



ENABLING VIRTUAL COLLABORATION IN DIGITAL CULTURAL HERITAGE IN THE SEEM REGION

PANAYIOTIS CHARALAMBOUS* AND GEORGE ARTOPOULOS†

Abstract. It has been observed that many researchers in the humanities do not use digital tools to their full extent for their research. Some of the most pressing needs of researchers in Digital Cultural Heritage (DCH) are data storage and handling and large scale computing. Linking these researchers to experienced groups will significantly improve productivity and research innovation in DCH. This work presents our efforts in enabling virtual collaboration for research in the South East and Eastern Mediterranean region and more specifically the deployment of the Clowder CMS system and the development of extraction services to handle, manage and automatically process DCH data. We give technical descriptions of the system and provide some results and discussions of our efforts to enable virtual collaboration between regional level DCH researchers in the context of the Horizon 2020 funded VI-SEEM project.

Key words: Content Management System; Digital Cultural Heritage; Data Processing; Online Visualization

AMS subject classifications. 68M14

1. Introduction. Cultural heritage is a key factor of European identity. Heritage is anything that helps us collectively to better understand the present and think of our future. The Southeast Europe and Eastern Mediterranean region (SEEM) is renowned for its ancient civilizations. It is also an area of major socioeconomic and cultural developments during the medieval and early modern periods. The rich heritage in the region is in risk due to climate and human factors, such as war and conflicts, and the VI-SEEM project¹ is dedicated to its preservation, which includes activities such as heritage documentation, artefacts' analysis, conservation and preservation that spans from the conservation practices of objects to the scale of whole archaeological sites; to facilitate wider dissemination and provide access to knowledge for everyone, as well as to preserve artefacts in digital form by means of virtual reconstructions.

Learning about the history of a place is a good way of bringing communities together through a shared understanding of the unique cultural identity that heritage places give to an area. In our contemporary societies there is a pressing need to deal with issues of intercultural dialogue, social identity, and collective memory more than ever.

The most current need of researchers and scholars operating in the field of Digital Cultural Heritage (DCH) is how to produce quality from quantity, how to devise critical methodologies that produce meaning and generate knowledge out of big data, and VI-SEEM contributes with computation methods to help in achieving that. The process of interpretation is cross-disciplinary in nature and involves various faculties of human activity that rely on data processing, such as logical reasoning, associative analysis, descriptive capacity, linguistics and semiotic processes, decoding, and therefore cognition, abstraction and visualization, in order to reveal patterns and narratives, address the whole and provoke affection. After more than a decade of large-scale digitization processes spurring from most museums and libraries and archives, the next big challenge that all CH stakeholders are facing is to make sense, to add value and establish methods of interpretation and are common, comprehensive, sharable and easily applicable to the vast archives of data and complexity of digital assets in big data.

Throughout Europe, heritage sites, artefacts, texts, works of art are being electronically documented and subsequently archived. This is an on-going process due to the enormous number of artefacts and the continuous growth and development of digitizing technologies. However, the available datasets and related repositories remain fragmented, of varied quality, while access to data is still widely limited. One major effort to unite all data is the European Commission's effort for a digital library for European culture under the name of Europeana [16].

*Associate Research Scientist, CaSToRC, The Cyprus Institute, Cyprus. (ps.charalambous@cyi.ac.cy).

†Assistant Professor, STARC, The Cyprus Institute, Cyprus. (g.artopoulos@cyi.ac.cy).

¹<https://vi-seem.eu/>

Challenges of using computational tools for heritage studies. The majority of researchers in the humanities do not use digital tools to their full extent for their research. Large potential is identified for research groups that have not used large scale computing before. Linking these to experienced groups will significantly improve productivity and research innovation in DCH. Data storage and handling is one of the most pressing and challenging needs of the Cultural Heritage community. The VI-SEEM project focuses on the provision of tools and resources for regional scientists to cast their data into Content Management Systems (CMS), hence offering the stepping-stone to join larger initiatives on the longer run.

Arguably the field of digital cultural heritage has still to undergo the computation paradigm shift that characterises other fields of human activity and the sciences, such as biology, climate, geography, physics and many more. Today we should be moving into a new era of computation in DCH that goes beyond digitisation of artefacts and into the interpretation of data. However the reality is different; medium to small-scale cultural operators and museums in the region do not have the knowledge, resources and capacity to digitise their huge collections. VI-SEEM offers training for those users to facilitate them on how to digitise their assets and then curate them in order for the produced big datasets of digital assets to be *accessible*, *findable* (*searchable*) and *interoperable* (for ingestion in larger repositories and databases), following the FAIR policies as defined in the H2020 roadmap. So VI-SEEM both enables research in the region and facilitates the integration of locally generated results into larger initiatives at the European level.

The latest developments in interacting with big data scientific visualizations rely on intensive data mining that necessitate a shift of computational tools from the traditional off-line computer cluster to High Performance Computing capable of real-time parallel processing of multiple inputs. State of the art facilities invest in bringing together humanities and science, creative industries, art and engineering, in order to study and disseminate, and ultimately contribute to the preservation of tangible and intangible heritage (cf. AlloSphere [2] and Media Lab Helsinki [24]). Additionally, science has benefited greatly of advanced visualization methods of complex systems - a process that relies heavily on HPC.

We start by giving an overview of related work (cf. Sect. 2), we give an emphasis on the needs of the VI-SEEM communities and how our choice of a CMS (Clowder) helps these communities (cf. Sect. 3), we continue with a description of our system (cf. Sect. 4), some example applications of VI-SEEM and its impact (cf. Sects 5 and 6) and end with some discussion and future directions (cf. Sect. 7).

2. Related Work. There are various definitions of these software environments and platforms for collaboration, which, according to their context, are described as Virtual Research Environments [35, 9], Science Gateways [52], or Digital Libraries [6]. In their more general form, they comprise digital infrastructure and services which enable research to take place [17]. These platforms respond to the aims of e-Infrastructures [27] and cyberinfrastructures [15]. They respond to the needs of the collaborating communities in various ways combining multiple approaches, features, services and protocols, including portals, repositories to content management systems, such as for example Clowder [13] (formerly named Medici [39]), which is used in this work. The latter case offers a wide variety of services and tools that are integrated in a comprehensive way for users to exploit the resources available and facilitate the access of data.

Literature supports that on average most researchers of the community are willing to share knowledge and information about their inquiries, as long as they are provided with a streamlined and intuitive experience [14]. It also highlights the differences between disciplines in the way scholars utilise the VRE or take advantage of its capacities, i.e., the epistemological approach of each field impacts the way and content of data / information is shared, archived, analysed and presented, etc. [7, 8]. In doing so, some researchers driven by the culture of the discipline might share code and/or data, whereas others might only use a VRE for training purposes. Others might prioritise security and access control to data (e.g., copyrighted cultural heritage assets of private collections in Museums), while for some scholars, ease of access to information and the learning curve of operating the environment is of high value. It is widely recognised though that different roles and occupation of researchers necessitates different needs from a VRE, and therefore varied features and tools. Additionally, flexibility in, and control of, the level of security and user access, as well as the duration and familiarity of interacting with the VRE all play important roles to the success and adoption of the platform by communities of users.

Therefore, the major challenges that VREs need to overcome in order to penetrate a variety of fields and become sustainable and inclusive of research communities can be attributed to *accessibility*, *wealth of information*

(richness of data), *cybersecurity*, *findability* and *interoperability*. The literature also points to the complexity resulting from integrating big data resources and the difficulties that arise when combining data from various sources / archives, an issue that highlights the importance of taking measures for providing interoperability of data and the associated metadata [23, 29, 33, 34]. This measure is facilitated greatly by generating semantic structures of metadata that enable interaction with and query of the digital assets of each repository integrated in the VRE [37, 40, 42].

This paper presents how in the context of VREs the adaptation of a flexible CMS, such as Clowder, and the further customisation of its features to the needs of the regional communities can benefit research in the SEEM area, and promote collaboration for the preservation of the invaluable heritage of the region. Shared access and collaborative interaction with vast amounts of data is ever more important across a wide range of disciplines who seek for creative interdisciplinary discourse and investigating cross-disciplinary inquiries. Therefore providing customised access and control of large sets of data in a meaningful way; i.e., addressing the particularities of each discipline involved, is of paramount importance for the further development of Digital Heritage. This dynamic interface between data of knowledge and multiple users requires extensive processing power. Enabling archaeologists and historians, social and political scientists, engineers and natural scientists to access the same set of data in order to collaborate for the hands-on investigation of the links between nature (e.g., natural systems - weather and geo-physics) and culture (human artefacts - tangible and intangible) is one of the major challenges of society's computational futures.

3. Needs of the DCH community of SEEM and Clowder. VI-SEEM aims at strengthening links among key players in the field bringing users currently working autonomously together. Large potential is identified for research groups that have not used large scale computing before. There is a great potential in linking these groups with major activities in Europe, and thus offer access to the immense CH data in the region to pan-European initiatives.

In this context, Clowder responds to the needs of the DCH communities in the region and aims to provide the stepping-stone to join larger initiatives on the longer run. In particular Clowder is a content management system designed to support any data format and multiple research domains. It enables users to access and operate HPC infrastructure and provides a data management system for the following services and activities:

- data and associated metadata curation with user controlled access, file versioning, user authentication and assignment of Personal Identifiers (PIDs) to digital assets;
- online 3D visualization;
- curation and online access to geolocated data;
- cloud storage space; and,
- HPC processing services.

Some of the currently provided services and features on Clowder include:

- creation of digital repositories;
- management and safe access to data;
- searching and metadata integration;
- trained convolutional neural networks;
- optical character recognition for scanned documents (currently English);
- data of material analysis for conservation purposes;
- mapping metadata of archives and repositories for the creation of Digital Libraries; and
- tools for the creation of virtual museums using the Unity Game Engine².

Expected impact of using Clowder in VI-SEEM. As digitization of cultural heritage artefacts progresses by the museums of Europe and access to their digital archives is provided to an ever growing number of people from all around the globe, operating Digital Libraries and facilitating data-mining technologies for large repositories requires excessive amounts of computing power that only a HPC can offer. Furthermore, Grid-based solutions to inter-connect various library systems should be designed in order to allow end-users to search for content from a unique portal. The impact of the VI-SEEM VRE services and specifically Clowder features presented in this article is envisioned to benefit the following research inquiries:

²<https://unity3d.com/>

- *Digital libraries and interactive visualization of Cultural heritage.* Cultural heritage methodologies deal to a large extent with storage and analysis of artefacts and past knowledge. Applications include the management of large collections of scanned books and documents (like these of the Banatica Virtual Library application, cf. Sect. 5), as well as of dynamic file formats such as Reflectance Transformation Imaging (RTI) (e.g., the data from the Centre for the Study of Ancient Documents at Oxford University[50]). Providing to the user communities in the region access to these collections offers great opportunities for breakthrough contributions to art, historical and archaeological inquiries.
- *Image classification, feature extraction and machine learning techniques for image and video analysis.* Exploiting the computational capacity of HPC greatly benefits these methods as they require the analysis of large datasets. Additionally, artefact and built heritage structures' reconstruction (3D modelling) by means of photogrammetric techniques, such as structure-from-motion[19, 28], which rely heavily on image matching and feature extraction, benefit greatly from HPC infrastructure. VI-SEEM has been actively enabling and supporting these research activities which contribute to the presentation of sensitive or threatened heritage in the region (see online 3D Database System for Endangered architectural and archaeological Heritage in the south Eastern MEediterRAnea Area (EpHEMERA) [1]) due to the active engagement of various regional research groups³. Remote sensing image analysis is used for land cover and land use classification, built-up and clear land area detection, monitoring of urban growth, monitoring of natural disasters, etc. This is essential for assessing risks and policy making in the area of environmental and heritage protection. Communities in the region use feature learning for image classification in remote sensing, and geophysical analysis of earth subsurface. Electrical Resistivity Tomography (ERT) comprises one of the most important modern techniques of near surface applied geophysics; HPC infrastructure enables for accurate automated resistivity modelling and inversion schemes.
- *Immersive and interactive visualisation of archaeological sites, artefacts and virtual visits to museums,* such as the VirMuf application that is pursued by the Biblioteca Alexandrina, Egypt [51]. This is a dynamically growing area of research that exploits the capacity of grid- and cloud- computing for real-time rendering of sophisticated visual representations of inaccessible, sensitive, remote or destroyed objects and structures. Interaction opportunities are offered to the visitor of the virtual space for research and educational purposes [3]. Advanced human computer interfaces are developed to enable users to better interact with information and knowledge. Also spatially distributed narratives, storytelling, and elaborate playful ways (e.g., serious games) are developed in order to sustain longer user engagement in the virtual space [4].

DCH Data in the VI-SEEM project. Data, in the case of VI-SEEM and more specifically in the field of Digital Cultural Heritage, can be of very diverse types. More specifically users can upload entire datasets or individual files of:

- Scanned books and their metadata
- 3D Models
- Image, video, text and sound files and their metadata, organised in collections.
- Advanced documentation data, such as Reflectance Transformation Imaging, and analysis of material properties of structures, works of art and artefacts.
- Code and workflows with sample files to share computational tools and methods (e.g., trained convolutional neural networks, photogrammetric techniques, interactive real time rendering environments for virtual museums and more).
- Semantic referencing of metadata.

A total of 17 applications of user communities are currently serviced in the region, and they have been offered 30,000 CPU-hours and 3,030,000 GPU-hours at 2 HPC sites, and 51 Cloud VMs at 7 Cloud sites in total, reaching approx. 18 TB of storage space - consumed by all applications.

³e.g., Foundation for Research and Technology Hellas, Science and Technology in Archaeology Research Centera and The Cyprus Institute

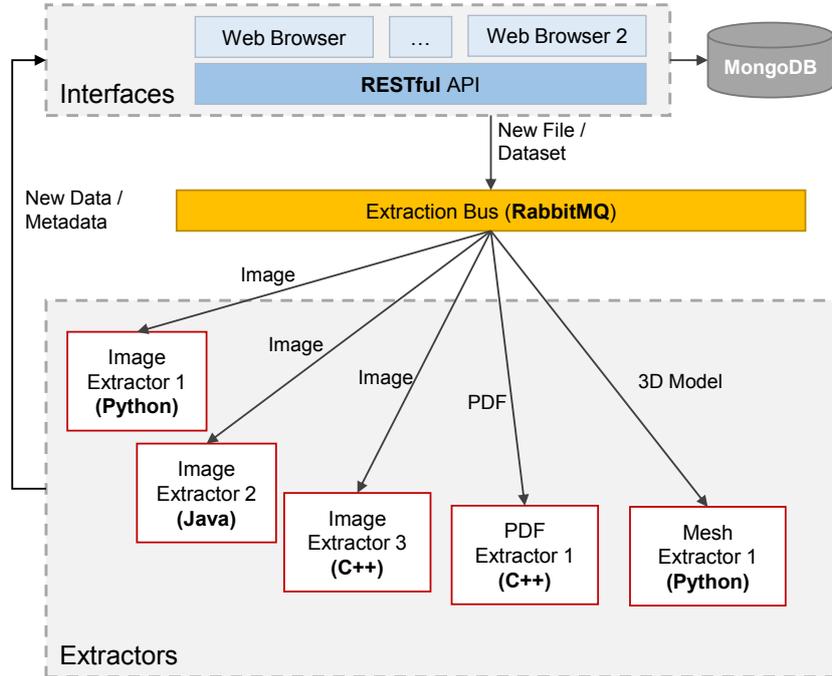


FIG. 4.1. **Clowder Architecture.** Users of the Clowder CMS can upload data using either a web interface or the RESTful API. Depending on the type (dataset, single file) and filetype of the data, the data are forwarded to **extraction services** which process the data to generate both new data and metadata that are associated with the source data. These new data are then processed similarly to the input data.

4. Methodology/Framework. We built our Digital Cultural Heritage management system on top of Clowder, which is a Web 2.0-based general multimedia content management system capable of semantic content management and service/cloud-based workflows [13]. It supports a broad range of research techniques and allows for community data management. Clowder provides scalable storage and media processing, simple straightforward user interfaces, search, social annotation capabilities, user management, preprocessing and previewing/visualization of various types data and metadata extension and manipulation. All of these features allows for the secure searchable access to large amounts of DCH data satisfying the needs of the DCH communities. At the core of Clowder are extraction services that allow for the *preprocessing*, *processing* and generation of *previews* for the data; these services are developed through collaboration of the various participating organizations.

4.1. Functionality. When users add new data to the system, whether this is through web front-end, or through the RESTful API, preprocessing is off-loaded to *extraction services* (cf. Fig. 4.1). These extraction services attempt to both extract metadata and generate new data based on the type of the data, e.g., to create image previews for videos or 3D files. These metadata are then associated with the uploaded data and presented to the user in the Clowder web interface. Newly generated data are uploaded back on the platform resulting in the call of different extraction services and so forth. We note that users can manually add and define other metadata at a later stage by using the web interface (or the RESTful API). Metadata can be defined at both the file and dataset level.

Users can upload and manage datasets in a variety of formats such as 3D, RTI, images, videos, text and audio (cf. Sect. 3); more formats can easily be integrated. Previews of large datasets in a variety of formats are also extracted and viewed to avoid the need of downloading the whole content on the user's system or finding the needed software to examine the contents of a file.

Clowder's scalability/parallelization, flexibility, and robustness, as well as its overall performance, are improved by decoupling the extraction services from the main server; i.e., multiple instances of the same extractor

can run on different machines in a distributed manner. We note that extraction services can be developed in a variety of programming languages and systems as long as they use the RabbitMQ message broker for communication with the Clowder instance (cf. Sect. 4.2.3). Currently, most of the extractors are written in Python and Java.

In the following paragraphs we give a brief overview of the technologies that are at the core of Clowder (cf. Sect. 4.2), how data and access control are managed (cf. Sects 4.3 and 4.4), and finally we give a description of the currently deployed extraction services (cf. Sect. 4.5).

4.2. Supporting technologies for Clowder. Clowder relies on various technologies to get the required flexibility to handle heterogeneous DCH data and metadata (cf. Fig. 4.1). These include the web server written using the Play framework (cf. Sect. 4.2.1), the MongoDB Database Management System (DBMS) (cf. Sect. 4.2.2) and the RabbitMQ Message Broker (cf. Sect. 4.2.3).

4.2.1. Web Server. The web server is built using the Play web application framework⁴ which supports both Java and Scala [25]. The Play framework provides minimal and predictable resource consumption which is really important for highly scalable applications. The server uses the model-view-controller (MVC) architectural pattern. It relies on a number of plugins for communication with the RabbitMQ broker and the MongoDB database, and user authentication. It uses dynamic HTML (ver.5) for webpage generation (e.g., views of the data) according to the results of input processing, search, etc. The models are closely associated with collections in the database. Preprocessors and scripts (i.e., previewers) running on users' browsers communicate with the server using a REST api [18].

4.2.2. Data Storage: MongoDB DBMS. The NoSQL MongoDB DBMS system [32] is used for the storage of both data and metadata in a flexible manner. It is a schema-less database [10, 21]; i.e., it does not require a rigid schema for the duration of the lifetime of the database, it does not enforce data type limitations, it can store both structured and unstructured data and administrators do not need to add additional layers on top to abstract the relational model into a more user friendly object oriented format. The choice of a schema-less database allows for more flexibility in handling the heterogeneous nature of DCH data and easier expansion such as community-generated metadata and new data formats.

4.2.3. Communication: RabbitMQ Message Broker. The role of the RabbitMQ message broker [36] is to take preprocessing messages from the web server that are sent once a dataset or file is uploaded and distribute them to the extractors that can then handle the jobs (cf. Fig. 4.1). The role of a message broker is to mediate the communication between applications [20, 26, 38]; this is done by validating, transforming and routing messages. RabbitMQ implements the Advanced Message Queuing Protocol (AMQP) [22]. In the case of Clowder, any extractor that is implemented must register one or more delivery queues on RabbitMQ the moment it is activated. Each queue is associated with a particular routing key set, which defines which routing keys a job can have in order for it to be routed to that queue. The extractor then continuously listens to the queue and acts accordingly.

4.3. Data Organization. Users can organize their data using a plethora of approaches; *datasets*, *collections* and *spaces* (cf. Fig. 4.2):

- Datasets contain related data; e.g., scanned books from a specific era or 3D models of objects scanned from an architectural side. Users can add metadata, tags or even comment on the data on a per file or dataset level.
- Collections are sets of related datasets such as sets of scanned books and audio transcripts of these books.
- Spaces can contain many datasets and collections and in addition users with different roles that can access and/or modify these data (cf. Sect. 4.4).

4.4. Access Management. Users can select the level of access to the data that they upload. This can be set using a variety of approaches. At a coarse level, data can be set as *public* or *private*; public data can be seen and downloaded by everyone that has access to the system, private data are restricted to selected users.

⁴<https://www.playframework.com/>

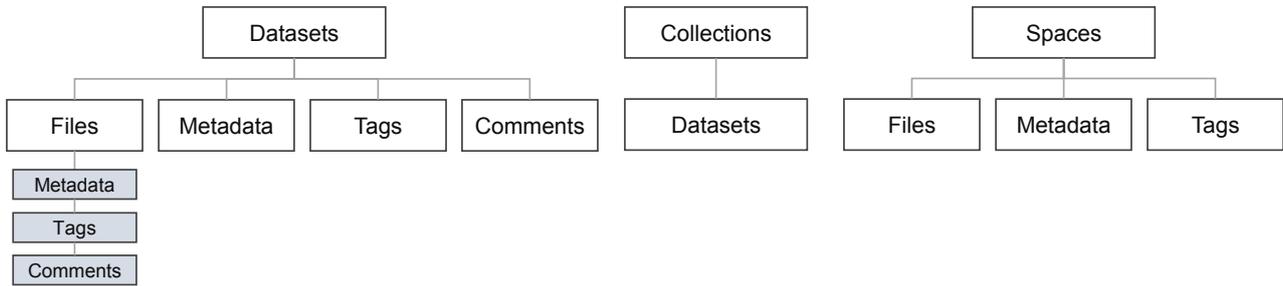


FIG. 4.2. **Data Organization in Clowder.** Data can be organized in datasets, collections of datasets and spaces associating users and their level of access to the data.

Additionally, in a finer level of control, data can be associated with *specific users* using spaces (cf. Fig. 4.3); these users can be assigned different roles such as administrators, editors or viewers of the data. The administrator of the system can define new roles and also controls who can register on the platform minimizing in the process misbehaving users (as much as possible).

4.5. The Vi-SEEM Instance of Clowder. We made the decision to create a separate Docker container⁵ for each one of the extractors, the Clowder instance, MongoDB and the RabbitMQ broker⁶. Additionally, most of the containers run on a single Virtual Machine (VM); extractors that need more processing are deployed on separate VMs. Having separate containers ensures that software is isolated and that the platform can be migrated with minimal effort and minimal conflicts between dependencies of different software. Additionally, code for the extractors and the docker containers for the project are available to the VI-SEEM community through the project's code repository [47].

Deployed Extractors. Depending on the VI-SEEM application, we develop extractors that do specialized processing on the data. More specifically, we develop(ed) extractors for:

1. extraction and importing of metadata from the Banatica collection of books
2. three-dimensional (3-D) inversion of surface Electrical Resistivity Tomography
3. automatic image georeferencing tomography (ERT) data in order to automatically determine a 3-D resistivity subsurface model using the AutoGR-Toolkit[5]
4. compressed file handling, such as contents, extraction of content and running of other extractors based on contents,
5. Reflectance Transformation Imaging (RTI) previewing and 3D model generation
6. optical character recognition of English documents

Additionally, we employ several of the readily available extractors, such as generation of previews for images, video preview generation, etc.

5. Examples of novel DCH research activities enabled by VI-SEEM. Some of the most visible research activities supported by Clowder that are also facilitated through VI-SEEM infrastructure are described below.

DataCrowds. Today more people are living in urban environments than in rural areas. It is forecasted that 70% of the global population will be living in cities by 2050. This intense urbanisation poses huge challenges in overcrowding, segregation, demographics and use of resources. The main goal of this project is to innovate in the unified area of research that is occupied with the transdisciplinary study of crowds in built environments. This project envisions a web-accessible, social platform that will allow researchers from very diverse fields, such as Crowd Simulation, Urban Modeling and Simulation, Pedestrian Dynamics, Computer Graphics, Social

⁵<https://www.docker.com/>

⁶A container is a lightweight, stand-alone, executable package that includes whatever a software package needs to run (e.g., code, executables, libraries, tools and settings).

Name	Description	Permissions	Edit / Delete
Admin	Admin Role	Space: View Create Delete Edit Resource To Space: Add Staging Area: Edit Dataset: View Create Delete Edit Resource To Dataset: Add	
Editor	Editor Role	Space: View Create Delete Edit Resource To Space: Add Staging Area: Edit	

FIG. 4.3. **User Roles.** The administrator can assign user roles; each can have different access to data and the owners of data can set the role of each user in a space.

Dynamics and Architecture to collaborate, share data and take advantage of each fields breakthroughs in order to contribute more accurate crowd simulations for the future sustainability of urban environments. As a first step in the implementation of this project, tracked data of crowds from various sources (such as the ones used in [11, 12, 31]) are being uploaded to the Clowder platform of the Vi-SEEM project.

PETRA. The “PETRA: Petra Painting Conservation Project”, which is pursued in collaboration with the Synchrotron-Light For Experimental Science And Applications In The Middle East (SESAME)⁷ and is developed for the Department of Antiquities of Jordan by the Department of Optics and Atomic Physics at the Technical University Berlin. PETRA provides documentation, condition assessment, and characterization of Nabataean wall paintings and painted marble sculptures from Petra, with a focus on its gilded wall paintings. Characterisation methods include 2D and 3D Micro-XRF, Micro-XANES, handheld XRF, handheld FTIR in addition to various complementary lab-based characterisation techniques. An important aim of this project is to survey painted material in Petra, e.g., collecting historical and recent research material about painted walls and sculpture in Petra, including photos, descriptive documents, analysis data, etc. This activity involves not only in-situ survey of the remaining intact painted walls as well as painted fragments and painted marble sculpture in Petra, but also the study and analysis of the painted material (condition assessment) and objects. In-situ and ex-situ analysis work is taking place in Petra and Berlin. The use of Clowder to store, access, share, link and compare the data, and even visualise them at a later stage, will certainly strengthen this research, time-wise and money-wise.

HaPPen. Another application that is currently under development is “High Performance Photogrammetry (HaPPen)” pursued at the Science and Technology Research Center at the Cyprus Institute. The installation of photogrammetric tools for running structure-from-motion methods for the digital reconstruction of monuments and artefacts from large collections of high-resolution photographs, i.e., of massive datasets of images, acquired from underwater, terrestrial and aerial survey systems, can be better served by HPC infrastructures. This is a computationally intensive process of repetitive image matching and feature extraction operations, and is currently widely used by DCH communities, where budget constraints and requests for high accurate models are ever rising. The project will test and implement a set of commercial and open source software to be used for image based 3D reconstruction processes. It will also conduct an assessment on performance and usability of the available software and methods, and benchmarks will be provided/shared to the DCH communities for further exploitation.

3DInv and AutoGR. The same computational logic (i.e., feature extraction) is exploited by yet another set of significant for the region applications that involves the massive georeferencing of aerial images [43], and the 3D reconstruction of subsurface conditions and structures by large sets of imaging data including satellite images and electrical resistivity tomography[44], respectively (cf. Fig. 5.1). These applications are pursued by different groups of the Foundation for Research and Technology Hellas (FORTH), Institute for Mediterranean Studies (IMS), Laboratory of Geophysical Satellite Remote Sensing and ArchaeoEnvironment (GeoSat ReSeArch Lab). Electrical Resistivity Tomography (ERT) involves the reconstruction of a subsurface resistivity distribution

⁷<http://www.sesame.org.jo/sesame/>

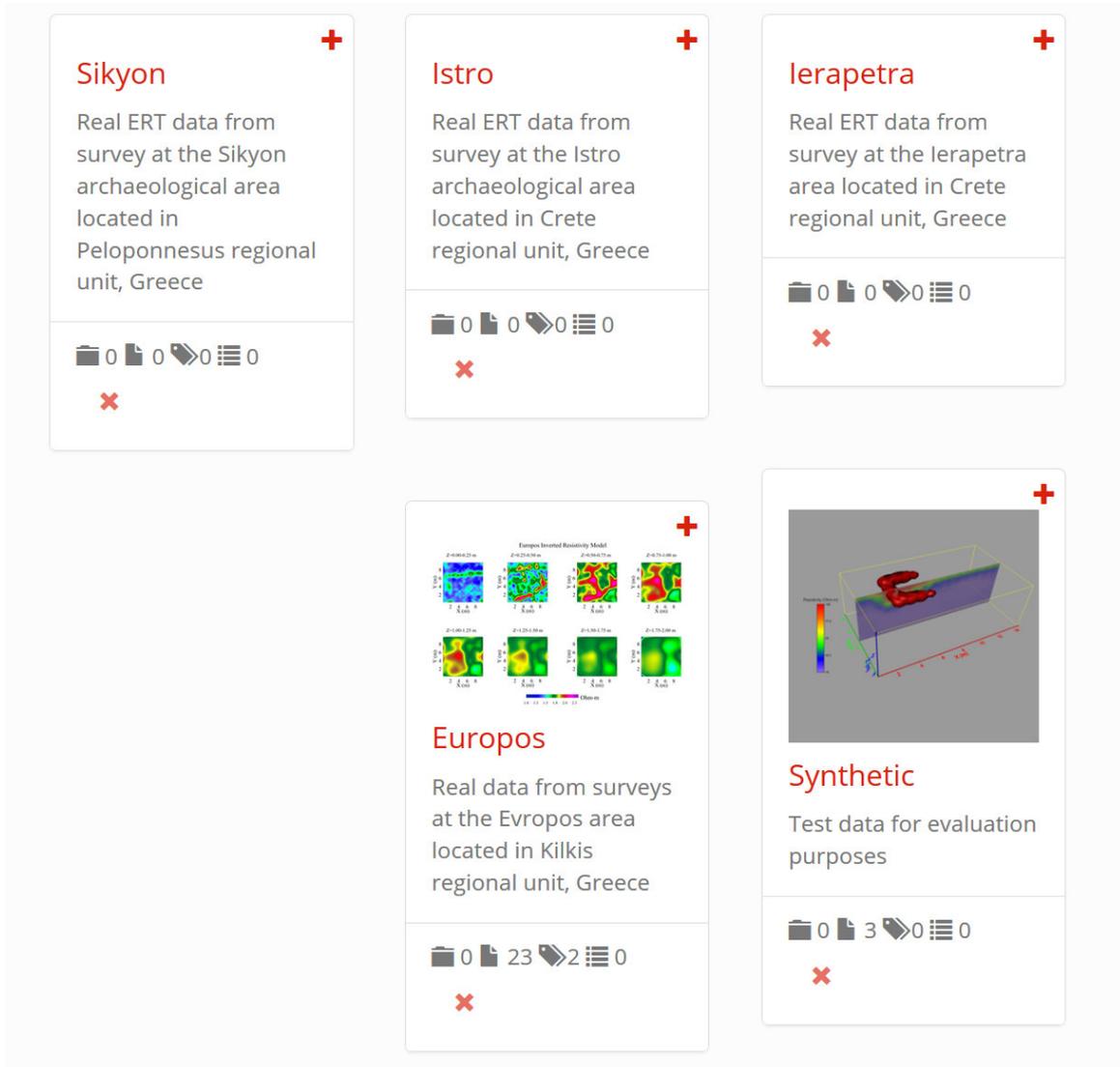


FIG. 5.1. **3DInv Datasets.** Datasets of electrical resistivity tomography; a specialized extractor is run and processes these data to generate correspondances between images.

for revealing finer archaeological details hidden in the original data through the reconstruction of truly 3-D resistivity models of the hidden archaeological relics. The knowledge gained and information acquired by the interpretation of the experimental data is expected to contribute to the update of the relevant policies and the revision of management plans of archaeological sites in Greece.

In all these applications VI-SEEM provides access to HPC infrastructure for running the software but at the same time Clowder offers to the users the opportunity to store, access, share, link, compare and visualise the data. These descriptions showcase only but a few of the wide range of applications that are currently under development and are briefly featured on the VI-SEEM DCH collaboration platform that is enabled by the use of Clowder (<http://dchrepo.vi-seem.eu>).

6. The impact of VI-SEEM to DCH inquiries. An application that already resulted a significant contribution of the VI-SEEM project in the DCH communities in the region is the Banatica Virtual Library (cf. Fig. 6.1), which is developed by the West University of Timisoara in collaboration with the IT Department of

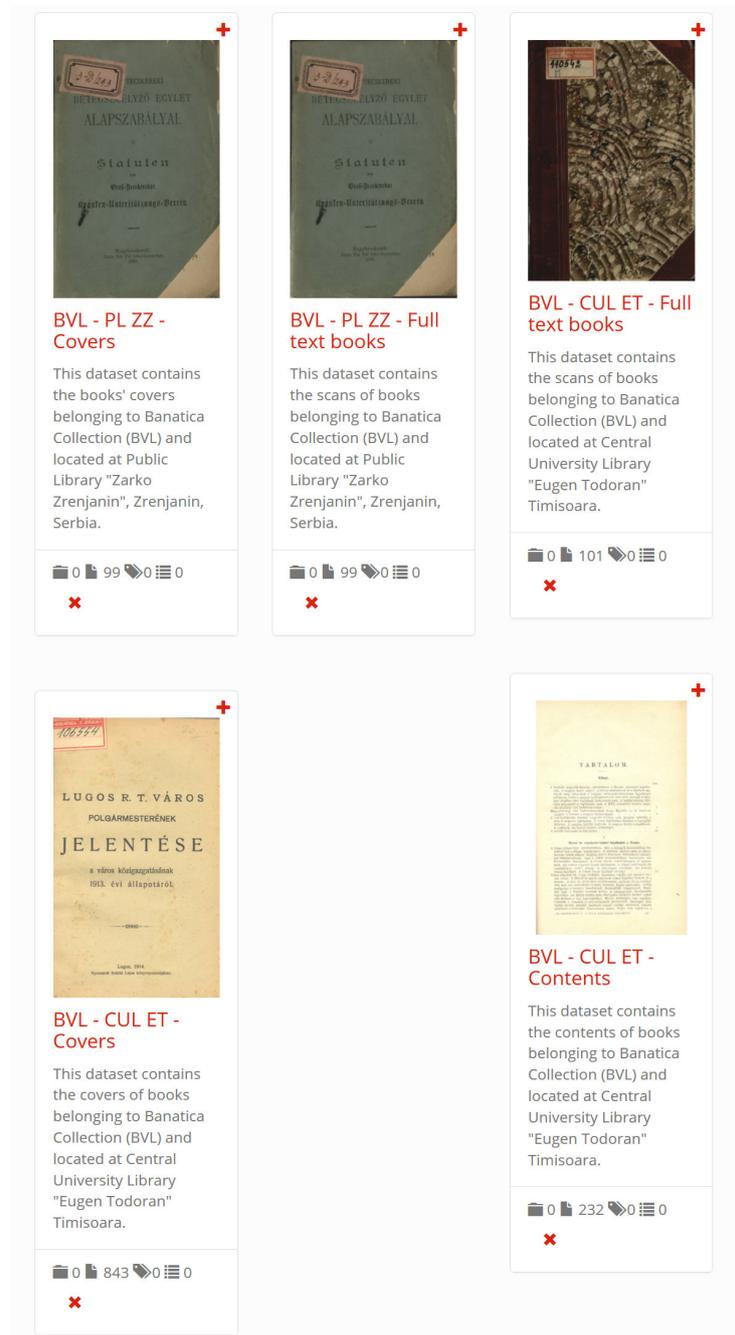


FIG. 6.1. *The Virtual Banatica Library collection on Clowder.*

the Central University Library "Eugen Todoran" Timisoara. This application result is significant as it combined the use of HPC for running computationally intensive processes and Clowder for storage, access and sharing of the resulting data [46]. In doing so the Banatica Virtual Library makes rare and old publications accessible again for a wider scientific audience in the SEE region and enables further machine processing to be applied on the entire manuscript collection.

The BANATICA collection gathers together all the printed products considered monographs (e.g., brochu-

res, books, yearbooks, calendars in volumes, prints with an individual cover, atlases, book-like printed scores etc.), which represent documentation sources for the culture and civilization on the Banat region. This collection was jointly created by the VI-SEEM partner, Central University Library "Eugen Todoran" Timisoara (Romania), and "Zarko Zrenjanin" public library (Serbia) throughout the *Biblio-Ident* IPA funded project. The collection comprises over 1000 bibliographic descriptions and 200 full-text scanned books. On Clowder the entire collection was organized into five datasets (cf. Fig. 6.1): two containing the covers of the publications from Banatica collection, two datasets each of each containing 100 full book scans (one for books owned by BCUT and another one for books hosted by Public Library Zarko Zrenjanin), plus an additional dataset of table of contents for the books. During the upload process, metadata was added to the documents to enable searching and association of the data.

In order to make the content machine readable, an environment to run optical character recognition (OCR) on the documents of the collection was setup by the developers. In the first stage of processing, noise was removed from the documents, and the scanned photographs of the pages were sharpened. The second stage involved the OCR process with the use of an open source engine [41]. In parallel the code and scripts behind this OCR pipeline were uploaded on the VI-SEEM's code repository [49]. The developers shared their experience which pointed to the CPU intensity of the process - the initial benchmark for a dataset of 200 digitized prints took 4 days (on a dedicated Virtual Machine). Eventually in order to speed up the process, the operation was replicated on multiple VMs, each getting a subset of PDF documents in a round robin fashion, a master being responsible for distributing the work to multiple workers that run the processing pipeline.

7. Discussion / Future Directions. The overarching goal of the VI-SEEM project is to facilitate cross-fertilisation of research activities between fields and disciplines in order to promote and accelerate interdisciplinary inquiries in the region. The VRE portal of the project, as well as Clowder for the DCH communities, will hopefully contribute cross-thematic activities between the three scientific communities of the project. The services that the VI-SEEM provides to the communities in order to enable interdisciplinary inquiries include data visualization, simulation data, data analytics and processing, geographic description of datasets and curation (e.g., Levante, Balkan regions, etc.), analytical studies and access to source code and the relevant training material. First examples of initialised activities that could result in interdisciplinary research involve the following cases:

- Impact of climate change on the experience of built heritage: visualise the impact of climate anomalies on historic sites and landscapes (under development by The Cyprus Institute);
- Remote sensing and preservation of heritage (developer: FORTH);
- Computer vision and documentation of heritage (developer: FORTH and The Cyprus Institute);
- Machine learning and documentation of heritage, e.g., CNN for satellite images [48] (developer: University of Banja Luka, Faculty of Electrical Engineering);
- Impact of climate on tangible heritage: for conservation purposes (PETRA); and,
- Climate and life sciences for the study of the impact of climate change on evolution (e.g., the "Aharoni" Digitized Collections: Past, present and future of the southern Levant biodiversity by the National Natural History Collections, at the Hebrew University of Jerusalem).

Collaboration of the DCH and Climate communities. The first successful cross-disciplinary research activity between the Digital Cultural Heritage and Climate communities that capitalised on recourses offered by the VI-SEEM project was recently exhibited at a popular international venue. The Seoul International Biennale on Architecture & Urbanism was a large-scale public event organized by the Seoul Metropolitan Government and Seoul Design Foundation and received 4m visitors over the course of its duration. Titled "Imminent Commons", the Biennale provided a forum for debate to policy makers, experts and citizens at large.

Following the UN's World Urbanization Prospect Report of 2014, 54% of the world's population now live in metropolitan areas. By 2050, this percentage will increase to 86% in advanced countries, and 64% in developing nations. Already now, the MENA (Middle East and North Africa) region, renown for its wealth of cultural heritage, ancient civilizations' monuments and major sociocultural developments during medieval and early modern times, is experiencing a high degree of urbanization. Climate change will have particularly strong manifestations in the lived experience of urban settings (e.g., Lelieveld et al., 2014 [30]), and will pose great challenges to the material integrity as well as use of built heritage in these environments.



FIG. 7.1. Visualization of the extreme dust event that took place on September 8, 2015 as seen in virtual reality using Nicosia simulation model. (Credits: Georgios Artopoulos, Theodoros Christoudias, Panayiotis Charalambous, Colter Wehmeier, Charalambos Ioannou, Charis Iacovou, Harry Varnava, Adriana Bruggeman, Panos Hadjinicolaou, Katerina Charalambous, Jonilda Kushta).

Nicosia, the capital of the Republic of Cyprus and the only major inland city of this Eastern Mediterranean island has been continuously inhabited for over 4500 years. Estimated to become a climate change *'hotspot'* in the foreseeable future, the people of this city already face the effects of the region's changing weather patterns and climate trends. The presented collaborative research activity involved an interactive audiovisual exhibit of immersive simulations that illustrate possible futures of this city, visualising forthcoming conditions of heat, dust and floods using scientific data of climate observations and (computationally) simulated projections (cf. Fig. 7.1). The long-term objective of this activity is to contribute towards the achievement of an integrated climate change adaptation strategy for all of the evolving 'hot spot' cities of the region, and to safeguard the well-being of people living in these locations including both their social structures and the conservation of the built environment.

Finally an application that shows great potential for drawing links with other fields and disciplines within the VI-SEEM project is the "Aharoni" online digitized collection [45], which aims at creating a digital repository for presenting and preserving the greatest Levantine faunal collection from the beginning of the 20th century. The application is pursued by the National Natural History Museum at the Hebrew University⁸ and focuses on the promotion and study of collections and archival content of the unique fauna (avian, amphibian, reptiles and mammalian species) of the Levant region, and are the sole direct evidence of that region's biodiversity. The 3D documentation with the support and training of VI-SEEM and online access on Clowder to the generated digital models of the content of these collections of specimens, linked with all relevant metadata will serve as a high-quality database of the southern Levant fauna both for the academic community, as a key biological resource, and for the general public as a repository of knowledge about this unique region.

Most importantly though the developers of this application envision to exploit the capacity of the VI-SEEM VRE for pursuing interdisciplinary collaborative activities. They propose to complement the Clowder repository of specimens with the available analytic platforms in biodiversity science, such as the Open Tree of Life, iDigBio, Lifemapper, Arbor and other complex post-tree analyses (e.g. niche modeling, niche diversification). The aim of this interdisciplinary activity is the use of HPC support for the analysis of the big data of biological studies in order to better understand the complex patterns conniving biodiversity loss in order to promote future land management and wildlife conservation programs.

⁸<https://nnhc.huji.ac.il/>

Acknowledgments. This work was supported by the European Union’s Horizon 2020 research and innovation program, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

REFERENCES

- [1] ABATE, D., AVGOUSTI, A., FAKA, M., HERMON, S., BAKIRTZIS, N., AND CHRISTOFI, P., *An online 3D database system for endangered architectural and archaeological heritage in the South-Eastern Mediterranean*, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W3, 1-8, <https://doi.org/10.5194/isprs-archives-XLII-2-W3-1-2017>, (2017).
- [2] <http://www.allosphere.ucsb.edu/>, *AlloSphere Research Group*, (Page retrieved 12 March 2018).
- [3] ARTOPOULOS, G. AND BAKIRTZIS, N., *Virtual Narratives For Complex Urban Realities: historic Nicosia as Museum*, Electronic Media and Visual Arts / Elektronische Medien und Kunst, Kultur, Historie (e-book; open access) (Germany: Heidelberg University), 190–99, (2006).
- [4] ARTOPOULOS, G. AND CONDORCET, E., *House of Affects ’ Time, immersion and play in digital design for spatially experienced interactive narrative.*, Digital Creativity Journal, vol. 17,4, 213–20, (2006).
- [5] <http://ims.forth.gr/AutoGR>, *AutoGR Toolkit*, (Page retrieved 12 March 2017).
- [6] CANDELA, L., CASTELLI, D., AND PAGANO, P., *History, Evolution and Impact of Digital Libraries*, In I. Iglezakis, T.-E. Synodinou, & S. Kapidakis, *E-Publishing and Digital Libraries: Legal and Organizational Issues* (pp. 1–30), IGI Global. (2011).
- [7] CANDELA, L., CASTELLI, D., PAGANO, P., AND SIMI, M., *From Heterogeneous Information Spaces to Virtual Documents.*, Digital Libraries: Implementing Strategies and Sharing Experiences, 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 12–15, 2005, Proceedings, Springer, (2005).
- [8] CANDELA, L., CASTELLI, D., AND THANOS, C., *Making Digital Library Content Interoperable* Digital Libraries - 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010, 13–25, (2010).
- [9] CARUSI, A., AND REIMER, T., *Virtual Research Environment Collaborative Landscape Study*, JISC (2010).
- [10] CATELL, R., *Scalable SQL and NoSQL data stores*, SIGMOD Rec. 39, 4 (May 2011), 12–27, (2011).
- [11] CHARALAMBOUS, P., AND CHRYSANTHOU, Y., *The PAG Crowd: A Graph Based Approach for Efficient Data-Driven Crowd Simulation.*, In Computer Graphics Forum (Vol. 33, No. 8, pp. 95–108). (2014).
- [12] CHARALAMBOUS, P., KARAMOUZAS, I., GUY, S. J., AND CHRYSANTHOU, Y., *A Data-Driven Framework for Visual Crowd Analysis.*, In Computer Graphics Forum (Vol. 33, No. 7, pp. 41–50). (2014).
- [13] *CLOWDER: Open Source Data Management for Research.*, (Page retrieved 15 December 2017), <https://clowder.ncsa.illinois.edu>
- [14] CONNAWAY, L.S. AND DICKEY, T.J., *Common Themes Identified in an Analysis of JISC Virtual Research Environment and Digital Repository Projects.*, OCLC Research, (2009).
- [15] CYBERINFRASTRUCTURE COUNCIL., *Cyberinfrastructure Vision for the 21st Century Discovery*, National Science Foundation, (2007).
- [16] *Europeana Collections*, (Page retrieved 12 March 2018). <https://www.europeana.eu/portal/en>
- [17] FRASER, M. *Virtual research environments: overview and activity.*, Ariadne, (44), (2005).
- [18] FIELDING, R. T., AND TAYLOR, R. N. *Architectural styles and the design of network-based software architectures (p. 151)*, Doctoral dissertation: University of California, Irvine. (2000).
- [19] FORSYTH, D., AND PONCE, J. *Computer vision: a modern approach.*, Upper Saddle River, NJ; London: Prentice Hall. (2011).
- [20] GAMMA, E.; HELM R., JOHNSON, R., AND VLISSIDES, J., *Design Patterns: Elements of Reusable Object-Oriented Software.*, Addison-Wesley, (1995).
- [21] HAN, J., HAIHONG, E., LE, G., AND DU, J., *Survey on NoSQL database*, In Pervasive computing and applications (ICPCA), 2011 6th international conference on (pp. 363–366), IEEE, (2011).
- [22] O’HARA, JOHN, *Towards a Commodity Enterprise Middleware*, Queue, Vol. 5, No. 4, pp. 48–55, (2007).
- [23] HEATH, T., AND BIZER, C., *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool, (2011).
- [24] *Media Lab Helsinki*, (Page retrieved 12 March 2018). <https://medialab.aalto.fi/research>
- [25] HILTON, P., BAKKER, E., AND CANEDO, F. (EARLY ACCESS EDITION), *Play for Scala*, Shelter Island, NY: Manning, (2012).
- [26] HOHPE, G., AND WOLF, B., *Enterprise integration patterns: Designing, building, and deploying messaging solutions.*, Addison-Wesley Professional. (2004).
- [27] E-INFRASTRUCTURE REFLECTION GROUP, *Blue Paper*, E-IRG, (2010).
- [28] KOENDERINK, J. J., AND VAN DOORN, A. J., *Affine structure from motion.*, JOSA A, 8(2), 377–385. (1991).
- [29] LE BOEUF, P., DO’RR, M., ORE, C.E., AND STEAD, S., *Definition of the CIDOC Conceptual Reference Model*, Available at http://www.cidoc-crm.org/official_release_cidoc.html, (2012).
- [30] LELIEVELD, J., HADJINICOLAOU, P., KOSTOPOULOU, E., GIANNAKOPOULOS, C., POZZER, A., TANARHTE, M. AND TYRLIS E., *Model projected heat extremes and air pollution in the eastern Mediterranean and Middle East in the twenty-first century*, Reg Environ Change, 14, 1937–1949. (2014).
- [31] LERNER, A., CHRYSANTHOU, Y., AND LISCHINSKI, D., *Crowds by example.*, In Computer Graphics Forum (Vol. 26, No. 3, pp. 655-664). Blackwell Publishing Ltd. (2007).
- [32] *MongoDB*, (Page retrieved 18 December 2017). <https://www.mongodb.com/>
- [33] PAEPCKE, A., CHANG, C. K., WINOGRAD, T., AND GARC’A-MOLINA, H., *Interoperability for Digital Libraries Worldwide*, Communications of the ACM, 41, 33–42. (1998).

- [34] PARK, J., AND RAM, S., *Information Systems Interoperability: What Lies Beneath?*, ACM Transactions on Information Systems, 22, 595–632. (2004).
- [35] PICCOLI, G., AHMAD, R., AND IVES, B., *Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training.*, MIS quarterly, 401–426. (2001).
- [36] *RabbitMQ*, (Page retrieved 18 December 2017). <https://www.rabbitmq.com/>
- [37] RONZINO, P., NICCOLUCCI, F., AND D’ANDREA A., *Built Heritage metadata schemas and the integration of architectural datasets using CIDOC-CRM*, Online proceedings of the International Conference Built Heritage 2013, Monitoring Conservation Management, 18–19 November 2013, Milan. (2013).
- [38] SCHMIDT, D. C., STAL, M., ROHNERT, H., AND BUSCHMANN, F., *Pattern-Oriented Software Architecture, Patterns for Concurrent and Networked Objects (Vol. 2)*, John Wiley & Sons. (2013).
- [39] SOPHOCLEOUS, C., MARINI, L., GEORGIOU, R., ELFARARGY, M., AND MCHENRY, K., *Medici 2: A scalable content management system for cultural heritage datasets.*, Code4Lib Journal, (2017).
- [40] STEPHENS, R.T., *Utilizing metadata as a knowledge communication tool*, In Professional Communication Conference, PCC 2004, International Proceedings, (2004).
- [41] *Tesseract OCR*, (Page retrieved 12 March 2017). <https://github.com/tesseract-ocr/tesseract>
- [42] VASSALLO, V. AND PICCININNO, M., *Aggregating Content for Europeana: A Workflow to Support Content Providers*, Lecture Notes in Computer Science, Volume 7489, Theory and Practice of Digital Libraries, Rasmussen and F. Loizides (eds.), Springer. (2012).
- [43] *ViSEEM, 3DInv Application*, Georeferencing of Aerial Images, (Page retrieved 12 March 2017). <http://dchrepo.vi-seem.eu/datasets/591184d5e4b03cc97586975d>
- [44] *ViSEEM, 3DInv Application*, Electrical Resistivity Tomography Reconstruction, (Page retrieved 12 March 2017). <http://dchrepo.vi-seem.eu/datasets/58e4a6e6e4b06113f7bf4c81>
- [45] *ViSEEM, “Aharoni” Collection*, (Page retrieved 12 March 2017). <http://dchrepo.vi-seem.eu/spaces/59ec8ebce4b013b4ad5aa0db>
- [46] *ViSEEM, Banatica Virtual Library*, (Page retrieved 12 March 2017). <http://dchrepo.vi-seem.eu/spaces/58f1bf66e4b02fd8f8f22e16>
- [47] *ViSEEM Code, Clowder Code Repository*, (Page retrieved 12 March 2017). <https://code.vi-seem.eu/totis77/extractors-cyi>
- [48] *ViSEEM Code, Convolutional Neural Networks for Satellite Images*, (Page retrieved 12 March 2017). <https://code.vi-seem.eu/Risojevic/cnn-features.git>
- [49] *ViSEEM Code, Distributed OCR*, (Page retrieved 12 March 2017). <https://code.vi-seem.eu/bogconst/viseem-distributed-ocr>
- [50] *ViSEEM Project, Digital Cultural Heritage*, (Page retrieved 12 March 2017). <https://vi-seem.eu/cultura-heritage/>
- [51] *VirMuf*, (Page retrieved 12 March 2017). <http://dchrepo.vi-seem.eu/spaces/5900b3fde4b02fd8247bab63>
- [52] WILKINS-DIEHR, N. *Special Issue: Science Gateways - Common Community Interfaces to Grid Resources.*, Concurrency and Computation: Practice and Experience, 19 (6), 743–749, (2007).

Edited by: Aneta Karaivanova

Received: Dec 22, 2017

Accepted: Mar 12, 2018