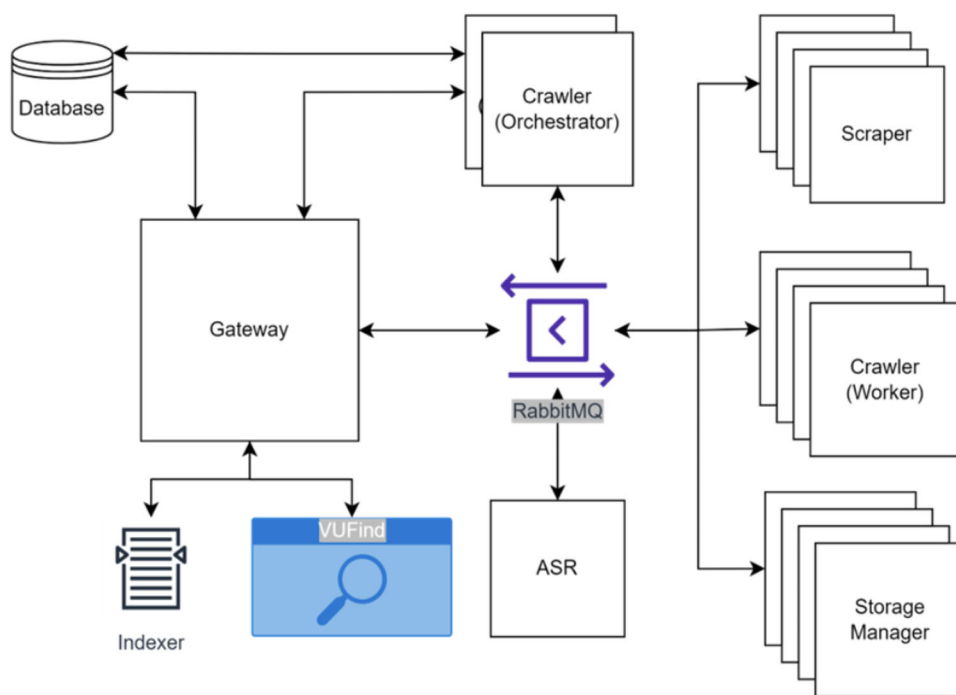
6.1 Crawling resources: retrieving metadata and content

Data collection and its extraction has been a daunting task since the conception of this project. It was a requirement not only to obtain a coherent and diverse archive of all the digital cultural and historical material, but also to augment it with the aid of modern technologies to be readily available to the general public. The final result is derived from a two-step process that first discovers such material, often referred to as crawling, and then extracts any important information or media, also known as scraping. It is worth noting that the developed crawlers are able to utilize XML, RSS, and OAI-PMH prototypes, but most of the sources do not support such prototypes and even when they do, they do not contain all the necessary information.

6.1.1 Crawling

Crawling is the process of traversing a website in order to find information that could be scraped and is best described from a graph theory point of view. Each node represents a web page while edges represent links between them. It is a directed graph that allows for easy traversal given a node.

Fig. 2 Overview of the OAVA platform architecture



While, initially, the web is a cyclic directed graph of nodes V , it can be turned into acyclic when considering the set of edges E as unique with respect to their set and target:

$$\begin{aligned}
 E = \{ & (x, y) \mid (x, y) \in V^2, \\
 & x \neq y, \\
 & z \in V, \\
 & z \neq x, \\
 & (z, y) \notin V^2 \\
 & (y, x) \notin V^2 \}
 \end{aligned}
 \tag{1}$$

Given this simplification, it is much easier to crawl through pages, without worrying about loops. Therefore, crawling is defined by a starting node and a maximum depth. Other restrictions have also been set in place to avoid false positives such as discarding pages that do not belong in the same domain as the starting node unless explicitly allowed by the administrator.

From a technical standpoint, this simplification is represented by a URL caching mechanism that will not allow the same web page to be scraped more than once. This has a massive impact on the design of the platform and what a web page actually represents. It implies that the content of any web page is immutable and any subsequent visitation would not yield different results. This is not entirely true when considering the example of news outlets that will often update an existing article, but it is a necessary design choice to significantly reduce computing requirements as otherwise, the platform would have to reprocess web pages that have already

been admitted into the system without actually gaining anything as far as new information is concerned except for rare circumstances. It is computationally expensive to automatically detect such updates, but there are in place mechanisms that can be triggered in a defined interval.

6.1.2 Scraping

Scraping is the process of distilling a web page’s contents into usable data. In essence, a headless browser is used to navigate the web and retrieve content in HTML form which is then processed. A large percentage of that content has no value for the purposes of the platform so CSS rules are used to query the HTML document for elements with relevant information which are then mapped to the respective fields of the platform’s metadata model.

There are inherent difficulties when dealing with such varied and often technologically dated sources. They all have to be treated as different cases, essentially making a profile for each website, to define very specific rules for data extraction. In exceptional cases, the entirety of the information, with regards to a specific data point, was spread over a number of different web pages which had to be co-related by the scraper. What started as a simple scraping procedure, based on CSS rules, quickly turned into a convoluted system that requires the administrator to define the extraction steps. Such steps involve both data extraction based on the HTML code provided, and procedures, such as network call interceptions, custom JavaScript injection, and user input emulation.

A large number of the sources that were crawled had content that was created for Flash Player, a now deprecated technology. The cost and effort that would be required in order to extract any information from such applications was deemed far out of the scope of this project and these are the only nuggets of information that were left behind.

6.2 Automatic speech recognition

One of the crucial tasks of the OAVA platform is to automatically transcribe files that contain audio, using deep learning and more specifically automatic speech recognition techniques. Utilizing the extracted text, the platform can then provide additional functionalities for search and retrieval. To accomplish this, automatic speech recognition models were utilized for both English and Greek audio content.

In the case of the English language, there is a plethora of available pre-trained models and architectures. The most effective speech recognition model was selected after comparing the performance of several pre-trained models with diverse architectures on the same dataset. Specifically, the word error rate (WER) evaluation metric was calculated on a test set (librispeech-dev-other) of the Librispeech [58] dataset for every pre-trained model. The models were trained on combinations of several publicly available English speech datasets. The vocabulary in which every model is trained plays a crucial role in the performance of the model. By the term vocabulary, we mean all the structural sets of symbols that can be predicted by a model. Such symbols include letters of the alphabet, punctuation marks, spaces, and numbers, and models trained with such characters are Jasper and Quartznet. In order to reduce the complexity of the alphabet, capital letters and most punctuation marks, as well as numbers have been removed and as a result, these models can only predict lowercase characters of the English alphabet. Apart from the aforementioned models that are trained with characters, the rest of the models can be trained with syllables, word roots, or even whole words. More specifically, the way to extract such a vocabulary that is based on Byte-Pair Encoding (BPE) [63] is done with the help of sentencepiece library [42]. Byte-Pair Encoding is a form of data compression where the most common character pairs are replaced by a new character that does not exist in the vocabulary and ensures that the most common words are represented in the vocabulary as a new symbol while rare words are broken into two or more subwords. The number of characters in the alphabet can be specified when training a tokenizer. The models use pre-trained tokenizers bpe_v128 and bpe_v1024 with 128 and 1024 characters, respectively. Finally, the char tokenizer is the tokenizer that splits words into simple letters, apostrophes, and spaces.

Table 2 shows the evaluation results from various ASR models. The model that displayed the lowest WER value

Table 2 Evaluation results of various architectures for the English model

Architecture	Model	Vocabulary	WER
CTC	Jasper	char	0.102
CTC	Quartznet	char	0.113
CTC	CitriNet	bpe_1024	0.077
RNN-Transducer	Contextnet	bpe_1024	0.038
CTC	Conformer	bpe_128	0.037
RNN-Transducer	Conformer	bpe_1024	0.029

Bold values represent the results with lowest word error rate (WER)

on the Librispeech development set is a Conformer based on RNN-Transducer architecture. This model consists of around 0.6B parameters and is trained on a composite dataset comprised of 11 publicly available datasets (Librispeech 960h of English speech [58], Fisher Corpus [5], Switchboard-1 Dataset [27], WSJ-0 and WSJ-1 [15], National Speech Corpus (Part 1, Part 6) [3], VCTK, VoxPopuli (EN) [65], Europarl-ASR (EN) [36], Multilingual Librispeech (MLS EN) - 2,000 hrs subset [60], Mozilla Common Voice (v8.0) [1], People's Speech - 12,000 hrs subset [25]). In order to further improve the WER, we measure the performance of the acoustic model with the help of a language model (LM). An N-gram model with $n=4$ was trained for the language model using the kenlm [33] library on a subset of Wikipedia edited by huggingface.² This model can be used with beam search decoders on top of the acoustic model to produce more accurate predictions. The beam search decoder would incorporate the scores produced by the N-gram LM into its score calculations as shown in equation 2:

$$\begin{aligned} final\ score &= acoustic\ score \\ &+ \alpha * LM\ score \\ &+ \beta * sequence\ length \end{aligned} \quad (2)$$

where *acoustic score* is the score predicted by the acoustic encoder and *LM score* is the one estimated by the LM. The parameter α specifies the amount of importance to place on the N-gram language model, and β is a penalty term to consider the sequence length in the scores. Larger α means more importance on the LM and less importance on the acoustic model. Negative values for β will penalize longer sequences and make the decoder predict shorter letter sequences, while positive values would result in longer words. In exploratory experiments, we observed that the best performance is obtained with $\alpha = 0.5$ and $\beta = 0.5$ Table 3 indicates there was no significant improvement of the WER metric, as the model's performance is already very high.

² <https://huggingface.co/datasets/wikipedia>.

Table 3 Comparison of best-performing acoustic model with and without a language model

Architecture	Model	LM	WER
RNN-Transducer	Conformer	No	0.0294
RNN-Transducer	Conformer	Yes	0.0279

Bold values represent the results with lowest word error rate (WER)

In contrast with the English language, there are limited ASR resources for the Greek language. Training such models requires large-scale speech recognition datasets, which consist of speech in audio format and their corresponding text transcription. A few publicly available examples of Greek datasets are the Mozilla Common Voice Greek subset³ (28 h), a Transcribed book⁴ (4 h), and the Greek TEDx speech dataset⁵ (20 h). It is apparent that the size of this corpus is very small compared to English datasets. This might impose many limitations on the derived model.

One important limitation is the potential underlying bias that might occur in a such limited set of data. Such bias might lead to poor model performance. An obvious solution to this is to obtain more diverse data for training, but this process is labor-intensive and time-consuming. We estimated that one hour of audio transcription accounts for approximately 8 man-hours. To mitigate this limitation and expand the dataset, we proceed to the manual annotation of more than 80 h of additional data. The new annotated corpus consists of diverse audio from the Greek parliament channel⁶ (60 h), audio from Greek news channels (3 h), audio from university open lectures (8 h), audio from the Onassis foundation event (3 h), audio from NGO events (8 h), and audio from library events (2 h). These sources are selected due to their diverse discussion scope, the speaker diversity, and their availability for public use.

The final dataset consists of 136 h of audio data and corresponding transcriptions. The dataset size is still smaller than the English dataset. This means that underlying biases can still be present. To address this issue, we utilize the transfer learning approach, a common approach in deep learning model training. Specifically, we begin by employing a pre-trained English model as the foundation for training our Greek model. This is proven to be beneficial to the model performance, as the deep learning training process manages to utilize the English model. Our preliminary experiments support this claim, as the model derived using transfer learning achieved better performance than the one trained without this approach.

³ <https://commonvoice.mozilla.org/el/datasets>.

⁴ <https://www.kaggle.com/datasets/bryanpark/greek-single-speaker-speech-dataset>.

⁵ <http://www.openslr.org/100>.

⁶ <https://www.hellenicparliament.gr/Enimerosi/Vouli-Tileorasi>.

Table 4 Evaluation of different architectures in two datasets

Architecture	Model	Vocabulary	WER
(a) Evaluation in <i>cultural</i> dataset			
CTC	Jasper	char	0.31
CTC	QuartzNet	char	0.29
CTC	Citrinet	bpe_128	0.27
RNN-Transducer	ContextNet	bpe_128	0.26
RNN-Transducer	Conformer	bpe_128	0.23
CTC	Conformer	bpe_128	0.24
(b) Evaluation in <i>news</i> dataset			
CTC	Jasper	char	0.27
CTC	QuartzNet	char	0.25
CTC	Citrinet	bpe_128	0.25
RNN-Transducer	ContextNet	bpe_128	0.24
RNN-Transducer	Conformer	bpe_128	0.20
CTC	Conformer	bpe_128	0.20

Bold values represent the results with lowest word error rate (WER)

To train and evaluate the model we split the dataset into different sets. The training set consists of the publicly available Greek datasets, combined with the data from the Greek parliament channel and data from university open lectures, with a total duration of 120 h. For the evaluation set, we use two separate datasets that are different from the training datasets and have different scopes. This can help us better understand the performance of the model and discover potential biases on unseen data from different domains. We denote the first as the “*cultural dataset*” and consists of the data from the Onassis Foundation event, the data from NGO events, and the data from library events, with a total duration of 13 h. We denote the second as the “*news dataset*” and consist of data from Greek news channels with a total duration of 3 h. Note that the main goal of OAVA is to develop general-purpose ASR models that can work well in various domains. Evaluating the model on two holdout datasets related to news and culture will provide us with a strong indication of the model’s performance on data that belong to a domain different from that of the training data.

We trained the same ASR architectures as the English version. Table 4 shows the evaluation results of the models in these two datasets.

It is apparent that CTC architectures have higher error rates compared to RNN-Transducer architectures. One exception is the CTC Conformer architecture which competes with the RNN-Transducer counterparts in both datasets. The best model in the cultural dataset Table 4a evaluations is the RNN-Transducer Conformer with 0.23 WER. For the news dataset Table 4b, the RNN-Transducer of Conformer performs on par with the CTC version with 0.20 WER.

Furthermore, the overall WER of the transducer model is further improved by adding a language model as a post-

Table 5 Evaluation of best-performing acoustic models with and without a language model in *cultural* and *news* datasets

Dataset	Architecture	Model	Vocabulary	WER
<i>cultural</i>	RNN-Transducer	Conformer	bpe_128	0.18
<i>cultural</i>	CTC	Conformer	bpe_128	0.19
<i>news</i>	RNN-Transducer	Conformer	bpe_128	0.16
<i>news</i>	CTC	Conformer	bpe_128	0.16

processing step. Following the same procedure as in English, a 4-gram language model is trained using millions of sentences from the CC100-Greek dataset [66]. In order to obtain the *final score* we utilized equation 2 with $\alpha = 1.3$ and $\beta = 1.7$ showing the most prominent results after exploratory experiments. Table 5 contains the results of the best-performing models with the language model processing. We can observe that using the language models has a huge impact on the performance of the models as it improves the WER by almost 0.04 in every setting.

Because the CTC architectures are faster than the RNN-Transducer ones, we chose the former as the Greek ASR model. We further analyze the ASR system performance and the required resources in Sect. 6.2.3.

The evaluation shows that the Greek model's performance, in terms of WER, is not as good as the English model's. This finding is understandable given the limited availability of training data, and is commonly observed with low-resource languages, even in commercial applications. For example, Azure ASR services claim that a WER of 0.20 is acceptable for low-resource languages, but additional training can still provide benefits.⁷ In our experiments, the best-performing model scores 0.18 and 0.16 WER in “*news*” and “*cultural*” sets respectively. This is strong evidence that the model can provide acceptable transcription in out-of-training samples. Although this evaluation seems promising, it is apparent that it can benefit from training with additional data. This is something we will consider for future work.

6.2.1 Auxiliary tasks

Apart from the models that are responsible for the speech-to-text conversion, additional effort is made toward improving the output of these models. Examples of such efforts include grouping individual speakers by assigning unique identifiers to each one of them, identifying the gender of the speaker, adding punctuation to the output text, and capitalizing specific words. In this subsection, we will describe these auxiliary tasks.

Speaker diarization and gender identification For the speaker diarization task, SpeakerNet architecture is utilized.

⁷ <https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio>.

SpeakerNet [39] is a neural architecture for speaker recognition and speaker verification tasks. It is composed of residual blocks with 1D depth-wise separable convolutions, batch normalization, and ReLU layers. This architecture maps variable-length utterances to a fixed-length embedding and the embeddings from the same speaker are trained to have a small euclidean distance. A clustering process can then associate each utterance to one speaker with a particular identifier. For English, the pre-trained SpeakerNet is employed. For the training set, the VoxCeleb1 and VoxCeleb2 [47] datasets were used. This model achieves a 5.4% Diarization Error Rate (SER) in CALLHOME American English Speech dataset.

Additionally, SpeakerNet is expanded to support gender identification. To accomplish this, an additional dense layer is added to the architecture, keeping the rest of the weights frozen during training. The training process tries to optimize the categorical cross-entropy loss function using VoxCeleb2 dataset as the training set. VoxCeleb2 contains two thousand audio clips of women and three thousand audio clips of men, which are subsampled for class balance. Stochastic gradient descent was used for training the models with a batch size of 32 and a learning rate of $10 * e^{-4}$. The accuracy of the model to a small test set of 100 samples with the same distribution of classes as the training set is 97%. This model is not language dependent and can be used for both English and Greek.

Punctuation and capitalization Speech-to-text models are not usually trained to generate punctuation or capital letters. This makes text comprehension difficult. To mitigate this, related works train a language model to generate punctuation or capitalize words. Following this paradigm, for each word in the input text, the Punctuation and Capitalization model predicts a punctuation mark that should follow the word (if any). The model supports commas, periods, and question marks. Also, the model predicts if the word should be capitalized or not. To achieve this, we are jointly training two token-level classifiers on top of a pre-trained language model, such as BERT [7]. For the case of the English language, we choose a pre-trained model that has a 77% F1 score for the punctuation model and a 98% F1 score for the capitalization model in the Librispeech test set.

For Greek, there are no pre-trained models available, so we train our own version of the Punctuation and capitalization model. Initially, for the language model, we use a pre-trained BERT model in Greek [40]. For training data, we use the CC100-Greek Dataset [66], splitting it to train, validation and test sets with a 90-5-5 ratio. For each set, we preprocess the text, removing the punctuation and lowercasing all the letters. Then, for each sample, we generate the labels which consist of two tokens per word. The first shows the necessary punctuation after the word (.,;) or 0 if no punctuation is needed. The second indicates if the first letter needs to be capitalized with U, 0 otherwise. For each word the labels are the following 00, .0, ; 0, , 00U, .U, ; U, , U. For

Table 6 Punctuation and capitalization model results

Classifier	Recall (%)	Precision (%)	F1 (%)
Capitalization	96.03	96.62	96.32
Punctuation	75.18	79.71	77.37

example, the text “Τι χάνεις;” is transformed to the training sample pair “τι χάνεις” → “0U;0”. For training, we used the Adam optimizer, which optimizes the categorical cross-entropy loss and the batch size contained 2048 tokens. We trained the model for 20 epochs, and we keep the model with the smaller validation loss. The results of the punctuation and capitalization model in the test set are shown in Table 6. We notice that the capitalization classifier is better than the punctuation classifier. This is justified because the latter predicts four different classes (comma, period, question mark, and no-punctuation) in contrast with the former which predicts only two (capital, no-capital). Comparing the results with the English pre-trained model, we consider this model to be adequate for our use case.

6.2.2 ASR model inference

During training ASR models we use short segments of audio and aligned transcriptions as training samples, but in real-world applications, audio has no duration limitations. An ASR system has to be able to transcribe audio regardless of its duration. The problem is that running model inference on long audio files is usually restricted by the available system resources like the GPU memory (or RAM if GPU is not available). This restriction dictates the maximum length of audio that can be transcribed in one inference call. For example, Conformer-CTC models, use a combination of self-attention and convolution layers to achieve the best of the two methods, the self-attention layers can capture global interactions while the convolutions efficiently capture the local correlations. But the use of self-attention layers comes with a cost of increased memory usage at a quadratic rate with the input length. This means that transcribing long audio files with such models, is not working with traditional inference approaches.

To solve this issue, we split the audio into consecutive smaller chunks and run inference on each chunk, but we also care for audio context at either edge for more accurate transcription. To achieve this we use buffers and chunks. Buffer size is the length of audio on which model inference is run and chunk size is the length of new audio that is added to the buffer. An audio buffer is made up of a chunk of audio with some padded audio from the previous chunk. In order to make the best predictions with enough context for the beginning and end portions of the buffer, we only collect tokens for the middle portion of the buffer of length equal to the size

Table 7 Greek ASR system inference times

Duration	10 min		30 min	
	Yes	No	Yes	No
GPU				
Text to speech	4 s	67 s	11 s	179 s
LM	4 s	4 s	19 s	19 s
Diarization	4 s	47 s	10 s	162 s
Punct and cap	3 s	32 s	8 s	124 s
Total time	15 s	150 s	48 s	484 s

of each chunk. For our models, we use 25 s as the buffer size and use 10 s as the chunk size, so one hour of audio is broken into 360 chunks of 10 s each. Note, that inference with buffers is performed in batches. Depending on the model, hardware specifications, and inference time requirements we can use different batch sizes. For the Greek model, we use a batch size of 16. For the English version, we use a batch size of 8 due to the fact that the model requires more computer resources.

6.2.3 Performance and required resources

The minimum requirements for the automatic speech recognition system of the Greek language are at least 6GB of RAM, some disk space to store intermediate and final files, and a processor. The speed of editing and converting the audio file into text depends a lot on the speed and performance of the processor. Besides the processor, many procedures can be accelerated with a graphics card (GPU) which supports the CUDA parallel processing platform and has at least 5GB VRAM for the Greek speech recognition system. Table 7 shows some indicative completion times for various processes of the system. Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz and NVIDIA GeForce RTX 2060 SUPER graphics card were used to extract the times. The duration of the audio files is 10 and 30 min and contains continuous speech throughout the file. For each of these files, we conducted 2 experiments, where in the first case we used the graphics card to record the times and in the second case we do not.

It is apparent that using the GPU speeds up the speech-to-text, speaker grouping, and gender identification processes, which become more than 92% faster. Also, there is a speedup in the punctuation and capitalization process. This is because these processes are using deep learning models and can run very efficiently on the GPU. In contrast, the n-gram language model decoding process is CPU compatible only and as result, there is no GPU speedup.

The only difference between the English and Greek speech recognition model lies in the speech-to-text conversion process. We use a Conformer based on the RNN-Transducer architecture for the former and a CTC-based Conformer model for the latter. The Greek model has 120 million param-

eters, while the English model has 600 million and as a result, it requires more time and resources to process speech. More specifically, the minimum requirements for the English speech recognition system require at least 12GB of RAM, and 11GB of GPU VRAM. In a similar experimental setup to the Greek model, the English model processes an audio file of 10 min in 203.33 s without GPU and 26.18 s with GPU. The processing times are 610 and 78.54 s, respectively, for a 30-minute audio file.

7 User interface features

The user interface of the OAVA platform supports an abundance of features, like advanced search, auto-suggestions, sorting the results, related results, free text search, FAQs, and various searching capabilities. Advanced search for content with various filters, provides the ability to search for specific content by title, date, provider, subject section, license type, language, content type and creator. Auto-suggestions is a feature that automatically provides suggestions when composing a query and as a result encourages the user to discover new content, while it limits the scope of the search exclusively to available results and corrects spelling mistakes. Sorting the query results by relevance, ascending or descending upload date, alphabetical order of creators, alphabetical order of titles is a feature that makes navigation easier, while the suggestion of relevant content in each query result provides links between them. The feature that searches the text of a video, generated by text-to-speech models, for specific keywords and recommends available results can identify content that is indirectly related to the subject of the query. Therefore, searching the free text of a video can reveal much more information than a simple or even complex search using filters. The FAQ display page improves the user experience by providing information about the site and handing over suggestions for searching and analyzing content. In the end, the search bar support features like wildcard searches, fuzzy searches, proximity searches, range searches, boosting a term and boolean operators. Figure 3 illustrates an example that utilizes searching with boolean operators and providing results based on keywords detected in the text. The user interface also displays the automatically transcribed text of the media with some additional metadata.

8 Conclusion and discussion

As audiovisual content increases, aggregation services will become more popular and necessary. Existing aggregation services, such as Europeana, and the Digital Public Library of America, collaborate with providers who prepare their metadata for ingestion into the aggregation services' models

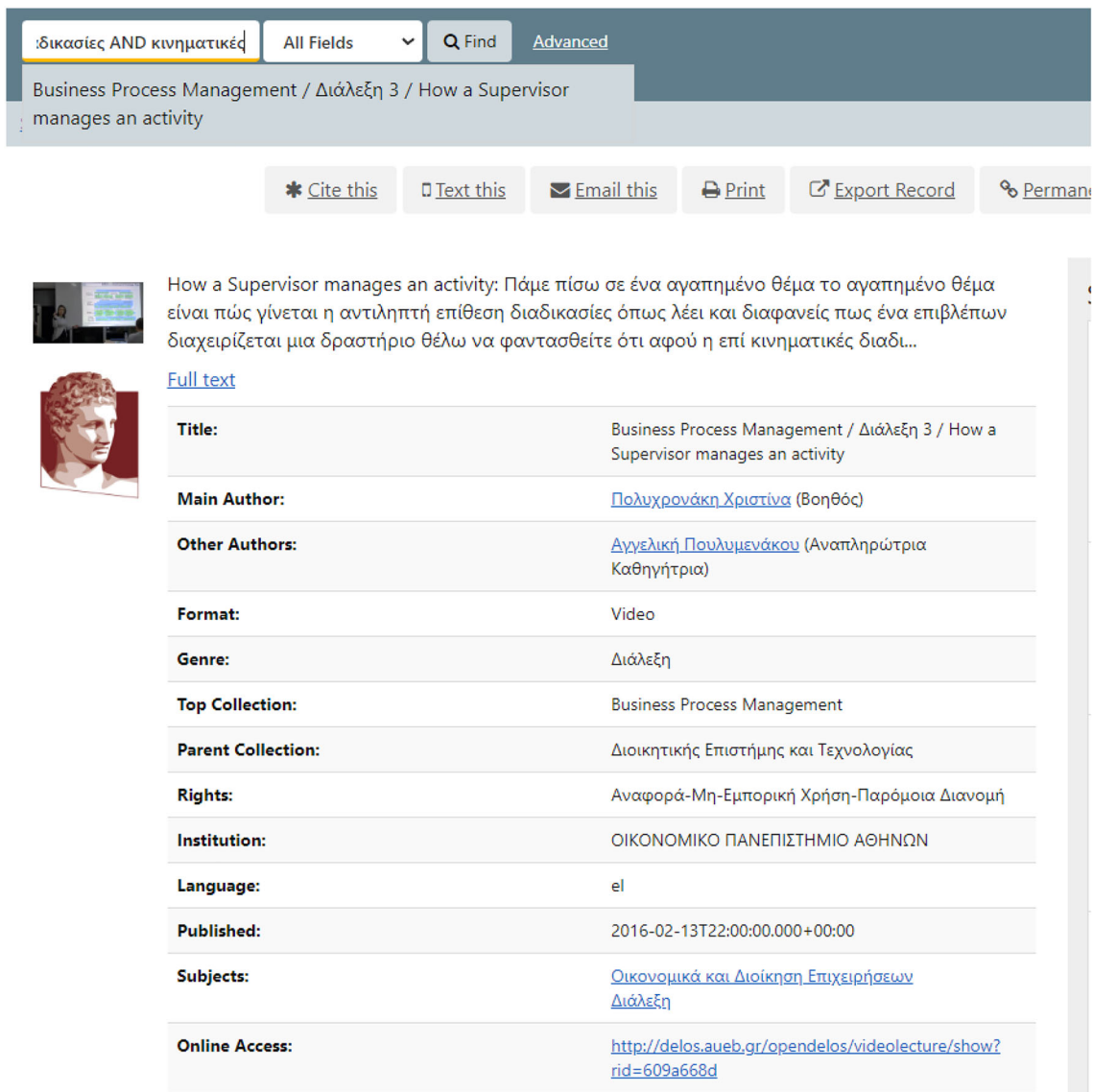
or schemas. The OAVA aggregation service aims to aggregate and enrich open-access audiovisual content but technical cooperation with the providers is not guaranteed, as is the case with the aforementioned aggregation services. Even in the few cases where interoperability standards were used (i.e., the OAI-PMH protocol) the services were not consistent, and the metadata provided was much less than expected. This finding is in accordance with the results in [37, 64]. Thus, for the OAVA project, the full extent of metadata was aggregated through crawling mechanisms and through the analysis of the profiles of existing providers. Provider profiling was necessary based on the results of [45] revealing that audiovisual providers in Greece use local metadata profiles or the minimum set of metadata imposed by the hosting streaming service, e.g., YouTube, Vimeo.

To ensure that the user will be able to access and search audiovisual resources from a single point in the same way, regardless of the provider and of the metadata schemes, the platform combines those schemes into one common enriched schema with local elements and value vocabularies. The metadata is also enriched by the OAVA - Automatic Speech Recognition process enhancing the findability of the aggregated audiovisual resources. The OAVA platform will export aggregated and enriched metadata in RDF using a subset of EBUCore classes and properties. Thus, the providers' content and metadata can be further disseminated in an interoperable linked open data format and be harnessed in new contexts. By using a subset of EBUCore, the OAVA model can be easily understood and utilized. The use of RDF enables formerly unstructured metadata to be semantically represented and consumed by third parties.

As cooperation with Greek audiovisual providers is not yet possible, the provider profiling was a necessary step. The aggregation and mapping of metadata from multiple providers using different application profiles was a challenging task. Considering that the aggregation is implemented through crawling and content harvesting, an additional challenge is that the providers often change the elements' labels or the elements themselves. Such changes impose the regular update of each provider's profile, and the adjustment of the mapping and the aggregation processes.

A limitation regarding the enrichment process is that additional values are assigned to the resources in a semi-automated way during the aggregation. As an example, in a provider's profile the OAVA team assigns the proper genre to a collection. During the aggregation, the selected genre is automatically assigned to all resources belonging to that collection. While in most cases the audiovisual resources belonging to the same collection are of the same genre, this may not be true for all collections.

To utilize audiovisual resources the OAVA platform makes use of the latest state-of-the-art automatic speech recognition technologies. OAVA platform currently supports Greek



δικασίες AND κινηματικέ All Fields Find Advanced

Business Process Management / Διάλεξη 3 / How a Supervisor manages an activity

Cite this Text this Email this Print Export Record Permanent

How a Supervisor manages an activity: Πάμε πίσω σε ένα αγαπημένο θέμα το αγαπημένο θέμα είναι πώς γίνεται η αντιληπτή επίθεση διαδικασίες όπως λείει και διαφανείς πως ένα επιβλέπων διαχειρίζεται μια δραστήριο θέλω να φαντασθείτε ότι αφού η επί κινηματικές διαδι...

[Full text](#)

Title:	Business Process Management / Διάλεξη 3 / How a Supervisor manages an activity
Main Author:	Πολυχρονάκη Χριστίνα (Βοηθός)
Other Authors:	Αγγελική Πουλυμενάκου (Αναπληρώτρια Καθηγήτρια)
Format:	Video
Genre:	Διάλεξη
Top Collection:	Business Process Management
Parent Collection:	Διοικητικής Επιστήμης και Τεχνολογίας
Rights:	Αναφορά-Μη-Εμπορική Χρήση-Παρόμοια Διανομή
Institution:	ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
Language:	el
Published:	2016-02-13T22:00:00.000+00:00
Subjects:	Οικονομικά και Διοίκηση Επιχειρήσεων Διάλεξη
Online Access:	http://delos.aueb.gr/opendelos/videolecture/show?rid=609a668d

Fig. 3 User interface of the OAVA platform

and English speech-to-text models. Specifically for Greek, we will continue to expand and improve the model mainly by expanding the training dataset which is very small compared to other models from more popular languages. Also, we will try to better detect and mitigate the underlying bias that might be present in the dataset. Among future plans is to experiment with the transcribed text and AI software for further metadata enrichment, like automatic classification and subject indexing.

The OAVA platform is expected to provide a unified way of searching Greek audiovisual content to serve teaching, research, and cultural actions. The aggregated and enriched metadata will improve users' search and retrieval tasks, and hopefully will be utilized in new contexts due to its pub-

lication as linked data. The OAVA platform is capable of exporting audiovisual resource descriptions as linked data. This linked data can be seen as a valuable by-product, which can be further processed and enriched (e.g., deduplication, entity matching). It is also anticipated that the OAVA platform will highlight the advantages of utilizing structured metadata, and will prompt Greek audiovisual providers to the adoption of the OAVA model.

Acknowledgements This work was supported by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE [project code: T2EDK-00526].

Funding Open access funding provided by HEAL-Link Greece.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ardila, R., Branson, M., Davis, K., et al.: Common voice: a massively-multilingual speech corpus (2019). arXiv preprint [arXiv:1912.06670](https://arxiv.org/abs/1912.06670)
- Barry, M., Sifton, D.: Towards a cross-Canadian digital library platform. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, pp 1–2 (2017)
- Bashir, B., Nasreen, N., Loan, F.A.: National digital library of India: an overview. *Library Philosophy and Practice* (e-journal) (2019). <https://digitalcommons.unl.edu/libphilprac/2601> (visited April 28, 2020)
- CIDOC (n.d.) Cidoc crm scope. <https://www.cidoc-crm.org/scope>. Last accessed on 2023-04-02
- Cieri, C., Miller, D., Walker, K.: The fisher corpus: a resource for the next generations of speech-to-text. In: LREC, pp 69–71 (2004)
- Deng, L., Li, X.: Machine learning paradigms for speech recognition: an overview. *IEEE Trans. Audio Speech Lang. Process.* **21**(5), 1060–1089 (2013). <https://doi.org/10.1109/TASL.2013.2244083>
- Devlin, J., Chang, M.W., Lee, K., et al.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
- DigitalNZ (n.d.a) About digitalnz. <https://digitalnz.org/about>. Last accessed on 2022-09-11
- DigitalNZ (n.d.b) Digitalnz metadata dictionary. https://docs.google.com/document/pub?id=1Z31_ckQWjnQQ4SzpORbClcIXUeO-Jd4jt-oZFuMcoQ. Last accessed on 2022-09-11
- DigitalNZ (n.d.c) Supplejack. <https://digitalnz.org/developers/supplejack>. Last accessed on 2022-09-11
- DPLA (n.d.a) About us from <https://dp.la/about>. Last accessed on 2022-09-11
- DPLA (n.d.b) Collection development guidelines. <https://pro.dp.la/hubs/collection-development-guidelines>. Last accessed on 2022-09-11
- DPLA (n.d.c) Metadata application profile. <https://pro.dp.la/hubs/metadata-application-profile>. Last accessed on 2022-09-11
- DPLA (n.d.d) Our hubs. <https://pro.dp.la/hubs/our-hubs>. Last accessed on 2022-09-11
- Drude, L., Heitkaemper, J., Boeddeker, C., et al.: Sms-wsj: database, performance measures, and baseline recipe for multi-channel source separation and recognition (2019). arXiv preprint [arXiv:1910.13934](https://arxiv.org/abs/1910.13934)
- EBU (n.d.) Tech 3293 ebu core metadata set (ebucore): specification v. 1.10., p. 7. <https://tech.ebu.ch/docs/tech/tech3293.pdf>. Last accessed on 2022-09-11
- Europeana (n.d.a) About. <https://www.europeana.eu/en/about-us>. Last accessed on 2022-09-11
- Europeana (n.d.b) Europeana data model. <https://pro.europeana.eu/page/edm-documentation>. Last accessed on 2022-09-11
- Europeana (n.d.c) Europeana semantic elements documentation. <https://pro.europeana.eu/page/ese-documentation>. Last accessed on 2022-09-11
- Europeana (n.d.d) For developers. <https://www.europeana.eu/en/for-developers>. Last accessed on 2022-09-11
- Europeana (n.d.e) For developers. <https://pro.europeana.eu/page/apis>. Last accessed on 2022-09-11
- Europeana (n.d.f) Metadata. <https://pro.europeana.eu/share-your-data/metadata>. Last accessed on 2022-09-11
- Fernie, K., Gavriliu, D., Angelis, S.: The carare metadata schema, v. 2.0. Europeana Carare project (2013)
- Frosini, L., Bardi, A., Manghi, P., et al.: An aggregation framework for digital humanities infrastructures: the parthenos experience. *Sci. Res. Inf. Technol.* **8**(1), 33–50 (2018)
- Galvez, D., Diamos, G., Ciro, J., et al.: The people's speech: a large-scale diverse English speech recognition dataset for commercial usage (2021). arXiv preprint [arXiv:2111.09344](https://arxiv.org/abs/2111.09344)
- GitHub (n.d.) Supplejack. <https://digitalnz.github.io/supplejack/>. Last accessed on 2022-09-11
- Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: telephone speech corpus for research and development. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE Computer Society, pp. 517–520 (1992)
- Graves, A.: Sequence transduction with recurrent neural networks (2012). arXiv preprint [arXiv:1211.3711](https://arxiv.org/abs/1211.3711)
- Graves, A., Fernandez, S., Gomez, F., et al.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning, pp. 369–376 (2006)
- Gregory, L., Williams, S.: On being a hub: some details behind providing metadata for the digital public library of America. *D-Lib Mag.* **20**(7/8), 25–32 (2014)
- Gulati, A., Qin, J., Chiu, C.C., et al.: Conformer: convolution-augmented transformer for speech recognition (2020). arXiv preprint [arXiv:2005.08100](https://arxiv.org/abs/2005.08100)
- Han, W., Zhang, Z., Zhang, Y., et al.: Contextnet: improving convolutional neural networks for automatic speech recognition with global context (2020). arXiv preprint [arXiv:2005.03191](https://arxiv.org/abs/2005.03191)
- Heafield, K.: Kenlm: faster and smaller language model queries. In: Proceedings of the sixth workshop on statistical machine translation, pp. 187–197 (2011)
- Holley, R.: Extending the scope of trove: addition of e-resources subscribed to by australian libraries. *D-Lib* 17(11/12) (2011)
- Holley, R.: Resource sharing in Australia: find and get in trove-making “getting” better (2011)
- Iranzo-Sánchez, J., Silvestre-Cerda, J.A., Jorge, J., et al.: Europarl-st: a multilingual corpus for speech translation of parliamentary debates. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 8229–8233 (2020)
- Kapidakis, S.: When a metadata provider task is successful. In: International Conference on Theory and Practice of Digital Libraries. Springer, pp. 544–552 (2017)
- Klijn, E., De Lusenet, Y.: Tracking the reel world. A survey of audiovisual collections in Europe. Amsterdam, training for audio-visual preservation in Europe (2008)
- Koluguri, N.R., Li, J., Lavrukhin, V., et al.: Speakernet: 1d depth-wise separable convolutional network for text-independent speaker recognition and verification (2020). arXiv preprint [arXiv:2010.12653](https://arxiv.org/abs/2010.12653)
- Koutsikakis, J., Chalkidis, I., Malakasiotis, P., et al.: Greek-bert: the Greeks visiting sesame street. In: 11th Hellenic Conference on Artificial Intelligence, pp. 110–117 (2020)
- Kriman, S., Beliaev, S., Ginsburg, B., et al.: Quartznet: deep automatic speech recognition with 1d time-channel separable convolutions. In: ICASSP 2020-2020 IEEE International Conference

- on Acoustics, Speech and Signal Processing (ICASSP).IEEE, pp. 6124–6128 (2020)
42. Kudo, T., Richardson, J.: Sentencepiece: a simple and language independent subword tokenizer and detokenizer for neural text processing (2018). arXiv preprint [arXiv:1808.06226](https://arxiv.org/abs/1808.06226)
 43. Li, J., Lavrukhin, V., Ginsburg, B., et al.: Jasper: an end-to-end convolutional neural acoustic model (2019). arXiv preprint [arXiv:1904.03288](https://arxiv.org/abs/1904.03288)
 44. Majumdar, S., Balam, J., Hrinchuk, O., et al.: Citrinet: closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition (2021). arXiv preprint [arXiv:2104.01721](https://arxiv.org/abs/2104.01721)
 45. Malliari, A., Nitsos, I., Zapounidou, S., et al.: Mapping audiovisual content providers and resources in Greece. *Int. J. Digit. Lib.* **23**, 217–227 (2022)
 46. Manghi, P., Artini, M., Atzori, C., et al.: The d-net software toolkit: a framework for the realization, maintenance, and operation of aggregative infrastructures. *Program* **48**(4), 322–354 (2014)
 47. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset (2017). arXiv preprint [arXiv:1706.08612](https://arxiv.org/abs/1706.08612)
 48. NCAMC (2018-2021) National centre of audiovisual media and communication. Who we are: establishment and mission. <https://www.ekome.media/who-we-are/>. Last accessed on 2023-04-02
 49. NDLI (n.d.a) Faq. <https://ndl.iitkgp.ac.in/faq>. Last accessed on 2022-09-11
 50. NDLI (n.d.b) National digital library of India. <https://ndl.iitkgp.ac.in/>. Last accessed on 2022-09-11
 51. NLA (n.d.a) Technical ecosystem. <https://trove.nla.gov.au/about/what-trove/technical-ecosystem>. Last accessed on 2022-09-11
 52. NLA (n.d.b) Technical specifications. <https://trove.nla.gov.au/technical-specifications>. Last accessed on 2022-09-11
 53. NLA (n.d.c) Trove content. <https://trove.nla.gov.au/about/what-trove/trove-content>. Last accessed on 2022-09-11
 54. NLA (n.d.d) Trove data dictionary. <https://trove.nla.gov.au/partners/partner-services/adding-collections-trove/trove-data-dictionary>. Last accessed on 2022-09-11
 55. NTUA (2006-2014) Introduction to mint. http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Introduction_to_MINT. Last accessed on 2023-04-02
 56. NTUA (2006-2014) Projects using mint. http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Projects_using_Mint. Last accessed on 2023-04-02
 57. Oesterlen, E.M.: (n.d.) Aggregation handbook, 3rd edition. <https://tech.ebu.ch/docs/tech/tech3293.pdf>. Last accessed on 2022-09-11
 58. Panayotov, V., Chen, G., Povey, D., et al.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5206–5210 (2015)
 59. Pitschmann, L.: (n.d.) Building sustainable collections of free third-party web resources. https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub98_57d70f70b208f.pdf. Last accessed on 2022-09-11
 60. Pratap, V., Xu, Q., Sriram, A., et al.: Mls: a large-scale multilingual dataset for speech research (2020). arXiv preprint [arXiv:2012.03411](https://arxiv.org/abs/2012.03411)
 61. Purday, J.: Think culture: Europeana. Eu from concept to construction (2009)
 62. Scholz, H.: (n.d.) Europeana publishing guide v1.8: a guide to the metadata and content requirements for data partners publishing material in Europeana collections. <https://europeana.atlassian.net/wiki/spaces/EF/pages/2059763713/EPF+-+Publishing+guidelines>. Last accessed on 2022-09-11
 63. Shibata, Y., Kida, T., Fukamachi, S., et al.: Byte pair encoding: a text compression scheme that accelerates pattern matching (1999)
 64. Togia, A., Koseoglou, E., Zapounidou, S., et al.: Open access infrastructure in Greece: current status, challenges and perspectives. *ELPUB* **2018**, 1–21 (2018)
 65. Wang, C., Riviere, M., Lee, A., et al.: Voxpopuli: a large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation (2021). arXiv preprint [arXiv:2101.00390](https://arxiv.org/abs/2101.00390)
 66. Wenzek, G., Lachaux, M.A., Conneau, A., et al.: Ccnet: extracting high quality monolingual datasets from web crawl data (2019). arXiv preprint [arXiv:1911.00359](https://arxiv.org/abs/1911.00359)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.