

## Journal of Integrated Information Management

Vol 7, No 1 (2022)

Jan-June 2022



### Managing digital archival collections by integrating Archivematica, AtoM and VuFind software: the University of Thessaly Historical Archive case

*Ioannis Clapsopoulos, Ioanna Laliotou, Stavros Doropoulos, Apostolos Fanakis, Vassilis Bourdakis, Yannis Stoyannidis, Ioulia Pentazou, George Kalaouzis, Theodore Selitsaniotis*

Copyright © 2022



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0](https://creativecommons.org/licenses/by-nc/4.0/).

#### To cite this article:

Clapsopoulos, I., Laliotou, I., Doropoulos, S., Fanakis, A., Bourdakis, V., Stoyannidis, Y., Pentazou, I., Kalaouzis, G., & Selitsaniotis, T. (2024). Managing digital archival collections by integrating Archivematica, AtoM and VuFind software: the University of Thessaly Historical Archive case. *Journal of Integrated Information Management*, 7(1), 7–18. Retrieved from <https://ejournals.epublishing.ekt.gr/index.php/jiim/article/view/37900>

# Managing digital archival collections by integrating Archivematica, AtoM and VuFind software: the University of Thessaly Historical Archive case

<sup>1</sup>Ioannis Clapsopoulos, <sup>1</sup>Ioanna Laliotou, <sup>2</sup>Stavros Doropoulos, <sup>2</sup>Apostolos Fanakis, <sup>1</sup>Vassilis Bourdakis, <sup>3</sup>Yannis Stoyannidis, <sup>1</sup>Ioulia Pentazou, <sup>1</sup>George Kalauzis and <sup>1</sup>Theodore Selitsaniotis  
<sup>1</sup>University of Thessaly, <sup>2</sup>DataScouting, <sup>3</sup>University of West Attica  
clib@uth.gr, laliotou@uth.gr, doro@datascouting.com, apostolof@datascouting.com, vas@uth.gr, ystoyannidis@uniwa.gr, pentazou@uth.gr, gkala@uth.gr, theosel@uth.gr

## Article Info

### Article history:

Received 10 April 2022

Received in revised form 15 May 2022

Accepted 30 May 2022

<https://doi.org/10.26265/ijim.v7i1.4506>

### Abstract:

**Purpose** – The great number of documents and records produced by universities leads to large archival collections within these Institutions. As a result, libraries and archival units apply new and innovative information systems to manage and exploit their archival collections for education and research purposes.

**Design/methodology/approach** – The current research took place under the THE.ME.DO.COM EPAnEK (2014-2020) project, which mainly aimed to develop two web platforms for the preservation, documentation, management, and dissemination of the University of Thessaly archival collections, which constituted its Historical Archive. For this purpose, the Archivematica, AtoM and VuFind open-source information systems were configured and integrated, while innovative metadata creation coding was created for the project.

**Findings** – Digital archival items were preserved and enriched with high-quality metadata in Archivematica, uploaded and described according to international standards in AtoM and distributed with smart search routines with VuFind. The integrated platform implementation resulted in an efficient automated approach to managing the University of Thessaly archives and archival collections, which comprised different digital object formats and included a high volume of digital items, making THE.ME.DO.COM a very successful project example.

**Originality/value** – The University of Thessaly Historical Archive was used as a paradigm for implementing a new way to deal with complex archival collections. Other University Libraries or Archival Services could efficiently utilise the project results.

**Index Terms** – Archives, Open-Source Systems, Digital Preservation, Documentation, Text and Data Mining, Thessaly

## I. INTRODUCTION

The rapid developments in digital technologies, which took place during the last few decades, have profoundly affected how digital collections are managed by the library and archive services worldwide. These changes have renewed the interest in how archival collections are used for practice, research and education in many scientific fields, including historical, cultural, and literary studies leading to an "archival turn in various disciplines" [1,2,3,4,5].

Universities produce a vast number of documents (administrative, research, event, cultural and other) during their daily operation, resulting respectively in large, constantly augmented archival collections. Furthermore, they may acquire additional, either personal or organisational, archives by various processes (purchasing, donations etc.). These collections are managed by University Libraries or by separate archival administrative units to be preserved and made available for education and research [5,6].

In 2018 the University of Thessaly Senate established the University's Historical Archive (HA) as a Central Library Department with a mandate of collecting, recording, preserving, organising, digitising, disseminating and managing the archival collections already owned or to be obtained by the University. For this purpose, the University Research Committee provided the funds for the lease and renovation of a building for housing the HA in the Tsalapata complex in Volos (Fig. 1), while the University purchased the initial technical infrastructure and equipment.

Within the same year, a research project titled "THESSALY MEMORY DOCUMENTATION AND COMMUNICATION (THE.ME.DO.COM)" was approved for funding under the

The authors would like to acknowledge funding for the current research from the single State Aid Action "RESEARCH - CREATE - INNOVATE" support measure of the Operational Programme Competitiveness,

Entrepreneurship and Innovation 2014-2020 (EPAnEK), which was co-funded by European Regional Development Fund of the European Union.

single State Aid Action "RESEARCH - CREATE - INNOVATE" support measure of the Operational Programme Competitiveness, Entrepreneurship and Innovation 2014-2020 (EPAnEK). THE.ME.DO.COM was co-funded by the European Regional Development Fund of the European Union, and the research results presented in the current paper were a part of the project's implementation. The project partners were the University of Thessaly and DataScouting (a software research and development

company). Three University units were involved in the project: the Library and Information Centre, the Department of History, Archaeology and Social Anthropology and the Department of Architecture. The primary aim of THE.ME.DO.COM was to create two new innovative integrated web-based platforms: one for preservation-management-dissemination of digital documents belonging to the University's archives and a second one for text and data mining of these documents.



**Fig. 1.** The University of Thessaly Historical Archive building (Central Library branch) in the Tsalapata complex in Volos

## II. METHODOLOGY

The University of Thessaly archives are comprised of:

- records from its administrative and academic units (Governing Committee, Senate, Rector's Office, Public Relations, Research Committee, Academic Departments, Administrative Divisions etc.),
- records from individual collections donated by external scholars or former academic staff (i.e. former rectors, professors emeriti, etc.),
- industrial records (brought in with respectively acquired local industry buildings) and
- records from previous higher education institutions, which either acceded to new academic departments or merged with the University of Thessaly.

These archival collections included documents in both physical and digital (already digitised or born-digital) formats.

The first step of the research process was to review the University's archival collections and decide which would be included in the THE.ME.DO.COM project collections. This selection was based on the fact that the relevant groups of documents were part of distinct archival collections or archives, which in turn belonged to the

University of Thessaly Historical Archive. Twelve archives from other University Institutions were comparatively studied to make that decision. These archives were selected according to the following guidelines:

- they represented typical cases coming from different continents (Europe, America, Australia)
- they were related to universities of different periods ('historic'/old and newer)

Based on the comparative study and the precondition that the respective documents should be directly or indirectly connected with the founding and operation of the University of Thessaly and the cultural heritage of the Thessaly region, the criteria for the selection of the material that would constitute the primary project collections were:

- representativeness of categories and material type or format
- emphasis on the operation and the historical evolution of the University of Thessaly
- physical document preservation condition
- the relative ease of access to the respective documents

Taking into account the criteria mentioned above, the

following archival collections were selected for digitisation, archival documentation and description:

- **Kitsos Makris Personal Archive** (personal archive of the renowned Greek folklorist K. Markis).
- **Professor Pandelis Lazaridis Personal Archive** (archival collection of the first president of the Governing Committee and first elected rector of the University of Thessaly).
- **Matsaggos Tobacco Industry Archive** (archive of one of the largest Greek tobacco industries, founded in the late 19th century and operating in Volos until 1971).
- **University of Thessaly Governing Committee Archive** (initial University governing body which operated from 1984 to 1999).

Additionally, during the second phase of the project's implementation, and following the merging of the University of Thessaly with the TEI of Thessaly and the TEI of Lamia, it was also decided to include the **Archive of the Pedagogical Academy of Lamia** to expand the Historical Archive coverage to the region of Central Greece, following the University's recent geographical and academic expansion.

The next steps of the initial research process included the formulation of detailed specifications for:

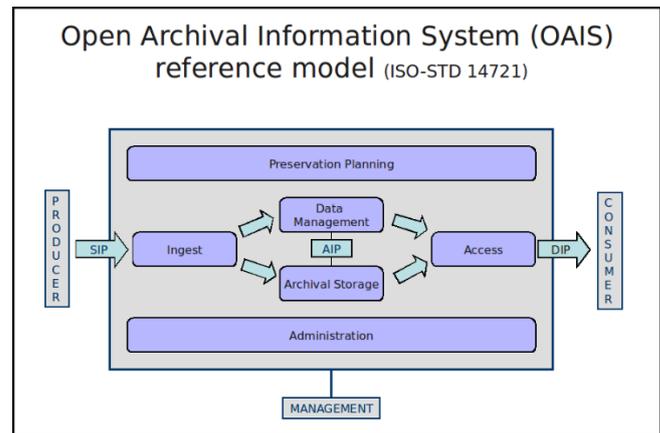
- *Digitisation and production of Optical Character Recognition (OCR) text files by document type.*
- *Archival classification and definition of levels of archival description for each collection, including instructions for the levels, extent and process of the archival description.*

Due to the large volume of the respective archival items, digitization that was carried out to representative documents of each collection has provided a sufficient number of digital documents to be utilised by the project platforms.

### III. TECHNOLOGIES AND SOFTWARE ARCHITECTURE

#### A. Digital Content Preservation Information System

The digital content preservation information system that has been chosen to meet the needs of the Historical Archive of the University of Thessaly was **Archivematica** (<https://www.archivematica.org>). This software is a free, open-source digital preservation system designed to maintain long-term access to digital memory objects. The software supports the Dublin Core and PREMIS (Preservation Metadata Maintenance Activity) metadata standards and is compatible with the ISO-OAIS (Open Archival Information System) functional model (Fig. 2). Additionally, it can be integrated with various software platforms and tools which manage digital objects, often used for archival collection management, including Access to Memory (AtoM), DSpace, CONTENTdm and ArchivesSpace [7].



**Fig. 2.** The OAIS reference model (*SIP = Submission Information Package, AIP = Archival Information Package, DIP = Dissemination Information Package*)

It can be seen in Fig. 2 that within the OAIS reference model, three kinds of information packages exist: the Submission Information Package (SIP), which is the information sent from the producer to the archive, the Archival Information Package (AIP); which is the information stored by the archive, and the Dissemination Information Package (DIP), which is the information sent to a user when requested. Archivematica is a software platform that realises the OAIS reference model.

Archivematica processes digital objects, which it transforms into Submission Information Packages (SIPs), then applies format policies and creates repository-independent Archival Information Packages (AIPs) and finally uploads Dissemination Information Packages (DIPs), containing descriptive metadata and access copies, to the selected access system of the installation (e.g., AtoM) [8].

Submission Information Packages (SIPs) always contain the digital object to be preserved and the necessary metadata about it and its content. The requirements and restrictions applied to the SIP content for each type of digital object are described in the preservation plan of each type in the Preservation Planning menu of Archivematica (Fig.3).

The Archivematica system aims to create Archive Information Packages (AIPs) based on international standards containing all the necessary information. The relative AIPs include a METS XML file with an implementation of the PREMIS retention metadata. It is pointed out that the "METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium [9]". METS focuses on relating the content of an item to the content and metadata of other digital items included in a digital library. These correlations constitute a process that takes place within repositories or between repositories. METS is a content and metadata management scheme that creates meta-metadata extracting and associating the semantic content

of digital documents [10].

The Dissemination Information Packages (DIPs) do not have a clear structure, and one or more AIPs can produce them. A DIP is a collection of digital content consisting of one or more digital files related to each other. These files may contain metadata and/or files that "bind" the individual pieces.

The process of storing a digital object in the Archivematica system to its final conversion involves multiple steps, which include:

- the conversion of the digital object into a SIP

- the subsequent conversion of SIP into AIP and DIP
- the storage of the AIP
- steps related to finding and retrieving DIPs from storage

The workflow within the system is visually shown in Fig.3. From the end-user perspective, all Archivematica functions (steps) take place within a web-based dashboard (Graphical User Interface: GUI), which can be accessed by logging in through a web browser.

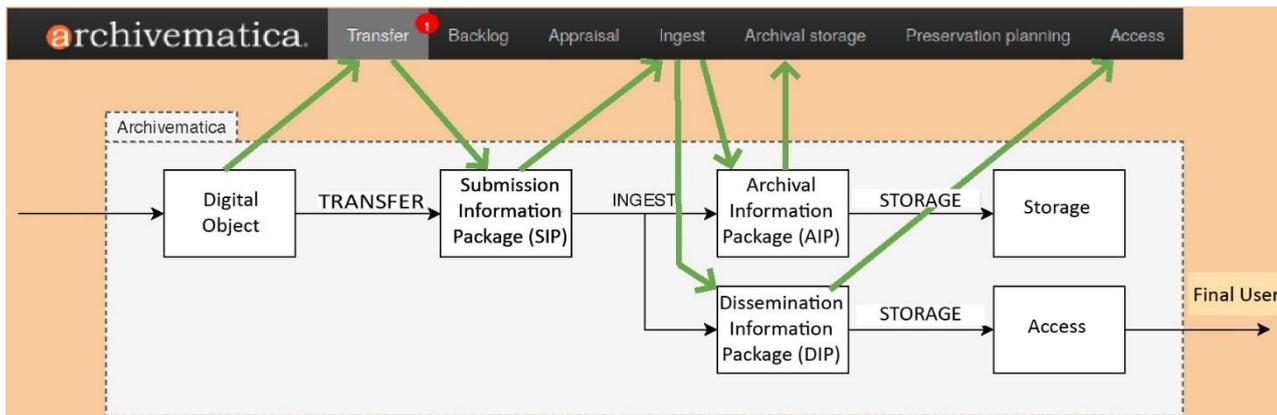


Fig. 3. Archivematica workflow

The steps listed above are directly related to the dashboard GUI. More specifically, the tabs which were used in the University of Thessaly Archivematica

implementation (digitalpreserv.lib.uth.gr) included:

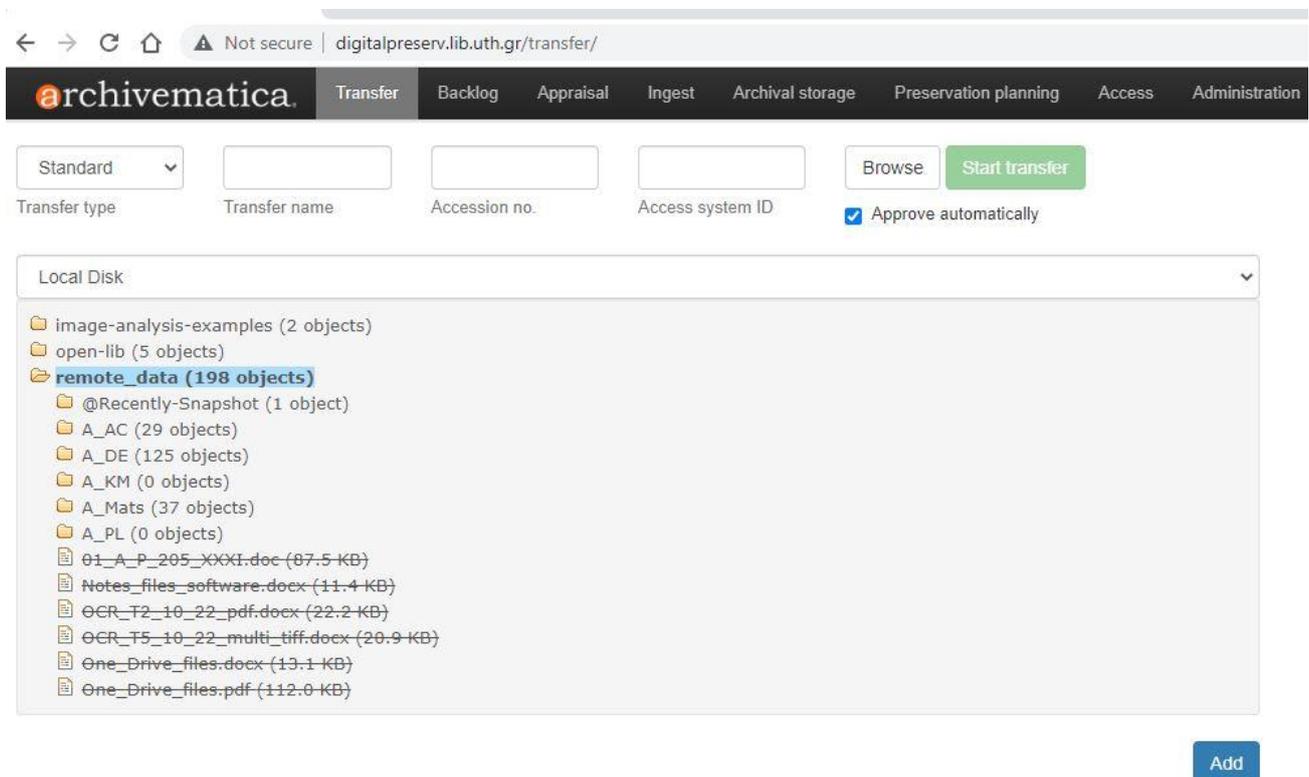


Fig.4. The Transfer tab in the University of Thessaly Archivematica implementation

- The **Transfer tab**: where users import digital objects into the system and convert them into SIPs. For the

current project, digital items could be uploaded to a virtual server hard disk (remote\_data), allowing the

end-users to upload as many items they wish per archival collection at any given period on their own without the need for intervention by system administrators. This has significantly accelerated the uploading process (Fig. 4).

- The **Ingest** tab (processing): where the results of the transport (SIP) become a package on which microservices run to convert them to AIP and generate the DIP (Fig. 5).
- The **Archival Storage** tab: where users can browse the board with archival material and perform searches
- The **Preservation planning** tab: where users can

define preservation requirements for each type of digital object. The file types supported in the system (Preservation planning tab) include groups of common file types like Audio, Database, Dataset, Image (Raster), Portable Document Format, Text (Plain), Text (Source Code), Video etc.

- The **Access** tab: during Ingest, the system can create access copies of the digital objects and package them into DIPs which will be uploaded to some access system (in this project, it was the AtOM system). In the Access tab, authorised users can see the DIPs that have been created and uploaded to the access system.

Submission Information Package	UUID	Ingest start time
Latin-english	f229aa66-1c83-49b2-a2ed-5e9a602cdcdb	2020-04-22 09:44
▶ Microservice: Normalize		
Job: Normalize [?]		Awaiting decision
Job: Resume after normalization file identification tool selected.		Completed successfully
Job: Move to select file ID tool		Completed successfully
Job: Identify file format		Completed successfully
Job: Do you want to perform file format identification?		Completed successfully
Job: Grant normalization options for no pre-existing DIP		Completed successfully
Job: Set remove preservation and access normalized files to renormalize link.		Completed successfully
Job: Check for Access directory		Completed successfully
Job: Check for Service directory		Completed successfully
Job: Identify manually normalized files		Completed successfully
▶ Microservice: Change SIP filenames		
▶ Microservice: Remove cache files		
▶ Microservice: Include default SIP processingMCP.xml		
▶ Microservice: Rename SIP directory with SIP UUID		
▶ Microservice: Verify SIP compliance		
asdf	5c7aed30-5471-493e-8b5d-7f4846313ac4	2020-04-21 08:20

Fig.5. The Ingest tab in the University of Thessaly Archivematica implementation

#### DataScouting Microservices

One of the main objectives of the current project was to develop software that would mine a text passage and extract some metadata about it. DataScouting developed specific microservices for that purpose which comprised:

- Language detection
- Semantic Analysis
- Automatic speech recognition
- Image analysis

The above microservices were added in the "Characterise and extract metadata" Archivematica microservice that runs on the Transfer step (Fig. 4). The results are saved in a metadata.csv file.

The Language detection microservice recognises the language of a text. It is used in text files (PDF) and in-text transcriptions (OCR) from images (PNG, TIF, JPG etc.). The result is the language identified as most likely for the text.

The Semantic Analysis microservice performs semantic analysis of text and is employed in text files (PDF), in-text

transcriptions (OCR) from images (PNG, TIF, JPG etc.), and in audio transcriptions from sound (MP3, WAV) or video (AVI, MP4, MKV, MOV) digital files. The result is a list of words that have a special meaning, such as persons, places, etc.

By combining the language detection and the semantic analysis microservices, the developed software finds the passage's language when it is not known (most texts are in the Greek language with some variations in English), the part of speech of each word and the nominal entities in the passage.

The automatic speech recognition microservice extracts text from sound files (MP3, WAV) or audio tracks of video files (AVI, MP4, MKV, MOV). The result is the transcribed text from the audio part of the video. For this purpose, a new tool was created at Archivematica, named ASR, while a new command named "Extract speech with ASR", which uses the above tool and defines a bash script, was added. The implemented script extracts the sound from files - video type - and executes a request to the endpoint of the previously described program, sending

the sound file. Provided that the export is successful, the script saves the text in XML format inside the METS file of the DIP. Otherwise, the script displays an error message.

The Archivematica system is hosted in the University of Thessaly Central Library infrastructure, while the installation of all the necessary applications (operating system, web servers, database servers and application servers and the OS) have been included.

The metadata produced by the DataScouting Microservices was thoroughly reviewed for different types of digital items uploaded to the Archivematica. The review showed that the microservices produced many subject terms (persons, places etc.) with a variation in success depending on the quality of the digitised item. However, it was noted that metadata terms should be reduced and subjected to quality control procedures to be fully exploited. In either case, it was concluded that they could be utilised as input in the archival management and text and data mining systems of THE.ME.DO.COM.

#### *B. Archival Management and Description Information System*

The archival management and description information system that has been chosen to support the documentation of the collections of the University of Thessaly Historical Archive was **Access to Memory (AtoM)** (<https://www.accesstomemory.org/>). The specific software is free and open source and supports the international standards of archival science, such as the standards of the International Council on Archives (ICA). The system is intended for online use with the aim of the archival description of digital objects based on international standards and access through a multilingual environment. AtoM is hosted in the University of Thessaly Central Library infrastructure, including installing all the necessary applications (operating system, web servers, database servers and application servers and the OS). The University of Thessaly installation is bilingual: the main interface is in the Greek language, in which the detailed archival descriptions are made, while selected description entries were translated into English (all the menus and field names are presented in both languages).

As in all modern information systems, AtoM consists of three levels: presentation, application, and data.

##### *Presentation Layer*

This level has direct contact with the users, regardless of their role. Access to the system was implemented via the world wide web (web) for all users.

##### *Application Layer*

This is the level at which all system services are performed and all its operating rules are implemented. The primary role of this layer is to interact with the data layer and feed the presentation layer with the appropriate information. The application layer is entirely modular (supporting different sub-systems), discretely implementing the basic functions of the system, such as importing and exporting content, indexing and storing it,

searching and navigating it., as well as the essential management functions, such as defining access rights, managing users, exporting statistics, etc.

##### *Data Layer*

This level includes the functional units that undertake the storage, organisation, classification, conversion, backup and, in general, the management of the system's content.

##### *Users*

AtoM users are classified according to their access and editing rights. In general, there are anonymous users, authenticated users and administrators. More specifically, anonymous users visit the system intending to locate archival material of interest to them and have access to the digital objects, if they exist, without the need to authenticate. Certified users are authenticated through a username and password and have various rights such as import/export, editing, etc. Finally, there is the administrator, who has almost all rights. The administrator can install/uninstall the system, import/export data, have full access to all data (read, modify, delete, import), change system preferences and appearance, and define new users and user groups.

##### *Metadata*

The AtoM archival description system supports international archival standards as it aims to standardise archival work. For the description of archival collections and the creation of finding aids, it supports the following international standards:

- ISAD(G): General International Standard Archival Description - 2nd edition.
- ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families, 2nd Edition.
- ISDIAH: International Standard for Describing Institutions with Archival Holdings.

Apart from the above International Council on Archives (ICA) standards, the system is flexible in using standards utilised by libraries, archives, etc. In addition, it supports the Dublin Core and MODS (Metadata Object Description Schema) metadata schemes.

Finally, the archival description system implements the Open Archives Initiative Protocol for Harvesting (OAI-PMH) to make it accessible to aggregators and increase its visibility and traffic.

##### *Submitting content to AtoM*

All users can log in to AtoM via its home page ([atomarchives.lib.uth.gr](http://atomarchives.lib.uth.gr)), which includes the header bar that provides access to a search box, the browse menu, the log-in button, the language menu, and the quick links menu (Fig. 6).

For users to be able to add or edit content and access the main menu in AtoM, they need to log in using the credentials supplied by the system administrator.

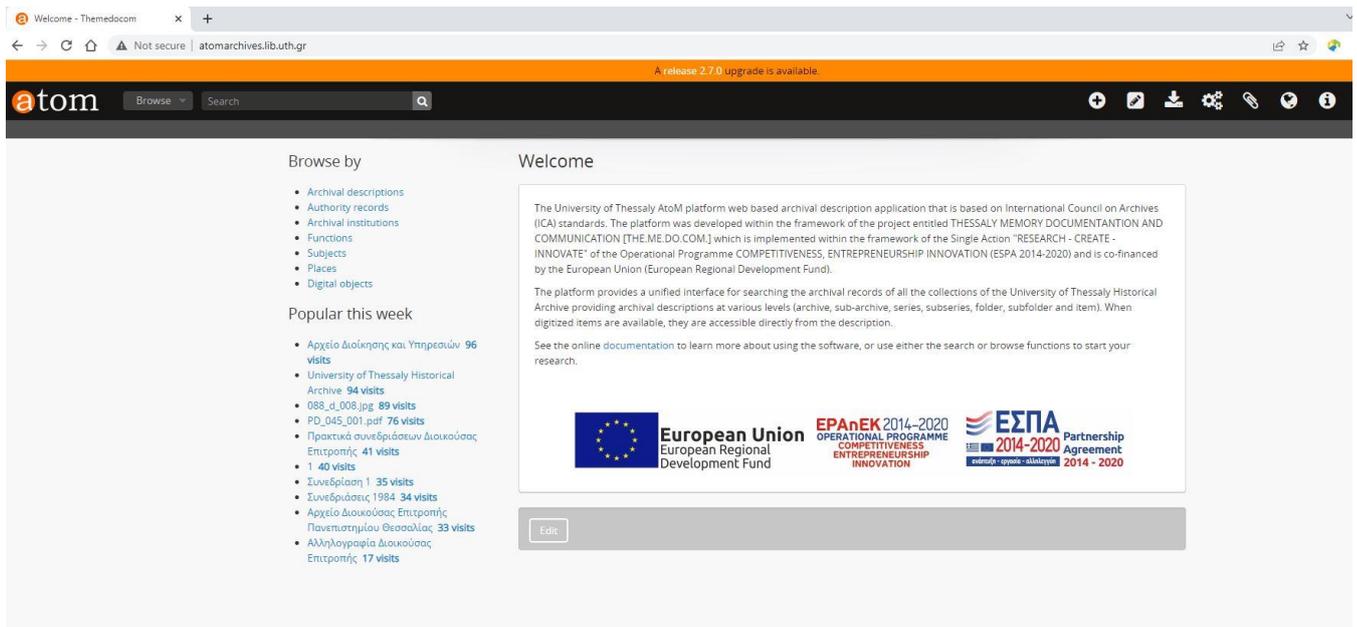


Fig.6. University of Thessaly AtoM home page (administrator view)

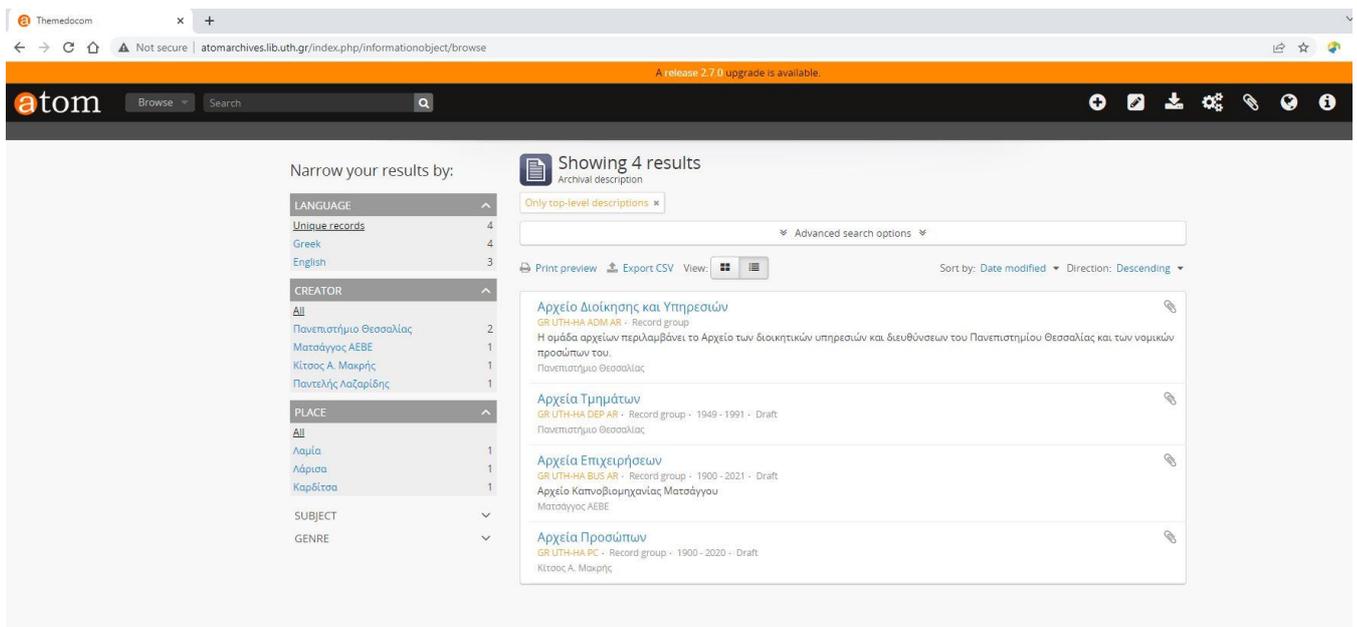


Fig.7. University of Thessaly AtoM Archives Group page

Entry of documents into the information system is done only by authorised users. Content is entered into the information system using forms based on international archival standards (ISAD (G), ISAAR (CPF), ISDIAH).

As already stated, the University of Thessaly Historical Archive (HA) comprises several archival collections. In order for the AtoM system to be able to accommodate all the archives or archival collections that may be later included in HA, the archives have been arranged in four distinct groups which are (Fig. 7):

- Archives of Administrative University Bodies and Units (currently, it contains the University of Thessaly Governing Committee Archive)
- Archives of University (academic) Departments

(currently, it contains the Archive of the Pedagogical Academy of Lamia)

- Business Archives (currently it contains the Matsaggos Tobacco Industry Archive)
- Personal Archives (Fig. 8) (currently it contains the Kitsos Makris and the Pandelis Lazaridis Archives)

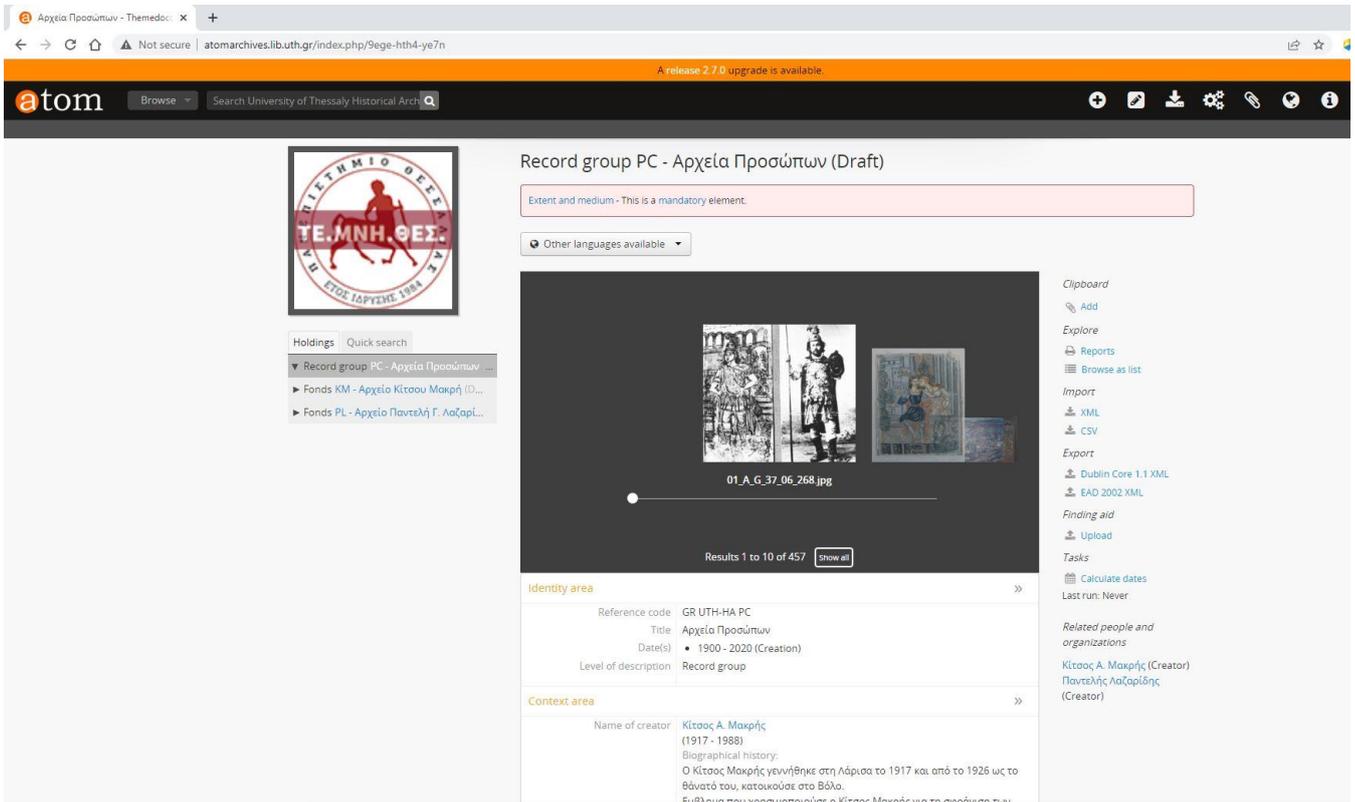


Fig.8. University of Thessaly AtoM Personal Archives Group page

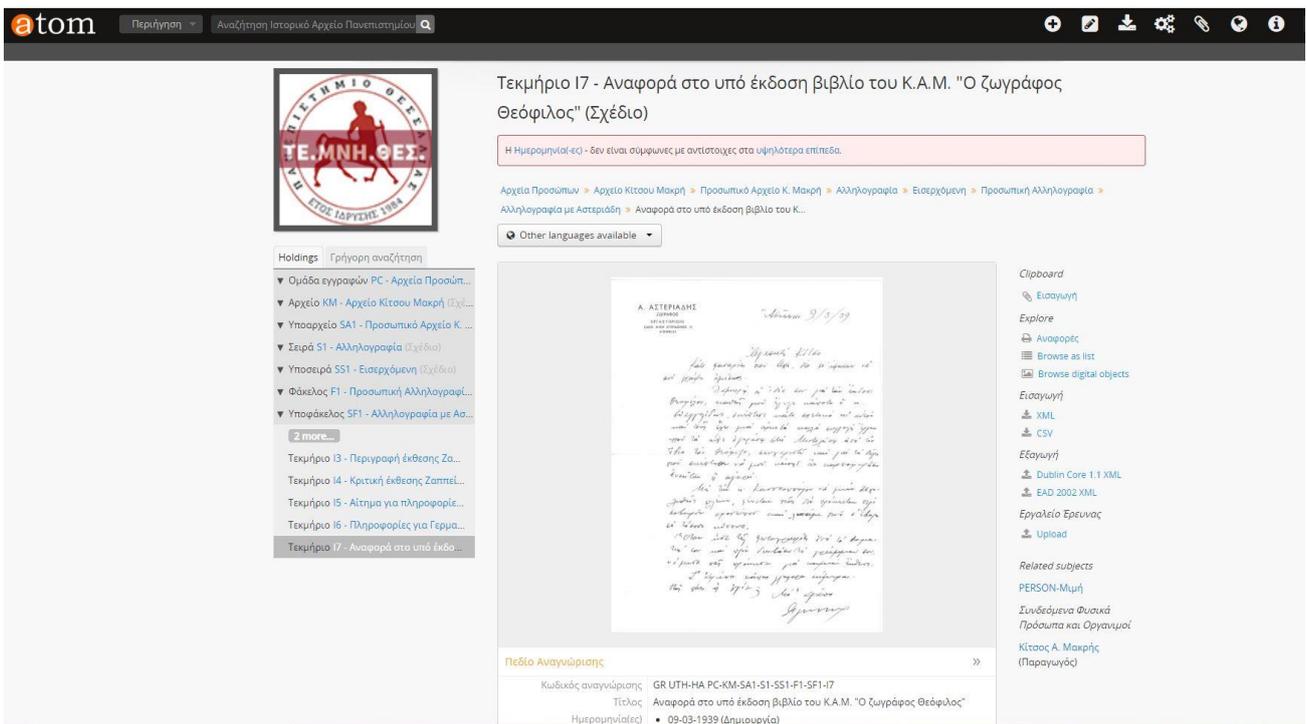


Fig. 9. University of Thessaly AtoM item description page (Greek interface)

Filling in all description fields was not mandatory. As already stated in the Methodology section, detailed instructions for the levels, extent and process of the archival description were created for the HA archives included in the AtoM platform. Description levels for the HA archives may include Fond, Subfond, Series, Subseries, File, Subfile and Item (Fig. 9). In all levels fields that were required to be filled in by the archivists were: Title and

Date (Identity Area), Name of Creator and Repository (Context area), Scope and Content (Content and structure area), and Language of the Material (Conditions of access and use area). On the item level, OCR and automatic speech recognition text produced by the DataScouting Microservices is copied in the Notes Area, while in the Access Points Area, the semantic analysis and image analysis metadata are entered automatically during the

uploading process from Archivematica to AtoM (Fig. 10).

The screenshot displays the AtoM interface for a document item. The left sidebar shows a navigation tree with 'Τεκμήριο 159 - Πρακτικά Συνεδρίασης...' selected. The main content area is divided into several sections, each with a title and a right-pointing arrow:

- Πεδίο Αναγνώρισης**: Contains fields for 'Κωδικός αναγνώρισης' (GR UTH-HA ADM AR-ADE-S1-SS5-F59-I59), 'Τίτλος' (Πρακτικά Συνεδρίασης 59), 'Ημερομηνία(ες)' (10-03-1988 (Δημιουργία)), 'Επίπεδο περιγραφής' (Τεκμήριο), and 'Μέγεθος και Υπόστρωμα' (7 σελ. χαρτιού).
- Πεδίο Πλαισίου Παραγωγής**: Contains 'Όνομα παραγωγού' (Διοικούσα Επιτροπή Πανεπιστημίου Θεσσαλίας) and 'Αποθετήριο' (Ιστορικό Αρχείο Πανεπιστημίου Θεσσαλίας).
- Πεδίο Περιεχομένου και Διάθρωσης**: Contains 'Παρουσίαση Περιεχομένου' (Περιλαμβάνει τα πρακτικά της 59ης συνεδρίασης της Διοικούσας Επιτροπής του Πανεπιστημίου Θεσσαλίας).
- Πεδίο Όρων Πρόσβασης και Χρήσης**: Contains 'Γλώσσα των Τεκμηρίων' (Ελληνικά, Ελληνικό) and 'Γραφή του υλικού' (Ελληνικά).
- Πεδίο Παρατηρήσεων**: Contains 'Σημείωση' (ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ ΔΙΟΙΚΟΥΣΑ ΕΠΙΤΡΟΠΗ ΠΡΑΚΤΙΚΟ ΕΥΝΕΑΡΙΑΕΗ 59/10.3.1988 Η Διοικούσα Επιτροπή του Πανεπιστημίου Θεσσαλίας, ύστερα από πρόσκληση του Προέδρου της Καθηγητή, κ. Παντελή Λαζαρίδη, συγκεντρώ-θηκε σήμερα 10 Μαρτίου 1988...).
- Σημεία πρόσβασης**: Contains 'Θέματα ως Σημεία πρόσβασης' (ORG-ΕΛΛΗΝΙΚΗ, ORG-ΑΗΜΟΚΡΑΤΙΑ, ORG-Πανεπιστήμιο, ORG-Θεσσαλίας, ORG-Διοικούσα, ORG-ΕΠΙΤΡΟΠΗ).

On the right side, there are options for 'Εισαγωγή' (XML, CSV), 'Εξαγωγή' (Dublin Core 1.1 XML, EAD 2002 XML), 'Εργαλείο Έρευνας' (Generate, Upload), and 'Related subjects' with a list of identifiers including ORG-ΕΛΛΗΝΙΚΗ, ORG-ΑΗΜΟΚΡΑΤΙΑ, ORG-Πανεπιστήμιο, ORG-Θεσσαλίας, ORG-Διοικούσα, ORG-ΕΠΙΤΡΟΠΗ, ORG-ΕΥΝΕΑΡΙΑΕΗ, ORG-Πανεπιστήμιο, ORG-Θεσσαλίας, PERSON-Παντελή, PERSON-Λαζαρίδη, ORG-Πανεπιστήμιο, ORG-ΕΠΙΤΡΟΠΗ, GPE-Αθήνα, GPE-Μητροπόλεως, PERSON-Παντελής, PERSON-Λαζαρίδης, PERSON-Ιωάννης, PERSON-Κουκιιάδης, PERSON-Κων, PERSON-Ακριτίδης, PERSON-Μαρία, PERSON-Ηλιού, PERSON-Μιχάλης, PERSON-Κουντούρης, PERSON-ΑΝΑΣΤΑΣΙΟΣ, PERSON-Κωτσάπουλος, PERSON-Αριστέλης, PERSON-ΑΜΠΡΟΥΛΗΣ, PERSON-Ιωάννης, PERSON-Παπακολλάου, PERSON-Εμμανουήλ, PERSON-Σταυρακίτης, PERSON-Εμμανουήλ.

Fig. 10. University of Thessaly AtoM item Access points

In addition, editing and/or deleting an existing description is possible.

At the same time, it is possible to automate the process of producing single indices of research tools per selected access point, such as natural persons, collective bodies, thematic headings etc.

Regarding bulk content import, the AtoM archive management system has import mechanisms which support widely used formats such as CSV and XML (EAD, EAC, Dublin Core XML, MODS XML) depending on the type (archive descriptions, established types, etc). It is also possible to import files in SKOS format.

The AtoM Archival Description System can extract archival descriptions, established terms, archival organisations and terms. Export can be done in widely used formats such as EAD 2002 XML, Dublin Core 1.1 XML, MODS XML (Fig. 8-10).

The export of the archival record with the respective standards is also accompanied by information about the associated established terms, the descriptions of the archival organizations and all its child records. Conventional terms are exported with EAC XML and SKOS files but through restricted access to authorised users.

The AtoM Archival Description System, in order to appeal to the general public, even to people who may not be familiar with the new technologies, employed a relatively simple graphical interface.

At the same time, the system is accessible to people with disabilities as the relevant internationally recognised

accessibility rules and guidelines concerning the development of accessible applications and services in a global web environment have been applied, in particular, the Web Content Accessibility Guidelines (WCAG) standard in order to meet at least level AA conformance.

The integration of Archivematica with AtoM serves as the platform for the preservation-management-dissemination of archival digital documents belonging to the University's archives as described in the THE.ME.DO.COM project. A similar approach of integrating Archivematica with AtoM was implemented in 2018 to organise and manage digitised materials from the Personal Fonds Rista Odavic kept by the Archives of Serbia [11]. However, this implementation utilised only the standard capabilities of Archivematica without creating additional microservices like the ones developed in the current project, and only one type of personal fond was tested and used.

### C. Text, data mining and dissemination platform

The process involved the design and implementation of an integrated software package through which textual and audiovisual documents would be enriched with machine learning and natural language processing techniques. These modules were built and incorporated as Archivematica microservices, which generate metadata from both the textual and audio-visual parts of the documents that are included in the DIP package

uploaded from Archivematica to AtoM.

The second project platform (archives.lib.uth.gr) is integrated with AtoM and is used for text and data mining purposes of digital documents (items) belonging to the University of Thessaly Historical Archive collections. The software consists of a content indexing and retrieval platform based on the open-source platform VuFind (https://vufind.org/vufind/), but also utilises a set of

modules that will allow the retrieval of additional metadata from the digital items. A special routine was created for the daily export of the metadata from all AtoM description levels (Fond, Subfond, Series, Subseries, File, Subfile and Item) in an XML file which is automatically imported into VuFind (Fig. 11-12).

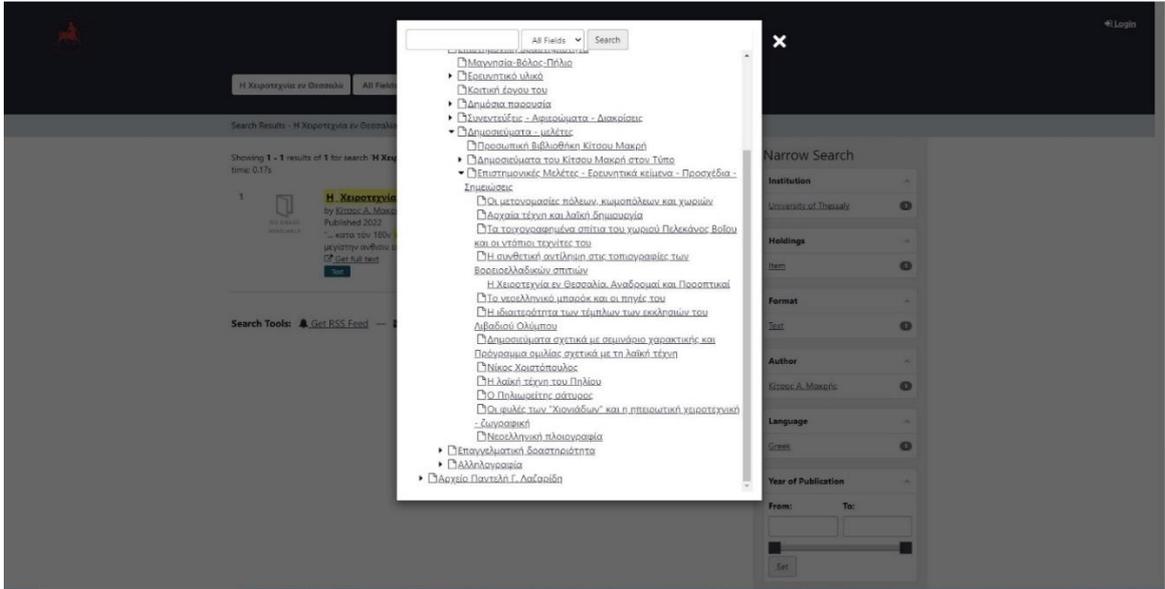


Fig. 11. Archival tree view following a thematic search on the VuFind University of Thessaly HA platform

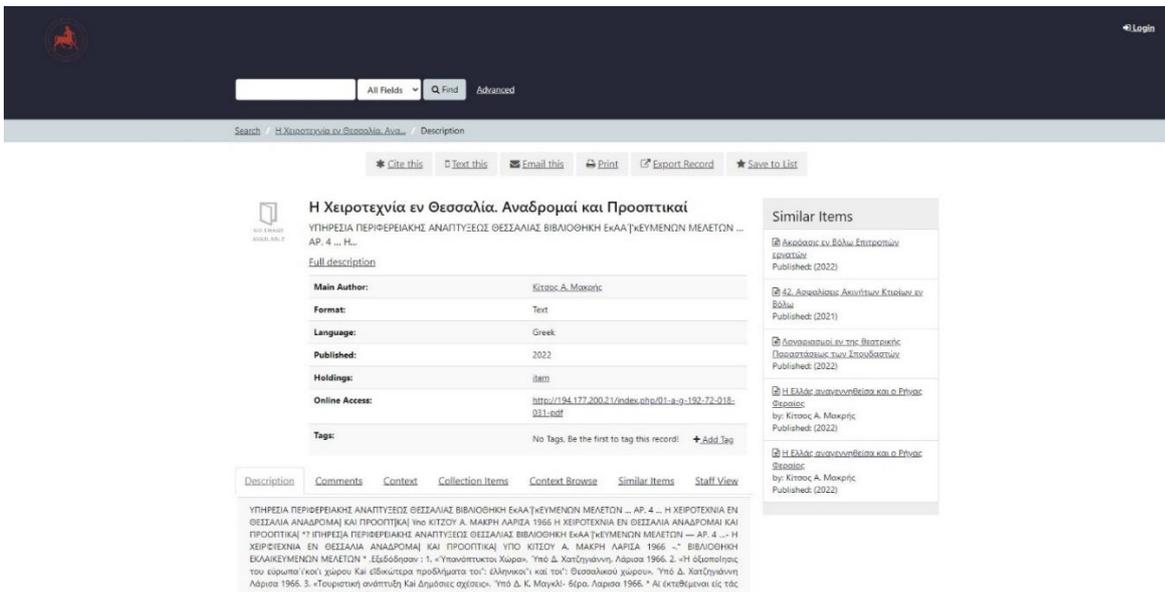


Fig. 12. Detailed item view on the VuFind University of Thessaly HA platform

The final platform will be accessible to the general public, and will be searchable through any metadata. Particular emphasis was placed on the visualisation of results through navigation tools. Additionally, the system is implementing the Open Archives Initiative Protocol for Harvesting (OAI-PMH) to make it accessible to suitable aggregators.

From the previous analysis of the University of Thessaly HA platforms, it seems more probable to utilise AtoM as an internal archival finding aid operated by specialised library staff, while the VuFind platform can be used as the

primary search portal for researchers and the general public. The relative decision will be made by the University Library after the completion of the THE.ME.DO.COM. research project.

#### IV. CONCLUSIONS

The main conclusions of developing two web platforms to preserve, manage, disseminate and analyse archival digital documents are summarised as follows:

The three open-source software applications used

(Archivematica-AtoM-VuFind) proved very successful in realising the main project goal of implementing the two above-mentioned platforms.

Archivematica was employed not only for creating preservation metadata of uploaded digital objects but also as a means of producing textual (OCR), speech-to-text and image-related metadata, and this was an innovative use of the software.

Integration between the three software applications was performed successfully and without significant problems.

The selected implementation showed that it was very effective for dealing with several archives and archival collections having divergent characteristics. The systems could also accommodate all major formats and large quantities of digital documents.

Future developments could include the production of additional digital objects metadata and the realisation of more efficient automated quality control mechanisms on descriptive metadata production.

In conclusion, the presented implementation is particularly effective for managing the diverse and voluminous University archives and archival collections.

## V. REFERENCES

- [1] Foster, H. 2004. "Regarding the 'archival turn' in various disciplines: An Archival Impulse", October 110, pp. 3–22.
- [2] Røssaak, E. (ed.). 2011. "The Archive in Motion: New Conceptions of the Archive in Contemporary Thought and New Media Practices". Studies from the National Library of Norway: Oslo, Norway.
- [3] Legg, K., Ellis, R.E and Hall, C. 2020. "Applying the seven principles of good practice: archives in the 21<sup>st</sup>-century University", Archives and Records, 41:2, 109-125, <https://doi.org/10.1080/23257962.2020.1728525>.
- [4] Colonna, S.E. and Lawrimore, E. 2019. "University archives and the slow fuse of possibility", The International Journal of Critical Pedagogy, Vol. 10, No. 1, 159-174. <https://libjournal.uncg.edu/ijcp/article/view/1517>
- [5] Fritz, A. 2018. "From collection silos to digital content hubs: digital project management in special collections and university archives", Project Management in the Library Workplace, Vol. 38, pp. 187-198. <https://doi.org/10.1108/S0732-06712018000038014>
- [6] Chrysanthopoulos, C., Drivas, I., Kouis, D., and Giannakopoulos, G. 2023. "University archives: the research road travelled and the one ahead". Global Knowledge, Memory and Communication, Vol. 72, No 1/2, pp. 44-68. <https://doi.org/10.1108/GKMC-08-2021-0128>.
- [7] Artefactual, Archivematica Development Wiki. (s. d.). Retrieved from: [https://wiki.archivematica.org/Main\\_Page](https://wiki.archivematica.org/Main_Page)
- [8] Artefactual, Archivematica Development Wiki. Documentation > Technical Architecture > Overview (s. d.). Retrieved from: <https://wiki.archivematica.org/Overview>
- [9] Library of Congress (2022). METS Metadata Encoding and Transmission Standard. Retrieved from: <http://www.loc.gov/standards/mets/>
- [10] Kyriaki-Manesi, D., and Koulouris, A. 2015. "Managing Digital Content" (in Greek, p. 94). Kallipos, Open Academic Editions. Retrieved from: <https://hdl.handle.net/11419/2496>
- [11] Bogosavljevic, M. 2018. "Digitization in archives - archivematica, the practical review of the software for

managing archival records on the example of the personal fonds Rista Odavic at the archives of Serbia". Moderna Arhivistika, 2018(1), 233-248.

## VI. AUTHORS



**Ioannis Clapsopoulos** holds a BSc in Geology from the University of Athens (Greece), a PhD from the University of Manchester (UK) and an MSc in Information & Library Management from Northumbria University (UK). His research interests include Information Literacy, Open Access to Documents and Data and the Development of Information Systems and Services relating to Libraries, Archives, and other cultural heritage organisations. He has been involved in more than 20 national and European projects and has published several articles in journals and conferences. He is the Director of the University of Thessaly Library and Information Centre, while he has also worked as adjunct lecturer at the same University.



**Ioanna Laliotou** studied History at the Department of History and Archeology of the University of Athens and holds an M.Soc. Sc. in Cultural Studies from the University of Birmingham (UK) and a PhD in History and Civilization from the European Institute in Florence. Her research interests include cultural history, subjectivity, mobility, migration, refugeeness, visions of the future, and utopia in contemporary society. At the same time, she has published books, several articles in journals, conferences and book chapters. She has participated in or coordinated several European and national projects and has published articles in journals and conferences. Currently, she is an Associate Professor in Contemporary History at the Department of History, Archaeology and Social Anthropology of the University of Thessaly.



**Stavros Doropoulos** is a graduate of the Computer Engineering Department of TEI of Central Macedonia and holds a Master's degree from the School of Informatics of the Aristotle University of Thessaloniki. He has many years of experience in both implementation and management of European and national IT R&D projects in the field of digital libraries, speech and video recognition, machine learning, big data analytics, software integration, media analysis, and he has published relevant articles in scientific journals. Currently, he is the Chief Information Officer & Software Engineer at the DataScouting software research and development company.



**Apostolos Fanakis** has participated in several R&D projects and was involved in installing and configuring the Archivematica, AtoM and VuFind platforms of the THE.ME.DO.COM EPAnEK (2014-2020) project. Currently, he works as a software engineer at the DataScouting software research and development company.



**Vassilis Bourdakis** studied Architecture at the National Technical University of Athens and completed his PhD at the University of Bath (UK) on the subject of building evaluation and performance. His research interests include Virtual Reality focusing on the communication of ideas /designs, the development of interactive virtual environments and their use as a visualisation tool for urban planning and policy-making and intelligent environments and their implications/application in real-life studies. He has participated in various European and national research projects and published over 50 articles in scientific journals and conference proceedings. Currently, he is a Professor at the Department of Architecture of the University of Thessaly.

Accreditation of Lifelong Learning Providers (EKEPIS), and he has extensive experience in managing library information systems. His interests include information systems for distance learning, digital libraries and digital repositories. Currently, he is responsible for the Integrated Library Information System (ILSAS) and the Institutional Repository of the University of Thessaly Library and Information Centre.



**Yannis Stoyannidis** studied History at the Department of History, Archaeology and Social Anthropology of the University of Thessaly, from which he received his Master of Arts (M.A.) and his PhD. His research interests include archives management, social history, urban history and the history of institutions, and he has published several articles in scientific journals and conference proceedings. He has participated in various research programs concerning social history, industrial heritage and archive management. Currently, he is an Assistant Professor at the Department of Archival, Library and Information Studies, University of West Attica.



**Ioulia Pentazou** studied history at the Aristotle University of Thessaloniki, completed her postgraduate studies at the University of Athens and received her PhD in digital design from the Department of Architecture and Engineering at the University of Thessaly. Her research interests focus on theoretical and applied aspects of digital design (visualization of historical information, big data management of historical science, digital archives, etc.). She has participated in many research projects concerning the teaching of history in digital environments, the visualization of the past and the representation of urban history in digital applications and technology, while she has published a book, book chapters and articles in journals and conferences. Currently, she is Assistant Professor at the Department of Culture, Creative Media, and Industries, University of Thessaly.



**George Kalaouzis** studied Electrical and Computer Engineering at the Aristotle University of Thessaloniki and received his M.Sc. in "State-of-the-Art Design and Analysis Methods in Industry" from the Department of Mechanical Engineering, University of Thessaly. He specialises in dynamic web development and programming in PHP, focusing on databases and interactive multimedia content development. He has participated in several research programs and published articles in conference proceedings. Currently, he works as a laboratory teaching staff at the Department of Architecture, University of Thessaly.



**Theodore Selitsaniotis** received his diploma in Informatics from the Department of Informatics of the Athens University of Economics and Business. He is a certified trainer of the National Centre for the