



Article

Hydria: An Online Data Lake for Multi-Faceted Analytics in the Cultural Heritage Domain

Kimon Deligiannis, Paraskevi Raftopoulou, Christos Tryfonopoulos *, Nikos Platis and Costas Vassilakis

Department of Informatics & Telecommunications, University of the Peloponnese, GR22131 Tripolis, Greece; deligiannis@uop.gr (K.D.); praftop@uop.gr (P.R.); nplatis@uop.gr (N.P.); costas@uop.gr (C.V.)

* Correspondence: trifon@uop.gr, Tel.: +30-2710-230-175

Received: 29 March 2020; Accepted: 15 April 2020; Published: 23 April 2020



Abstract: Advancements in cultural informatics have significantly influenced the way we perceive, analyze, communicate and understand culture. New data sources, such as social media, digitized cultural content, and Internet of Things (IoT) devices, have allowed us to enrich and customize the cultural experience, but at the same time have created an avalanche of new data that needs to be stored and appropriately managed in order to be of value. Although data management plays a central role in driving forward the cultural heritage domain, the solutions applied so far are fragmented, physically distributed, require specialized IT knowledge to deploy, and entail significant IT experience to operate even for trivial tasks. In this work, we present Hydria, an online data lake that allows users without any IT background to harvest, store, organize, analyze and share heterogeneous, multi-faceted cultural heritage data. Hydria provides a zero-administration, zero-cost, integrated framework that enables researchers, museum curators and other stakeholders within the cultural heritage domain to easily (i) deploy data acquisition services (like social media scrapers, focused web crawlers, dataset imports, questionnaire forms), (ii) design and manage versatile customizable data stores, (iii) share whole datasets or horizontal/vertical data shards with other stakeholders, (iv) search, filter and analyze data via an expressive yet simple-to-use graphical query engine and visualization tools, and (v) perform user management and access control operations on the stored data. To the best of our knowledge, this is the first solution in the literature that focuses on collecting, managing, analyzing, and sharing diverse, multi-faceted data in the cultural heritage domain and targets users without an IT background.

Keywords: cultural heritage; big data management; data lake; data store; analytics and visualization; open source

1. Introduction

In the last few years Cultural Informatics (CI) has surfaced as a new promising domain that constitutes the socio-technological approach to understand, represent, communicate and re-invent cultures and cultural institutions [1]. CI may also be used in a disruptive fashion, aiming to change the way we understand and experience our cultural heritage [2], by enabling us, for example, to create personalized museum experiences [3,4], to discover facets and stories from new or existing cultural heritage data [5–7], or to create inter-linked cultural data repositories [8–11]. While performing these tasks, CI are creating an avalanche of data, produced by a vast number of related activities such as profiling of or feedback from museum and cultural venue visitors [12–15], social media activity (e.g., posts and comments) related to cultural events [16–23], papers and specialized conferences on the topic [24–27], and raw data on cultural objects such as artifact descriptions [28–31]. This data is typically fragmented and distributed among the different stakeholders, while the data management solutions

that are involved vary greatly, ranging from simple spreadsheet files for the less tech savvy to typical data stores such as relational databases [32–34] or semantically richer knowledge bases [9,10,35,36].

From the above, we can conclude that data management is a key technological factor that drives the Cultural Heritage (CH) domain forward [35,37], but the data management solutions applied so far are fragmented, physically distributed, heterogeneous, non-aligned and require specialized IT knowledge to deploy and operate (e.g., [38–40]). Moreover, the asynchronous nature of the data acquisition process itself poses new challenges in the collection, organization, and processing of the relevant data [37]. Proposed solutions for data storage of cultural information (e.g., [9,10,35,41]) usually require significant computing infrastructure, which is not easy to obtain or maintain, and the constant support and active involvement of IT experts even for trivial tasks, like creating a graph from the given data, updating the produced statistics, or incorporating a new data source/set. Typically, reconfiguring an existing solution for reuse in another setup or setting one up from scratch for the specific cultural data management problem requires (i) time-consuming meetings between scientists of different disciplines trying to understand each other's needs and goals, and (ii) resource-consuming IT infrastructure that calls for outsourcing to IT specialists and regular maintenance/upgrades to keep up with technological requirements [42]. Due to these issues, a great number of stakeholders (such as small museums or humanities research groups) that lack the resources for infrastructure and/or computing expertise still rely on outdated approaches like (i) storing their data in spreadsheets or raw files, (ii) sharing their data with colleagues through email, cloud uploads of zip files, or even by snail mailing electronic copies in removable media, and (iii) analyzing the data via sub-standard tools and trial software. Such practices create, in turn, other concerns like data freshness/integrity issues due to versioning, issues with the significance of reported results due to data scarcity/fragmentation, and even ethical issues like unequal access to data and resources [43].

In this work, we present *Hydria*, a novel *free online data lake* meant for *acquiring, storing, organizing, analyzing* and *sharing* heterogeneous, multi-faceted cultural heritage data. *Hydria* and all the provided functionality (given in detail in the following sections) are fully developed by the authors by resorting to open source tools. The data lake architecture [44] adopted in the design of *Hydria* enables the direct incorporation of heterogeneous information that has been recorded in dispersed formats, while specialized processing engines ingest data without compromising the data structure, making it available for tasks such as visualization, mining, analytics and reporting. In this sense, the proposed system targets primarily the functional requirements posed by the cultural informatics domain, and enables researchers, curators and other stakeholders within the cultural informatics domain to easily acquire, manage and share data/knowledge within *Hydria*.

Thus, the *Hydria* system proposed in this paper is an innovative, integrated framework that enables users with *no prior IT knowledge* to (i) *setup and launch*, in an easy and transparent way, data acquisition services like topical focused crawlers, social media monitors, web scrapers and dataset imports, (ii) *collect* questionnaires and other types of user input data by resorting to several built-in and customizable data entry forms, (iii) *record, organize and manage* collected data by storing them in different data stores (called *data ponds* in the *Hydria* terminology), (iv) *share* whole data sets or horizontal/vertical data shards via a powerful publish/subscribe mechanism that notifies users when other data ponds store data of interest, (v) *search and analyze* data by using a powerful yet simple point-and-click mechanism that performs queries on the stored data and extracts the requested information in several formats/outputs such as histograms, pie charts, (heat) maps, (stacked) bars/columns, area charts and various file types (like CSV/TSV and raw), and (vi) perform basic and advanced *user management* tasks (such as manage users, assign user privileges and permissions, perform access control on data and data ponds). All services are designed for usage *by non-IT experts* and are configured/executed by resorting to step-by-step wizards, contain in-context explanations for the different system functions and provide online help with examples.

The contributions of this work are three-fold:

- (i) We put forward Hydria, an *online, free, zero-administration* data lake that offers both fundamental and advanced user and data/knowledge management functionality for big cultural data management. To the best of our knowledge, this is the first system that focuses on collecting, managing, analyzing, and sharing diverse, multi-faceted data in the cultural heritage domain and allows users without an IT background to deploy, populate, and manage their own data stores within minutes, alleviating the need to rely on expensive custom-made solutions that require IT infrastructure and skills to maintain.
- (ii) We present the *architectural solutions* behind the proposed system, discuss the individual module technologies and provide details on the module orchestration. We also describe several *novel services* that include automated data harvesting from the web and social media, integrated user input collection via standard and customizable data types, easy to perform data analysis and visualization, publish/subscribe functionality to facilitate sharing of different facets and data shards, and access control mechanisms.
- (iii) We advocate the appropriateness of our approach for the cultural heritage domain and showcase different scenarios that highlight its usefulness for cultural data management.

From the argumentation presented above, it becomes clear that a free online system in the form of a *data lake* that is meant for *acquiring, storing, organizing, analyzing* and *sharing* heterogeneous, multi-faceted cultural heritage data would be a valuable asset to several different cultural heritage applications such as museum curation, user study management, bibliographical analysis, dataset management, and data integration. Moreover, such a system would be an invaluable source to many cultural informatics projects that either lack the resources or lack the IT expertise to design and deploy their own software and/or hardware infrastructure.

The rest of the paper is organized as follows. Section 2 discusses related work. Subsequently, Section 3 presents the overall system architecture and outlines the different modules as well as the respective services, while Section 4 provides an indicative case study during the Alpha testing phase of Hydria within the TripMentor project [45]. Section 5 presents various application scenarios in the cultural heritage domain and discusses how different stakeholders may benefit from using Hydria. Finally, Section 6 concludes this article and provides future research directions.

2. Related Work

In this section, we overview related research approaches that (i) are associated with the data acquisition and knowledge extraction for the cultural heritage domain based on social media, (ii) present the most prominent solutions in information systems meant for cultural heritage, and (iii) include museum and/or user recommendation information systems intended to improve visitors experience in cultural venues.

2.1. Social Data Management in the Cultural Heritage Domain

As an ever-growing number of social networks users constantly post opinions about cultural venues (by publishing reviews, describing their perceived experiences, using check-ins, subscribing to upcoming events, etc.), high volumes of data/content of great interest to cultural heritage applications is generated within popular social media platforms. The work presented in [16,17] aims to bridge the social media and cultural heritage domains and shows a way to stimulate history reflection by assembling games, social networks, history, and culture. In [20] the authors introduce the notion of the *prosumer*; the term refers to people that, besides consuming information, also produce new content when visiting cultural sites. A prominent paradigm in this line is the HeritageGO system proposed in [18]. HeritageGO tries to convert raw cultural heritage data coming from countries with a vast amount of cultural PoIs into meaningful digital information; towards this effort, the authors present social networks users as the main data harvesting lever and use metric quality models to filter the acquired data.

The approaches presented in [19–23] focus on improving cultural tourism and enhancing the visitors' experience by enriching the information concerning historical sites, monuments and other cultural PoIs with social media content, which is uploaded by social networks users and obtained using web mining techniques. To do so, identification methods that use geotagged multimedia data from social networks or location-aware services and sensors (e.g., GPS) attached on tourists' mobile devices have been implemented, while classification tools are used to rank the most relevant cultural heritage landmarks with respect to the user context (e.g., location) to render smart interactions among tourists and the cultural surrounding. The work presented in [46] proposes a Twitter big data-centric solution, which acknowledges a collection of Key Performance Indicators (KPIs) focusing on the quantity metric evaluation of cultural heritage sensitivity as interpreted by Twitter users, by merging natural language processing, semantic methodologies, location reports and time inspection.

Regarding the use of multimedia content in the cultural heritage domain, [47] is a prominent work; the authors describe the main aspects of multimedia social networks (MSNs), present the interactive system GIVAS, and highlight its importance to archeologists, cultural heritage researchers and tourists, as it consists a multimedia cooperative framework for managing, exploring, visualizing and sharing cultural heritage data. In [30], the author discusses how published multimedia data (especially videos) concerning cultural heritage venues and gathered from social media are of great importance to field consultants for generating 3D models (by using structure for motion methods).

PATCH [48] is a portable system able to harvest cultural heritage content from distributed and heterogeneous sources (such as social networks), to supply its users with profitable and personalized information and services based on their interests and their surroundings, and to provide data management, retrieval and analysis services. This system is the most conceptually and functionally similar work to the Hydria data lake; however PATCH was designed for the needs of a specific project and applied to a particular research study, while our work is an online, free, zero-administration data lake that offers both fundamental and advanced user and data/knowledge management functionality in the cultural heritage domain, able to be customized for the requirements of any cultural heritage project, and addresses all users, without requiring any IT background/skills.

2.2. Information Systems for Cultural Heritage

Information management in the cultural heritage domain concerns a cycle of organizational activity: the acquisition of cultural content from one or more sources, the storage and distribution of this data to those who need to evaluate it, and its final disposition through archiving. Over the years, many solutions aiming at the management, sharing and analysis of cultural heritage information have been proposed, while other investigations have tried to classify the variety of software tools and systems associated with the vast amount of data in the cultural heritage domain. The authors in [35] perform an itemized categorization of software tools and systems used in the cultural heritage area, associated with both spatial and temporal data. The contribution in [42] aims at exploring and classifying knowledge organization systems that are used in the cultural heritage field, while it applies extensive qualitative evaluation to the most prominent ones.

The work presented in [38] introduces the notion of *smart space* as a software development approach that enables creating service-oriented information systems for emerging computing environments for the Internet of Things (IoT), and considers the different principles to semantic-driven design of service-oriented information systems. In a similar spirit, [39] presents the ExhiSTORY infrastructure and discusses how sensors and the IoT can be used in cultural heritage sites so that exhibits communicate with the visitors towards generating rich, personalized, coherent, and highly stimulating experiences. In [49], a number of separate streams and current systems functionalities are examined through the usage of the European EU-CHIC framework, in order to achieve optimal suggestions for enhancing the management of cultural heritage data. The CHIS project [36] points at constructing an information system to assist operations that involve different user types in the cultural heritage domain, offering a scientific advancement that can improve personalized services in a business

environment. The research in [50] focuses on mobile software development for cultural information educational purposes and presents how mobile device users can be well-informed about cultural heritage sites when they visit them.

On the basis of several studies carried out on cultural landscapes in a spatial-planning perspective, [51] discusses the potential and limits of Geographical Information Systems (GIS) for supporting the territorialization of multidisciplinary landscape analysis for the management of a site of the UNESCO world heritage list, and proposes an approach for a GIS responding to landscape-oriented studies. Two similar approaches that propose a 3D representation of cultural objects, in order to facilitate researchers in determining both the relationships between data and the spatial relationships between cultural information, are presented in [31,52].

Recommendation systems are very popular in many scientific domains; in the cultural heritage field, recommendation systems constitute powerful tools that may help users improve their experience in cultural venues/PoIs. The work in [53] proposes that guidelines and recommendations should be used in all cultural infrastructures in Poland, associated with technical perspectives of digitization (such as technical and structural metadata, rules series, parameters and formats). The work in [54] describes an info-mobility recommendation system, coined TAIS, that assists tourists while traveling. TAIS can interpret user actions, uncloak their preferences, and suggest cultural sites in respect to the users' current locations (while also providing possible transportation means). The approach in [40] concerns a (big data) architecture that is able to host applications that retrieve data of the cultural heritage field from distributed and heterogeneous repositories; the authors introduce an innovative user-focused recommendation method for cultural element proposal to be applied on top of the data management infrastructure. Finally, the work in [41] introduces a novel ontology-based user method, pointing at improving personalized suggestions and users' visit experience by learning their background and interests.

2.3. Information Systems for Museums

In recent decades, a great number of cultural institutions (e.g., museums or national archives) integrate information systems in order to catalogue and document their exhibits, disseminate cultural information from their web sites and/or deliver informal education to their audience. Moreover, many information systems applied in institutions use Virtual Reality (VR) and Augmented Reality (AR) technologies aiming at enhancing tourists' experiences. The Digital Diorama [55] is a Mixed Reality (MR) system applied to museums focusing on rendering more features than existing dioramas in museum exhibitions, by prefetching background information. The work in [56] aims to enrich visitors' experiences in museum exhibitions by introducing a multichannel information system. The work in [57] introduces a virtual informal education system for the well-known ancient illustration of "Qing-ming Festival by the Riverside", by using VR technology to generate a wide, captivating, and responsive virtual environment. The work in [58] elaborates on the installation and integration of information systems in museums, identifying four success factors for relevant projects, while stressing the fundamental differences between museums and commercial companies. The approach in [59] describes the formulation and the adaptation of an AR-based system tailored for museum supervision; this research aims to narrow the gap between man and machine by applying instinctive as well as user-friendly synergies in an omnipresent computing environment. In a similar direction, TOMS [60] is a collaborative semantic-based system developed to provide sharing services of a vast variety of cultural heritage multimedia content between national museums in Thailand.

Personalization systems emphasize on tailoring a service or a product in order to accommodate particular individual preferences; presently, many museums and cultural institutions adopt personalization systems in order to offer custom-fit guidance and thus improve visitors' experiences. The work in [13] proposes a multimedia information system that is able to support multiple display devices, is built on top of an application server hosting plentiful digital content, and is presented to visitors in respect to their particular requirements. In [14], the authors demonstrate Future Worlds,

a knowledgeable game-based environment for cooperative feasibility investigations in science museums that is able to dynamically identify and adjust visitors' specific preferences while touring in the exhibition. The approach in [15] puts forward a web intelligent virtual assistant-based service for virtual museum explorations that can advance suggestions in respect to the museum exhibits and tailored to user's choices. In a similar spirit, [61] discusses experimental results obtained towards personalizing a museum visit based on gaming, using an approach relying on users' cognitive style, social networks, and recommendations. In [62], the authors, trying to connect cultural heritage, games and social networks, design social network games to be used for accomplishing user profiling and supporting museum visits; the games are also presented in a generic framework in cultural heritage. Finally, in [3], the authors investigate the use of indirect profiling methods through a visitor quiz, in order to provide the visitor with specific museum content, identify key profiling issues, and discuss guidelines towards a generalized approach for the profiling needs of cultural institutions.

Other works adopt different technological approaches, such as location-aware or spatial methods, in order to provide particular tour guidelines to their visitors. In [63], the authors investigate the practical usage of GIS as a tool, while inspecting how museums can adapt GIS technologies in separate operating zones. The work in [64,65] demonstrates a similar approach of a 3D information system, developed to manage cultural heritage information, which provides information layers that link with the exterior environment of the artifacts, following a similar to the GIS solution, in order to allow relationships between individual items.

To the best of our knowledge, Hydria is the first system that focuses on collecting, managing, analyzing, and sharing diverse, multi-faceted data in the cultural heritage domain and allows users without an IT background to deploy, populate, and manage their own data stores within minutes, alleviating the need to rely on expensive custom-made solutions that require IT infrastructure and skills to maintain.

3. System Architecture

The *Hydria data lake* allows users to (i) harvest and/or import data from structured and semi-structured data sources, (ii) collect user input data by resorting to several built-in and customizable data entry forms, (iii) store and manage collected data by organizing them in different big data management data stores (called *data ponds* in the Hydria terminology), (iv) share whole data sets or horizontal/vertical data shards via a powerful publish/subscribe mechanism that notifies users when other data ponds store data of interest, (v) search, filter and analyze data by using a powerful yet simple point-and-click mechanism that performs queries on the stored data and extracts the requested information in several visual representations and outputs, and (vi) perform basic and advanced user management tasks on the stored data. In Hydria, data ponds are custom-made database collections that are used to conceptually group data within a specific cultural heritage application. Figure 1 provides a high-level view of the system architecture, of the different services and functionalities implemented, and their conceptual organization within the Hydria data lake. In what follows, we present in detail the different services and modules that comprise the Hydria ecosystem and briefly outline the functionality and added value of each module.

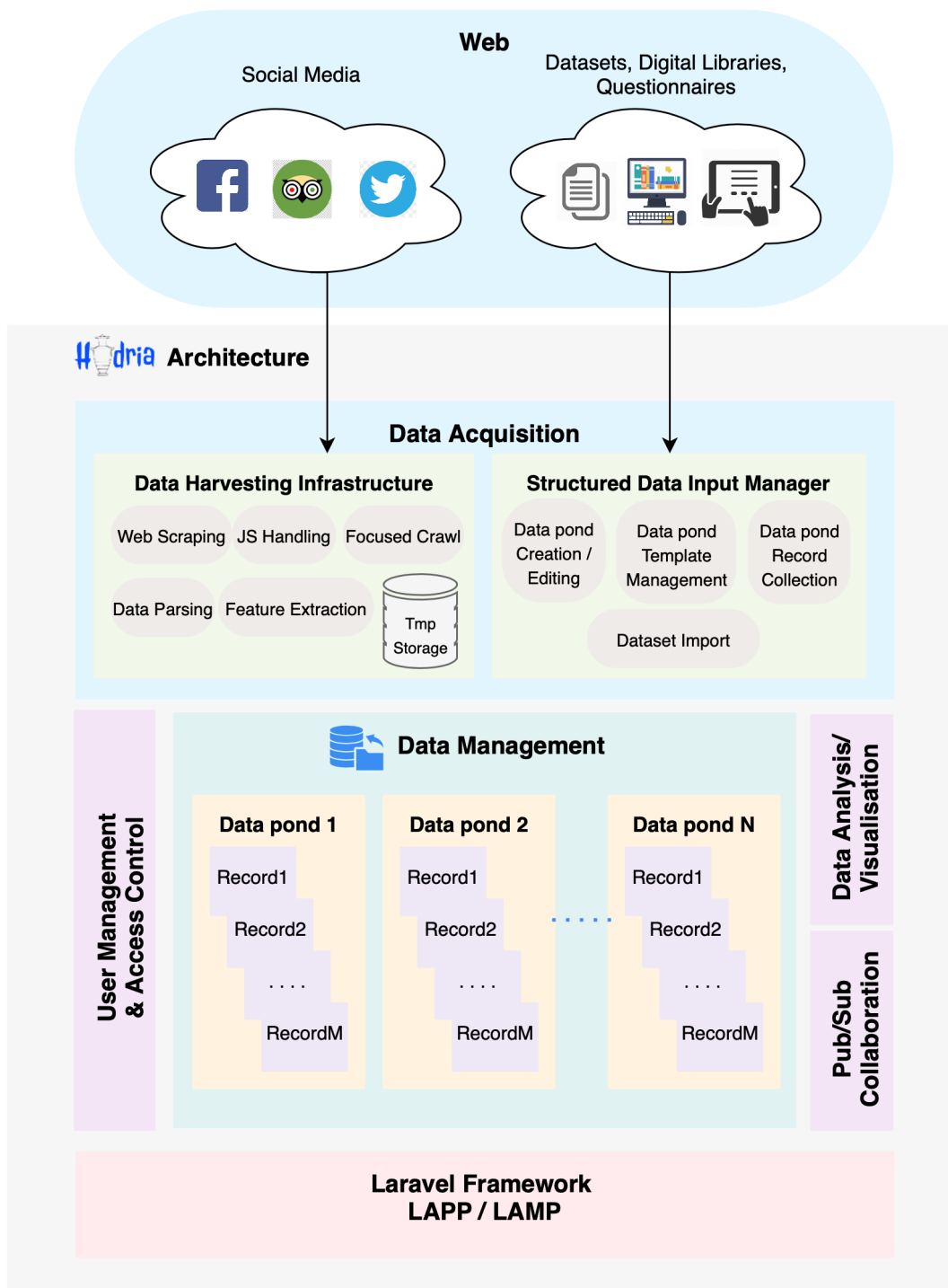


Figure 1. Hydria data lake architecture.

3.1. The Data Acquisition Module

Information of interest to many cultural heritage applications are present in many different websites ranging from online encyclopedias like Wikipedia [66], online digital libraries like Europeana [67] or DBLP [68], to portals and directories like governmental sites on culture (e.g., Odysseus [69], the portal of the Greek Ministry of Culture and Sports) and WikiCFP [70]. However, presently, a lot of data/content of great interest to cultural heritage applications may be also discovered within popular social media like Facebook, TripAdvisor or Twitter. A vast number of people (of any origin, language or educational profile) have accounts in (typically more than one) social media

platforms, and use them to post their opinion about cultural heritage venues by publishing reviews, to describe their perceived experiences by uploading posts, to designate favorite destinations and provide useful points of interest (PoIs) by using “check-ins”, or to keep up to date by subscribing to upcoming events hosted by different types of venues [71–73].

To cover this widespread need for data acquisition within the cultural heritage domain, Hydria provides a flexible yet powerful Data Acquisition module that is conceptually separated into two distinct submodules:

- (i) The *Data Harvesting submodule*, which allows Hydria users to setup and deploy automated data collection crawlers and web scrapers (spiders) that are able to navigate the web and popular social media platforms, discover and harvest content of interest, and store the harvested data to the Hydria data lake. The Data Harvesting submodule is discussed in more detail in Section 3.1.1.
- (ii) The *Structured Data Input submodule*, which allows Hydria users to import whole datasets into Hydria and collect user input data by exploiting several built-in and customizable data entry forms. The Structured Data Input submodule is elaborated on in Section 3.1.2.

3.1.1. The Data Harvesting Submodule

The Data Harvesting submodule implements several distinct services that may be invoked by Hydria users to initiate automated data collection. At the heart of this submodule lies the web scraping service that extends the basic components of the open source web scrapper Scrapy [74,75] for scraping social media content. Currently, the web scraping service supports two of the most popular social media platforms, Facebook and TripAdvisor, while support for more is under development. Setting up the web scraping service only involves providing the initial seed URLs from the aforementioned social media; the service automatically recognizes which social platform and what type of data (e.g., venues, PoIs, reviews) is targeted for data harvesting and launches the appropriate web scrapper instance. To tackle user privacy issues that may arise during spidering, the web scraping service provides no spidering options for individual users (i.e., one cannot scrape user pages) and supports only the collection of aggregate values and fields that are general enough so that no person may be identifiable by reasonable means. The data flow process is controlled by the scrapper execution engine that is responsible for the data flow between all components of the Hydria system and is summarized below:

1. The engine receives a *scrape* request, which is a custom-made class that is used to parse responses and extract scraped data, and pushes it to the scheduler for later use; then the engine expects to read scheduled tasks from the Scheduler, in order to process them further.
2. The scheduler performs scheduling on the available requests and returns to the engine the next request to be further processed (downloaded). Then, the engine forwards the request to the downloader through an appropriate message broker. When the page download is completed, the downloader creates a response and forwards it back to the engine via the message broker.
3. Once the engine gets the response, it moves it to the spider for processing through the message broker. When the spider processing is over, the scraped data are returned and *new* requests are sent to the engine via the message broker.
4. The engine initially passes the scraped data to the item pipeline, which is the software that is in charge of processing the data after they have been extracted by the spiders, then dispatches the processed request to the scheduler and subsequently awaits the next request to scrape.
5. The above steps are performed iteratively, until no more scheduled requests are available. Note that all the extracted items are temporarily pushed in a persistent local data store (one per initiated scraper).

This procedure can be applied straightforwardly on plain HTML pages that are received from the server; however, to enhance the user experience, practically all major social media sites employ JavaScript, making their content more interactive. This practice, however, poses challenges to the

scraping procedure; to overcome these challenges, we have also developed a JavaScript handling service by adapting and integrating the Selenium library [76]. The Selenium library provides a useful tool for data harvesting and web scraping: the Selenium renderer uses a web browser engine to render a given URL and mimics human behavior on the web page. This allows the web scraping (and the crawling) service to interact with JavaScript functions that exist on the target website (e.g., infinite scrolling) and avoid unnecessary hold-ups in the spidering process. The JavaScript handling service uses the Google Chrome web driver for the Selenium renderer.

Apart from spidering popular structured social media platforms, the Data Harvesting submodule contains also a *focused crawl* service that is designed to perform *thematic* crawls on the clear web with the purpose to discover new resources that may contain cultural heritage data of interest. From the Hydria user side, setting up the focused crawling service only involves providing a relevant (to the task) query (e.g., “cultural heritage” or “archeological museum”) that will be used to produce the initial crawl seeds. The underlying crawling infrastructure is based on the ACHE crawler [77], one of the most popular focused crawlers [78] available, which prioritizes URLs in the crawl frontier and categorizes the crawled pages as relevant or irrelevant using machine learning-based techniques. To direct the crawl towards topically relevant websites (i.e., websites with content relevant to cultural heritage) we use an SVM classifier, which is trained by resorting to an equal number of positive and negative examples of websites that are used as input to the *model builder* component of ACHE. Subsequently, the *seed finder* [79] component is used to aid the process of locating initial seeds for the focused crawl on the clear web; this is achieved by combining the pre-built classification model with the topic-related user-provided query discussed above. Since the crawled websites may be anything from blog posts to organizational web pages and do not have a predetermined structure (unlike the social media pages), the collected content is only parsed to remove HTML markup and is stored as raw text in the Hydria data lake.

Finally, the Data Harvesting submodule contains two auxiliary services (namely data parsing and feature extraction) that are used to extract textual content and features from the harvested websites.

3.1.2. The Structured Data Input Submodule

Apart from the crawling and web scraping functionality that were presented in the previous section, Hydria also supports structured data input via the relevant submodule. Structured data input supports several services that provide users with the necessary functionality to (i) import data from a structured CSV/XML/JSON formatted dataset, (ii) create (in a stand-alone data pond) a questionnaire-style form that may be used in surveys and the associated user answers, and (iii) reuse all or part of these questionnaires via the creation and management of templates. Each of the aforementioned services is associated with a data pond that is created to allow users to manage the stored data (more details about data pond creation and management are given in Section 3.2). To achieve this functionality, the Structured Data Input submodule implements several distinct services as follows.

The *Dataset Import service* is used to automatically load/store a structure dataset into a Hydria data pond; the service automatically matches the columns of a CSV/XML/JSON tagged file with the pre-specified data pond fields and for each data item (typically a row in the CSV file or element under the root of the XML/JSON document) a new record is created and stored in Hydria under the corresponding data pond. Notice that this service may be also used as a separate stand-alone step for importing harvested web/social media content that has undergone processing outside of Hydria; i.e., when the harvesting task is over, the user may select to process the downloaded content outside of Hydria and subsequently manually load the result of this intermediate processing into a separate data pond.

The rest of the available services target the creation of questionnaire-style forms that allow Hydria users to create and store data collection tasks that involve electronic input of end-users into structured forms (e.g., surveys, end-user evaluations, museum experience records, etc.). In order to facilitate

data pond creation and reuse of common parts between constructed questionnaires, Hydria supports (via an appropriate template management service) the creation and use of *templates* that can be shared or reused between different data ponds. This functionality is native in Hydria and is tightly coupled with (i) the creation, maintenance and analysis of data ponds (described in detail in Section 3.2) and (ii) the access control mechanism for the different data ponds (described in detail in Section 3.5).

3.2. The Data Management Module

The Data Management module supervises the creation, editing, organization and management of the data ponds, and performs all necessary storage and retrieval operations to the database back-end (i.e., manages the stored data related with a specific data pond or a specific record). It supports a flexible, adaptive and intuitive way for designing and composing a data pond or a data pond template.

The Data Management module employs several different services that allow Hydria users to create and edit data ponds. Such activities are supported by an easy-to-use wizard mechanism that guides the Hydria user through the whole process. Building a new data pond/template involves defining a title and a description for it; subsequently the user specifies the different fields for the data pond (i.e., the attributes to be stored) by providing for each field its textual description and its type. According to the selected attribute type, hidden fields or dialogs appear for inserting more specific information about the attribute (e.g., if the attribute is of type *multiple choice*, the user has to fill up a list of values or select one of the existing template lists). The available data types that currently Hydria supports are as follows: title (this field is not fillable, although it is used to separate data pond sections), text, integer, decimal, date, multiple choice, picture drawing, image file, and complex data types.

Complex data types are a construct provided to allow for more efficient modelling of cases where a group of fields appear multiple times within a document, across different documents in the same data pond, or even across documents in different data ponds. Examples of such cases may be interpretations of cultural items (with each interpretation having an author, a summary, and extended analysis and supporting documents, and each cultural item being potentially subject to multiple interpretations), or company addresses (each address consists of a street name, a number, a city, a zip code and a country, and a single company may have multiple addresses). To introduce a complex data type, a user needs to provide the specification of a recurring attribute with more than one fields. Please note that the complex data type definition may be subsequently modified to change the attribute order and/or edit or delete a specific attribute by using the corresponding controls on the wizard; any changes are reflected to data ponds using the modified complex data type. The advantages of creating and using complex data types include better data modelling, increased flexibility in the design of data ponds with complex/recurring attributes, elevated knowledge capture capabilities and, consequently, the ability to formulate more expressive and semantically rich queries.

To ensure data consistency across data ponds and to enhance data integrity and input validation, Hydria natively supports the following two features:

- (i) Hydria allows users to *share* data pond templates by supporting the reuse of all or a part of data pond fields (e.g., demographic data in questionnaires) across different data ponds. To promote this functionality, the data pond creation service prompts the user to consider reusing one of the available data pond templates before creating a new data pond.
- (ii) Hydria provides users with the ability to *dynamically* create, store and edit drop-down lists of elements. To do so, the user specifies a unique name for the drop-down list and enters the list elements. Subsequently, when defining a multiple choice field (i.e., attribute), the user needs to set the data type to multiple choice and either select one of the stored drop-down lists from the pop-up window or dynamically create a new one that is thereafter stored along with the other drop-down lists for further (re)use.

Finally, notice that Hydria has no direct policy on how one stores or uses the stored data. This is particularly important in the case of evolving data(sets) where the user has many different options

to store, monitor, or analyze data evolution. In particular, she has the option to create different data ponds for different snapshots of the data, use different fields to model data evolution within the data pond, or store solely the differences between various data snapshots.

3.3. The Data Analysis Module

The Data Analysis module supports search, filtering and analysis of the data stored in each data pond of the Hydria data lake. It provides a powerful yet easy-to-use data manipulation and query mechanism that allows users to formulate queries against the data ponds involving selection, projection, grouping and ordering operations through simple point-and-click interaction and without requiring any background knowledge of SQL (Figure 2). The provided mechanism also contains in-context explanations for the different data analysis elements and provides online help with examples. This functionality is targeted mainly towards users with limited relevant experience. Besides analyzing each data pond, the module also offers data extraction and visualization functionality in a variety of different formats such as histograms, pie charts, (heat) maps, (stacked) bars/columns, area/mekko/bubble charts, scatter plots, and various file types (like CSV/TSV and raw text). The implemented data visualization component involves a three-step process where the user (i) defines the type of the chart to be exported, (ii) specifies the base dataset by selecting the data pond and the data pond field(s) that will be used to create the chart, and (iii) may apply filtering conditions (restrictions) on the chosen dataset.

The screenshot shows a web interface titled "Charts & Statistics". It contains several configuration options:

- Select chart type:** A dropdown menu with "Bar" selected.
- Select datastore:** A dropdown menu with "testing datastore" selected.
- Select question:** A dropdown menu with "age" selected.
- Filtering options:**
 - Add (where) value condition?
 - Group data by value?
 - Order the data in ascending or descending order?
- Condition field:** A field labeled "Select condition for value:" with a dropdown arrow and a text input containing "30".
- Second question options:**
 - Add another question dataset to the chart?
 - Select question 2:** A dropdown menu with "total exhibition review" selected.
 - Add (where) value condition?
 - Group data by value?
 - Order the data in ascending or descending order?
- Show chart:** A blue button at the bottom right.

Figure 2. Filter and project records.

3.4. The Publish/Subscribe Module

The Publish/Subscribe module is essentially an efficient and easy-to-use collaboration tool that allows users to share their collected datasets (or parts of them), as well as to discover and access datasets of other users within the Hydria ecosystem. This tool is not intended for use by the Hydria system end-users (i.e., people participating in a Hydria survey, or museum visitors that provide

feedback via a Hydria questionnaire), but rather targets other user categories like curators/super-users (see Section 3.5 for a detailed description of the Hydria user roles). Using this functionality involves the following two-step process:

1. A user may use the data pond search functionality to look for data ponds stored within the Hydria data lake that satisfy a given keyword query; after selecting one or more data ponds that are included in the result, she may send a subscription request to the owner of the specific data pond(s) asking for permission to access the data pond's schema definition, i.e., the list of attributes of the data pond, their descriptions and data types. If the owner of the data pond accepts the subscription request, the user is eligible to access the data pond's schema definition.
2. After examining the data pond schema, the user may decide to request access to specific attributes of the data pond at record level. In this case, she may select one or more attributes of the targeted data pond and send a follow-up subscription request to the owner of the data pond. Once the owner receives the new request, she is able to either deny, accept the request as is, or remove any of the attributes that should not be shared at record level, and confirm the sharing of the remaining ones.

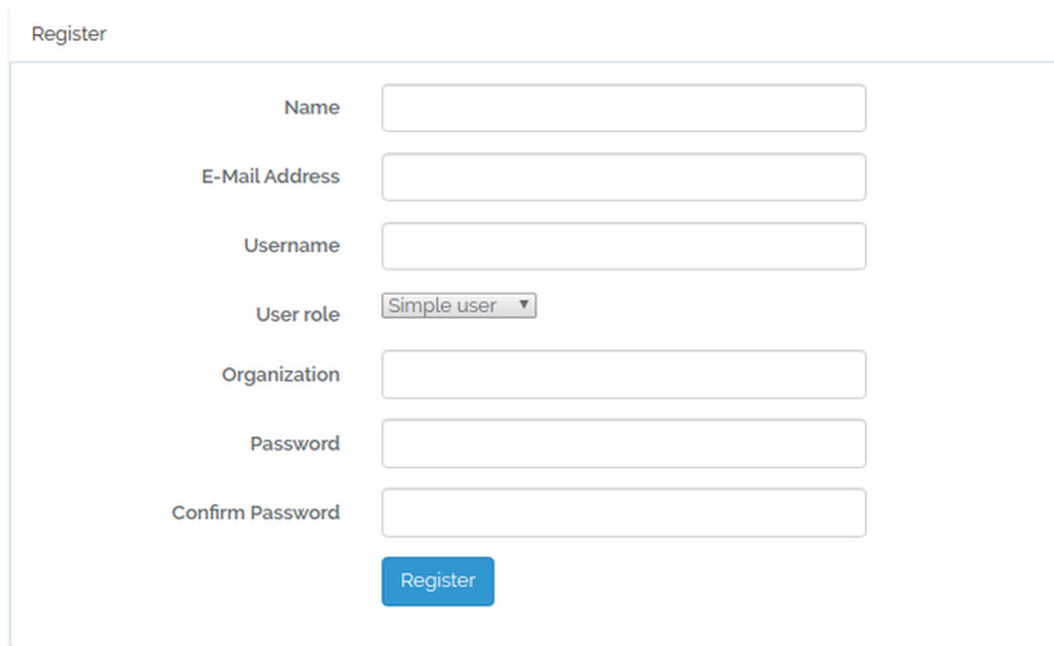
When both stages have been completed, the user is able to access the subscribed records for preview, incorporate them into her own data ponds, or create charts and visualizations using the data analysis and visualization tool. Moreover, the subscribed user will be notified whenever new records that match her subscription are incorporated in the subscribed data pond.

3.5. The User Management Module

The User Management module is responsible for performing basic and advanced user management tasks, such as manage users in the Hydria community, assign user privileges and permissions, and perform access control on data ponds and the stored data within them. It supports the following three types of users:

- (i) *System administrators*, who are able to access, create, edit, preview, delete and filter all the data ponds and records that are stored in the data lake; they are also capable of managing all types of Hydria users, as well as the user-role assignments.
- (ii) *Power users*, who are typically curators in charge of one or more data ponds in the Hydria ecosystem. They are typically able to create new data ponds, initiate the collection of data from different online sources such as social media, the web, existing datasets, or end-users via questionnaires and surveys. They have access privileges in the data ponds that they create and in the records within these data ponds; they may analyze, filter and visualize the stored data, and collaborate with other power users in Hydria in the context of data sharing. A power user may also request to link specific end-users to new or existing data ponds.
- (iii) *End-users*, who may participate in surveys and questionnaires issued by Hydria power users and view/edit their own data; an end-user is neither allowed to create new data ponds, nor to view data of other end-users within the same data pond. They may, however, use the analysis tools to perform limited analysis tasks on their own contributed data.

Figure 3 presents the UI for the creation of a new user from the administrator point of view in Hydria.



The image shows a web form titled "Register". It contains the following fields and elements:

- Name**: A text input field.
- E-Mail Address**: A text input field.
- Username**: A text input field.
- User role**: A dropdown menu with "Simple user" selected.
- Organization**: A text input field.
- Password**: A text input field.
- Confirm Password**: A text input field.
- Register**: A blue button at the bottom.

Figure 3. User creation form.

3.6. Implementation Aspects

Hydria has been entirely developed by open source software. The Data Acquisition module uses the Linux/Apache/MariaDB/PHP (LAMP) solution stack for temporary storage of the extracted data and was developed using Python tools [80]. The rest of the modules were built by using the Laravel Framework [81] and use the Linux/Apache/PostgreSQL/PHP (LAPP) solution stack as the back-end database infrastructure. Also, many of Hydria functionalities were developed using JavaScript/JQuery/AJAX.

4. The TripMentor Case Study

In the context of Alpha testing the Hydria platform, we made the system available to partners within the TripMentor project [45] that aims at creating a tourist guide for the region of Attica, Greece. The partners had varying levels of IT expertise, ranging from relatively experienced to very experienced, and were asked to use Hydria for collecting, storing, and managing data relevant to the project, by deploying the different services offered by Hydria and using the functionality provided. In the following, we report the results for this case study, as these were drawn from field observations as well as from the analysis of usage logs and collected data.

The TripMentor partners used Hydria to navigate within TripAdvisor [82] and Facebook [83] and detect content relevant to the tourism domain; the interest of focus was primarily points of interest (PoIs) located in the *Attica region in Greece*, due to the nature of the TripMentor project. The data retrieved was stored in the Hydria data lake, and the stakeholders involved used Hydria functionality to design the necessary data ponds, modify selected data records, mine and visualize information from the stored data, and export files for further analysis.

4.1. Data Harvesting

To crawl the web for content and data relevant to the TripMentor project, the integrated ACHE crawler was used within the Hydria environment. To set up the crawler, a number of relevant pages within the domain of interest were used as seed URLs; such URLs included tourist articles and blog posts related to cultural and tourist activities in the Attica region. To implement the page classification required by ACHE, different features of the page URL, title and content were exploited, which were

determined after an examination of the patterns followed by the pages in different social networks. For instance, for the TripAdvisor spider the following features were used:

- (i) The page URL was matched against the regular expression patterns `.*attraction/.*` and `.*attica/.*` (note that in Hydria regex expressions are case insensitive), using the `url_regex` classifier type, since the URLs of the PoIs in TripAdvisor start with the word *attraction* and contain the name of the region (*Attica*, in our case).
- (ii) The page body was matched against the regular expression pattern `.*greece/.*` using the `body_regex` classifier type, to ensure that the word *Attica* found in the URL actually refers to the Greek region (and not e.g., to the Attica city in the State of NY).

Similarly, for the Facebook crawlers patterns from the seed URLs were exploited; for instance, to identify the activities available in the city of Athens, Greece (which is located in the Attica region), the seed URLs start with the string *Things-to-do-in-*, which is followed by the city name (*Athens*), and this is in turn followed by the name of the country (*Greece*).

Notably, a multitude of additional operators are available for more complex classification tasks, which include matching of the title (`title_regex` classifier type), combination of regular expression matches through the AND and OR operators, or using machine-learning-based text classifiers (SVM, Random Forest) using the `title_regex` classifier type.

Having collected the seed URLs, six distinct spiders were developed, in an effort to cover a wide spectrum of events and many different scopes; four of them targeted data harvesting from the Facebook platform, while the remaining two were deployed over TripAdvisor. Table 1 summarizes the outcome of the data harvesting process.

Table 1. TripMentor-related data ponds records after Hydria harvesting process.

Data Ponds	Records
facebook_venues	10,405
facebook_posts	139,880
facebook_comments	203,523
facebook_events	150
tripadvisor_venues	6869
tripadvisor_user_reviews	298,769

4.1.1. Facebook Spiders

In this section, we discuss the deployed Facebook spiders for the TripMentor case study and present some initial statistics and insights on how the Hydria social media spiders may be used within the cultural informatics context. Please note that the spiders and data ponds were setup and deployed by the TripMentor project partners; our statistics and observations are based on usage logs and the schema analyzes of the data ponds used to store the collected data.

The first spider, which constitutes the initial setup within Hydria, was deployed over Facebook and extracted a total number of 10,405 different PoIs in the Attica region; a time period of approximately 48 hours was needed to conclude the data harvesting operation, and the collected PoIs are categorized as shown in Table 2. For each of these venues, the spider retrieved and stored in a Hydria data pond the following fields from the related Facebook profile pages: the venue name, the venue unique ID as stored in the Facebook platform, the hours/days of the week that the venue is open to visitors, the venue website, the venue phone number, the registered email address, the physical address of the venue, the total number of check-ins for the venue (i.e., the visitor traffic), the average review score of the venue, the venue category, and the geographical coordinates (latitude and longitude) of the venue. Please note that all the collected data were about venues and cultural events, and no personal or user-specific data were harvested or stored.

Table 2. Different PoIs in the Attica region extracted by the first FB spider.

Venue Category	# of PoIs
Arts and Entertainment	2116
Breakfast and Brunch Restaurants	114
Cafe	2854
Hotels	778
Landmarks	614
Museums	210
Parks and Outdoors	712
Restaurants	3007

Having retrieved the PoIs unique IDs from the previous process, a new spider that generated the venues profile page URLs using the venue IDs was created and launched within Hydria for a subset of the collected PoIs (approximately 1.7 K PoIs). The spider was then deployed and collected around 140 K posts related to the targeted PoIs in a time frame of around 168 h (one week). All retrieved posts were stored in a separate data pond within Hydria, along with the following metadata: The post unique ID, the profile source of the post, the profile that shared this post, the upload date of the post, the post text, the total number of reactions and the number for each individual reaction type (e.g., likes), and the URL of the post. Notice that posts and the collected post sources refer to venues and cultural events and are not related to persons or personal data, while the collected reactions correspond to aggregate numbers and cannot be traced back to individual users. In particular, regarding the information on the profile that shared the post, only the id of the profile was collected and transformed using an one-way function, hence the data cannot be associated with the original profile; however maintaining the ability to determine whether two posts were posted by the same user profile.

By using the generated venue profile page URLs, another spider was setup and executed within the Hydria environment; it employed 587 seed URLs and was able to retrieve around 240 K user comments within a time frame of 168 h (one week). The collected data was also stored in a separate Hydria data pond and contained fields like: the date that the comment was posted, the total number of the reactions in each comment, the comment text, and the comment URL. Notice that the collected comments and the related metadata do not contain user information or personal data; no user IDs were collected and our logs show that all user references in the comment text were deleted by using the Hydria text cleaning (regex-based) functionality.

Next, profiles of venues that are known as major event organizers were used as initial seeds to bootstrap and launch an event harvesting spider that was able to extract a few hundred upcoming events and their relevant event cards. The seed URLs were identified by the TripMentor stakeholders by manually inspecting the collected venues and using tacit knowledge regarding the major event organizers in the region of Attica. Again, the harvested data were stored in a separate Hydria data pond, with the fields stored within this data pond including the event name and date(s), the physical address of the event, the number of people interested to visit this event, the URL of the event, the unique identifier of the PoI where this event was found, the unique identifier of the event, and the text description of the event. Please note that only aggregate numbers on the number of individuals interested in the event are harvested and no personally identifiable information is either collected or stored within Hydria.

4.1.2. TripAdvisor Spiders

In this section, we discuss the deployed TripAdvisor spiders for the TripMentor case study; this set of spiders is used to showcase the versatility and usefulness of the data harvesting component. As with the Facebook spiders, we had no control over the TripAdvisor spiders deployed and the data ponds created; all setup, deployment and data manipulation was done by the TripMentor project partners. The statistics and metadata presented in this work were drawn from usage logs and the data pond schemas.

The first spider was setup and deployed over TripAdvisor aiming to extract PoIs in the Attica region; after a runtime of around 48 h, it collected information about 7 K different PoIs, belonging in a vast variety of different categories (as identified by the respective TripAdvisor field) including monuments, museums, landmarks, natural reserve sites, parks and water parks, different types of restaurants, cafes, etc. For each one of the collected PoIs, the following fields were stored in the Hydria data pond: the venue name in different languages, the overall venue review score, the total number of venue reviews, the ranking of the PoI with respect to other PoIs of the same category in the same broader area (e.g., “4th out of 10 restaurants within the district”), the categories that this PoI appears in, the physical address of the PoI, the PoI phone number, and the TripAdvisor URL of the PoI.

Subsequently, a spider to collect the individual user reviews (without the associated user information) for the 7 K venues/PoIs that were previously harvested was created. After setup and deployment, the spider was able to extract around 300 K individual user reviews in a time frame of around 120 h (five days); the fields of each review that were detected and stored in the respective Hydria data pond are: the review title, the review text, the date of the review, and the review score (in the TripAdvisor bubble format). For the purpose of better understanding the user background, the spider also collected anonymized information about each user that posted a review. This data does not contain any personally identifiable information and was limited on purpose to the following general fields and aggregate metrics: the user country of origin, the total number of user votes (rounded to the nearest ten), the total number of TripAdvisor contributions (rounded to the nearest ten), general user tags (like “history lover”), and generalized age ranges of users. Notice that this information is common among a vast number of TripAdvisor users and cannot be used to personally identify an individual.

The spider examples presented above show only a fraction of the functionality that is available within Hydria. Apart from focused crawlers to crawl the Web for relevant pages and Facebook or TripAdvisor spiders to harvest data from the respective social media sites, Hydria also provides Twitter monitors. These monitors use the Twitter [84] search or stream API to perform keyword-based filtering of published tweets; all retrieved tweets may be subsequently stored in appropriately configured Hydria data ponds for further processing.

4.2. Importing Datasets and Adding/Modifying Records

Besides automated data harvesting, Hydria also offers a file import service (as described in Section 3.1.2) that allows users to easily import their own datasets into a Hydria data pond. In our case study, we asked partners from the TripMentor project to use the file import tool to incorporate a new CSV dataset into Hydria. One of the project partners responded and reported that they used Hydria to store a home-brewed list of tourism stakeholders (who could be interested in the project results) in a Hydria data pond. The imported dataset consisted of several hundreds of individual records of companies and stakeholders operating in the tourism sector alongside their contact information, and was shared with the rest of the TripMentor partners by defining the appropriate access rights.

Subsequently, other project partners were able to browse the created data pond with the tourism-related companies and add or modify records as needed by filling out the different data pond fields, tagging records with notes for the data pond curator, and save any desired changes in the specific data pond. Figure 4 gives an overview of the aforementioned data pond; at the top of the figure, controls providing access to all available data pond functionality are presented to the user. Figure 5 shows the add record tool where the user may insert individual records, providing data for a multitude of fields of different types (free text; number; drop-down lists; complex types; and an image field).


Tourism Stakeholders users			Tourism Stakeholders all records			Import records from CSV			Add record		
Tourism Stakeholders											
A case study for tourism stakeholders in the Attica region in Greece											
No.	Question text										Answer type
1	Company name										Text
2	Company category (service type)										Multiple Choice
3	Days & Hours open										Text
4	Employees number										Integer
5	Contact Info										Complex Type
6	Company Branches Info										Complex Type
7	Customer service review										Decimal
8	Company founded										Date
9	Image file upload										Upload Image

Figure 4. A data pond example.

Tourism Stakeholders

Create new record

No.	Question/Title text	Answer																																								
1	Company name	Attica Travel Agency																																								
2	Company category (service type)	Transfer & Guiding Services																																								
3	Days & Hours open	Mon - Sat 09:00 - 17:30																																								
4	Employees number	35																																								
5	Contact Info	<table border="1"> <tr> <th>1</th> <th>Address:</th> <th>Phone:</th> <th>E-mail:</th> <th>Site:</th> </tr> <tr> <td></td> <td>Philelinon :</td> <td>+30 210 329</td> <td>info@atticat</td> <td>https://ww</td> </tr> </table>	1	Address:	Phone:	E-mail:	Site:		Philelinon :	+30 210 329	info@atticat	https://ww																														
1	Address:	Phone:	E-mail:	Site:																																						
	Philelinon :	+30 210 329	info@atticat	https://ww																																						
6	Company Branches Info	<table border="1"> <tr> <th>1</th> <th>Branch name:</th> <th>Address:</th> <th>Phone:</th> <th>Email:</th> </tr> <tr> <td></td> <td>Kifissia Offic</td> <td>Kifissias Av.</td> <td>+30 210 803</td> <td>kifissia@attic</td> </tr> <tr> <th>2</th> <th>Branch name:</th> <th>Address:</th> <th>Phone:</th> <th>Email:</th> </tr> <tr> <td></td> <td>Ag. Paraske</td> <td>Ag. Ioannou</td> <td>+30 210 600</td> <td>agiaparaskr</td> </tr> <tr> <th>3</th> <th>Branch name:</th> <th>Address:</th> <th>Phone:</th> <th>Email:</th> </tr> <tr> <td></td> <td>Glyfada Offic</td> <td>Grigoriou Li</td> <td>+30 210 894</td> <td>glyfada@att</td> </tr> <tr> <th>4</th> <th>Branch name:</th> <th>Address:</th> <th>Phone:</th> <th>Email:</th> </tr> <tr> <td></td> <td>Piraeus Offic</td> <td>Iroon Politel</td> <td>+30 210 41 6</td> <td>piraeus@att</td> </tr> </table>	1	Branch name:	Address:	Phone:	Email:		Kifissia Offic	Kifissias Av.	+30 210 803	kifissia@attic	2	Branch name:	Address:	Phone:	Email:		Ag. Paraske	Ag. Ioannou	+30 210 600	agiaparaskr	3	Branch name:	Address:	Phone:	Email:		Glyfada Offic	Grigoriou Li	+30 210 894	glyfada@att	4	Branch name:	Address:	Phone:	Email:		Piraeus Offic	Iroon Politel	+30 210 41 6	piraeus@att
1	Branch name:	Address:	Phone:	Email:																																						
	Kifissia Offic	Kifissias Av.	+30 210 803	kifissia@attic																																						
2	Branch name:	Address:	Phone:	Email:																																						
	Ag. Paraske	Ag. Ioannou	+30 210 600	agiaparaskr																																						
3	Branch name:	Address:	Phone:	Email:																																						
	Glyfada Offic	Grigoriou Li	+30 210 894	glyfada@att																																						
4	Branch name:	Address:	Phone:	Email:																																						
	Piraeus Offic	Iroon Politel	+30 210 41 6	piraeus@att																																						
7	Customer service review	4.5																																								
8	Company founded	01/01/2001																																								
9	Image file upload	Choose File atticaTravel.jpeg																																								



Write notes for this record (optional):

Figure 5. Manual record creation.

In this figure, we can observe the use of the complex data types feature, to design groups of input fields which may also be recurring. For example, in the aforementioned dataset the curator may select to use a complex data type to jointly represent longitude/latitude information (under the complex type named “coordinates”), or branch information (comprising fields “branch name”, “address”, “phone” and “email”). Additionally, the curator may use the latter complex type (branch information) as a recurring input field, to model contact information about a company that has multiple branches. Recurring input fields effectively model the master-detail relationships between parent and child objects (one-to-many relationships). In the future, we plan to support more complex data types, such as voice and video recording, time-series and streaming data.

4.3. Reusing Data Ponds and Data Pond Templates

To support the reuse of all or a part of a data pond (e.g., the contact details) between different data ponds, the notion of data pond templates is introduced in Hydria. When a curator creates a new or edits an existing data pond, Hydria prompts her to (re)use one of the available data pond templates; Figure 6 presents the editing of a data pond with seven different data type fields from the administrator’s UI.

Tourism Stakeholders Import template Assign users to this datastore

A case study for tourism stakeholders in the Attica region in Greece

Search for questions...

No.	Question text	Answer type	Actions																				
1	Company name	Text	Edit Delete																				
2	Company category (service type)	Multiple Choice	Edit Delete																				
3	Days & Hours open	Text	Edit Delete																				
4	Employees number	Integer	Edit Delete																				
5	Contact Info <table border="1"> <tr> <td>1</td> <td>Text: Address</td> <td>Text: Phone</td> <td>Text: E-mail</td> <td>Text: Site</td> </tr> </table>	1	Text: Address	Text: Phone	Text: E-mail	Text: Site	Complex Type	Edit Delete															
1	Text: Address	Text: Phone	Text: E-mail	Text: Site																			
6	Company Branches Info <table border="1"> <tr> <td>1</td> <td>Text: Branch name</td> <td>Text: Address</td> <td>Text: Phone</td> <td>Text: Email</td> </tr> <tr> <td>2</td> <td>Text: Branch name</td> <td>Text: Address</td> <td>Text: Phone</td> <td>Text: Email</td> </tr> <tr> <td>3</td> <td>Text: Branch name</td> <td>Text: Address</td> <td>Text: Phone</td> <td>Text: Email</td> </tr> <tr> <td>4</td> <td>Text: Branch name</td> <td>Text: Address</td> <td>Text: Phone</td> <td>Text: Email</td> </tr> </table>	1	Text: Branch name	Text: Address	Text: Phone	Text: Email	2	Text: Branch name	Text: Address	Text: Phone	Text: Email	3	Text: Branch name	Text: Address	Text: Phone	Text: Email	4	Text: Branch name	Text: Address	Text: Phone	Text: Email	Complex Type	Edit Delete
1	Text: Branch name	Text: Address	Text: Phone	Text: Email																			
2	Text: Branch name	Text: Address	Text: Phone	Text: Email																			
3	Text: Branch name	Text: Address	Text: Phone	Text: Email																			
4	Text: Branch name	Text: Address	Text: Phone	Text: Email																			
7	Customer service review	Decimal	Edit Delete																				
8	Company founded	Date	Edit Delete																				
9	Image file upload	Upload Image	Edit Delete																				

Your question text here Choose answer type ↓

[+ Save question](#)

Figure 6. Data pond editing.

4.4. Visualizing Information

Having populated the Hydria data ponds with data concerning different TripMentor needs (e.g., PoIs and related information from Facebook and TripAdvisor spiders, and tourism stakeholder data), the TripMentor partners were able to search, analyze, filter and visualize the stored data by using the powerful yet easy-to-use Hydria data analysis module. The analyzed and visualized information is presented in Figure 7 and is entirely produced using tools provided within the Hydria environment (i.e., no external visualization modules were used to create the graphs shown).

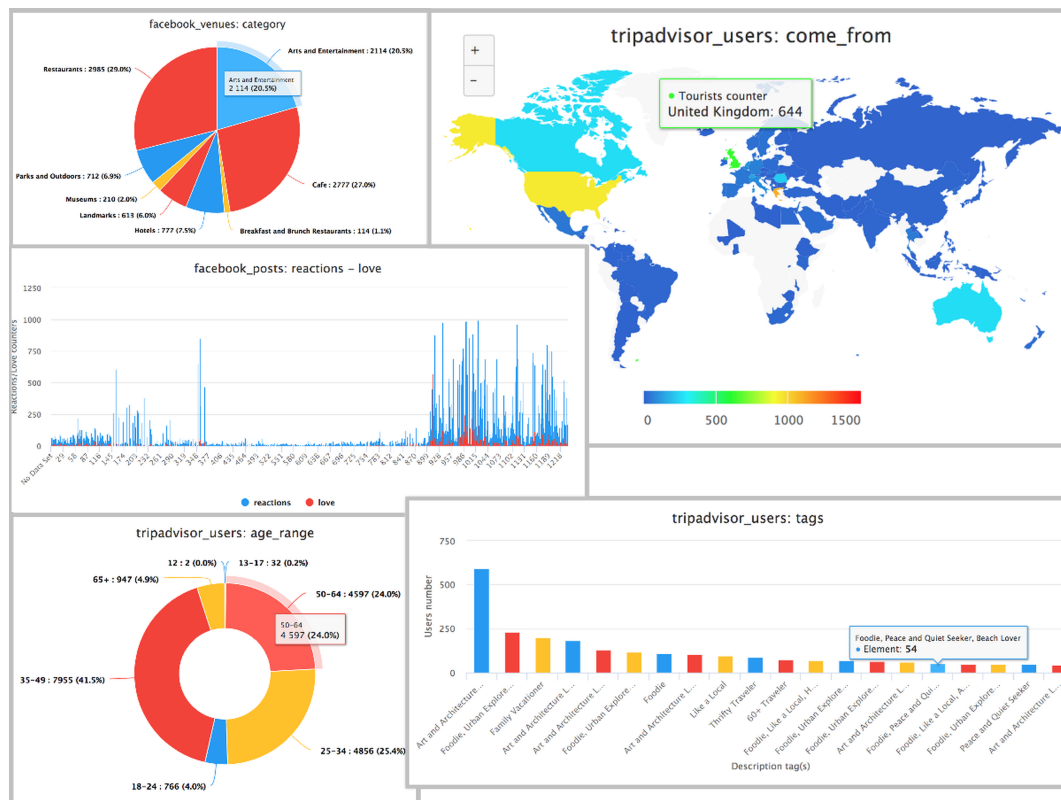


Figure 7. Graphs produced by the TripMentor data.

More specifically, the top left graph in Figure 7 presents a pie chart with the different PoIs in the Attica region grouped by venue category, while the graph in the center of Figure 7 provides a bar chart with the total number of Facebook reactions vs. the number of love reactions on posts created for the different venues in the Attica region. Both aforementioned graphs visualize the data harvested in the context of TripMentor from Facebook spidering. The rest of the graphs presented in Figure 7 visualize the data extracted from the TripAdvisor social network; the bottom left graph shows the age range of the users that checked in or commented about different cultural venues in the Attica region in a donut chart; the bottom right diagram presents a bar chart with the most popular (self-assigned) user tags of users who have visited at least one PoI in the Attica region in Greece; and the geo graph in the top-right corner of the figure presents the geographical location of origin (at country granularity) of the visitors of the PoIs within the respective data lake.

Notice that the visualizations in Figure 7 present only a meaningful sample (in the context of the examined case study) of the available data processing and visualization capabilities of Hydria; these capabilities extend to include numerous additional graph types (as enumerated in Section 3.3 above), export to different formats such as raw/CSV/TSV, and many more.

5. Indicative Application Scenarios for Hydria

In this section, we provide four distinct application scenarios that demonstrate the versatility of Hydria and highlight its usefulness in diverse cultural heritage setups.

5.1. Hydria for Curators

In this application scenario let us consider Auguste, a curator for a small regional museum who wants to collect data that include mentions of the museum he is responsible for from various online social networks, and perform simple sentiment analysis on these mentions to understand what the visitors like or dislike about the museum. Since the museum resources are not sufficient for maintaining an IT department or acquiring the necessary computational resources, Auguste resorts to manually skimming through scattered visitor reviews on various social media (e.g., Facebook, TripAdvisor, Google reviews) and regularly searching Twitter to get an overall feeling of the visitor opinions about his museum. Of course, as this is a constantly evolving process, he has to personally search for new useful reviews and take into account recent unexpected events (e.g., a power outage that disappointed many first time visitors) that may drive review scores adrift.

Clearly, Auguste would greatly benefit from an online, free, and powerful information system that would allow him to create and launch automated social media harvesting tools able to monitor popular social platforms and store content (e.g., posts, tweets, reviews) in a transparent way at an online, easily accessible data store. Such a system would allow him to access the stored data, perform the required opinion analysis, and easily visualize the results into different types of graphs to be used for press and media releases.

5.2. Hydria for Researchers

On another scenario, Nikki, member of a humanities research group, wants to perform an analysis of how particular events in European history are perceived by citizens of different European countries and record reflections of individuals on those events. This survey is part of a European effort and involves a large consortium of researchers of different disciplines; as Nikki's group is coordinating the survey, she is also responsible to set up the platform for the collection and processing of survey data. Nikki cannot use any of the popular survey creation tools as they do not support access control to the data and user management (each consortium partner should only have access to the data they collected, with the exception of the coordinator who should be able to access all data). She finally decides to contact an IT company and explain the needs and specificities of the survey. Subsequently, her group would need to buy and maintain a costly IT infrastructure onsite to host the developed solution and, possibly, hire an IT professional/company to keep both the system and the infrastructure up to date.

Clearly, Nikki and her group would benefit from accessing an online service that would allow her to create and deploy such a platform in a fast, free, and effortless way; this system would be a valuable tool *beyond anything currently supported*. After the platform creation, Nikki will be able to create and manage the users of the deployed platform and define access control policies. These users will then be able to login and either input questionnaire data, or directly provide the survey subjects with appropriate credentials that would identify the owner of the input. When the data collection phase is completed, the survey coordinator (i.e., Nikki's group) may use the available searching and filtering techniques to issue appropriate queries, export data of interest for analysis (e.g., to tools like SPSS), and visualize the findings of the analysis in an easy and intuitive way. Since Nikki's group is coordinating the survey, it has access to all inserted data, while other groups' access is restricted to the data policy enforced.

5.3. Hydria for Data Scientists

Sophia is a computer scientist working in an organization that provides support in the construction and maintenance of a collaborative home-brewed database for cultural heritage applications. In this

context, Sophia is responsible for the curation of the database and the enrichment/integration with existing Web resources like DBpedia [85] and various online thesauri (e.g., the Getty Art & Architecture Thesaurus [86]) using both manual/crowdsourced and automatic techniques [87,88]. As this knowledge base is continuously evolving, monitoring its quality over time becomes an essential task. Having access to an appropriate information system that is able to support publish/subscribe functionality for alerting of possible events or data inputs of interest would allow Sophia and other data scientists to subscribe (with appropriate textual and attribute constraints) and get notified about (i) spurious and/or unusual input in the collaborative database, (ii) the creation/evolution of different schemas used to represent various data facets, and (iii) the trending of specific terms or attributes in the database. Such functionality would be an invaluable tool that would simplify database moderation, as Sophia would, for example, be notified about (i) an unusual database update action that mistakenly records the “Benaki Museum” in Attica, Wyoming, instead of Attica, Greece, (ii) a new published dataset on European History of Art containing appropriate metadata for exhibits located in museums that are already using the system for storing collection specific information, or (iii) a new research topic.

5.4. *Hydria for End Users*

Finally, Matteo is an under-graduate student from an Information Studies department writing his thesis in contextual reasoning for cultural heritage applications. Matteo is mainly interested in retrieving scientific publications on his topic of interest, following the work of prominent researchers in the area, and studying the evolution of the field over time. Due to the particularities of his research field (i.e., focused but interdisciplinary topic, heavy mathematical background), he regularly resorts to online resources—like the DBLP digital library [68] and WikiCFP [70] portal—to search for new relevant areas, to study and map the evolution of the field in terms of scientific papers and related venues (like conferences and workshops). To do so, it is required to periodically download relevant datasets from these sites (e.g., the raw DBLP data of all indexed papers), filter them to maintain only relevant information and store them for further processing (e.g., perform timeline analysis on the new available data). Even though searching for interesting/related works this week turned up nothing, a search next week may return new information or even new datasets. Clearly, an information system that is able to (i) easily integrate several online digital sources, (ii) incorporate new datasets, (iii) analyze and visualize the analysis results, and (iv) capture his long-term information need (using publish/subscribe functionality) would be a valuable tool that would allow Matteo to save both time and effort.

6. Conclusions and Outlook

We have presented Hydria, the first online, free, zero-administration platform that offers both fundamental and advanced user and data/knowledge management functionality for big cultural data and targets users with little or no IT background. Hydria enables the direct incorporation of heterogeneous data that has been recorded in dispersed formats, while specialized processing engines ingest data without compromising the data structure, making it available for tasks such as visualization, mining, analytics and reporting. Thus, a system in the form of a data lake meant for acquiring, storing, organizing, analyzing and sharing multi-faceted cultural heritage data constitutes a valuable asset to several different cultural heritage applications such as museum curation, user study management, bibliographical analysis, dataset management, and data integration.

In this work we discussed the architectural solutions behind the proposed system, outlined the individual module technologies and provided details on the module orchestration. We also described several novel services that include automated data harvesting from the web and social media, integrated user input collection via standard and customizable data types, easy to perform data analysis and visualizations, publish/subscribe functionality to facilitate sharing of different facets and data shards, and access control mechanisms. Finally, we advocated the appropriateness of our approach for the cultural heritage domain and showcased different scenarios that highlight Hydria’s

usefulness for cultural data management. We continuously develop new functionality to support more import/export formats and more sophisticated data types, perform user studies to improve usability and document additional user needs, incorporate more data analysis tools and simplify the data analysis procedure, and incorporate versatile data streams from sensors and IoT devices.

Currently, Hydria is running on a commodity server and is dealing mainly with the TripMentor project needs; however, the long-term plan is to release it as a free service to any interested parties. This entails tackling several important issues that include scaling, platform viability and impact measurements. Regarding scaling, we plan to reshape Hydria before releasing a free public version of the system; the intended reshaping includes modifying some system components to be cloud-native, so as to provide resource elasticity and exploit the benefits emanating from cluster computing infrastructures. Platform viability and maintenance after the end of the project is a typical issue in applied research; we plan to actively pursue new funding that will allow us to continue the development and extension of Hydria. Moreover, the open source ecosystem of tools used to build Hydria allows us to release it as an open source project to the development community to further aid project maintainability. Finally, appropriate impact measurements are an important direction that will drive and affect both the large-scale deployment and the viability of the platform. Hydria, as also usually happens with many research prototypes, is not directly involved in generating revenue, so impact measurements could involve KPIs such as user base, data quality, application versatility, release efficiency, and system reliability.

The work presented in this paper has several implications for both practitioners and researchers. At practical level, it introduces a tool that empowers its users to access, interrelate, analyze, share and visualize multi-faceted data harvested from structured or semi-structured sources, through an intuitive graphical interface, without the requirement of any IT skills. While the implemented system targets the cultural information domain, it can be straightforwardly adapted to any domain where data sourced from social networks and the linked open data (LOD) cloud need to be harvested, managed, analyzed and shared, such as the marketing, political and social analysis domains. The Hydria data lake may also contribute the data pond contents to the LOD cloud, reciprocating from social networks and the LOD cloud through the provision of unified and integrated datasets. The sharing mechanisms of the Hydria system can be leveraged to provide persistent identifiers and automatically register data ponds (or parts of them) that are characterized as “public” to searchable directories, adhering thus to the FAIR data principles [89].

Regarding the research dimension, the architectural paradigm of Hydria, which is a key factor to its success, can be adopted in other classes of systems that offer services to non-IT experts, such as scientific data analysis systems or business data analysis systems. The proliferation of systems based on the architecture of Hydria will accelerate the development cycle of analysis and visualization algorithms that are suitable for non-IT experts, since the extension of the prospective user base will facilitate gathering of relevant requirements and allow the collection of richer testing and evaluation feedback.

Another research direction for the Hydria platform is to extend cooperation between users beyond data sharing, to include expertise finding among the users of Hydria for advice seeking or joint execution of tasks requiring diverse areas of expertise (e.g., multidisciplinary tasks). To this end, algorithms for expert identification can be developed for the Hydria platform, or appropriate existing algorithms can be identified and tuned (e.g., [90–92]). Expert searches may also extend outside the scope an Hydria installation (or a federated Hydria installations network), through the interfacing of the Hydria platform to expert hiring and crowdworking platforms [93] as well as the adoption and customization of algorithms for the synchronization of these tasks [94]. In all cases, all types of modules developed for the Hydria system (e.g., analysis or visualization algorithms, or components supporting expert identification and cooperation between users), as well as knowledge about best practices, can be stored in shared and searchable repositories, providing a dynamic, evolving and self-sustained ecosystem for the Hydria platform.

Author Contributions: Conceptualization, K.D., P.R., C.T. and C.V.; Methodology, K.D., P.R., C.T. and C.V.; Software, K.D., P.R., C.T., N.P. and C.V.; Visualization, N.P., K.D.; Validation, K.D., P.R., C.T. and C.V.; Writing—Original draft, K.D., P.R., C.T., C.V.; Writing—Review & editing, K.D., P.R., C.T., N.P. and C.V. All authors have read and agree to the published version of the manuscript.

Funding: This research has been co-financed by European Union and Greek national funds through the Operational Programme “Competitiveness, Entrepreneurship and Innovation”, under the call RESEARCH—CREATE—INNOVATE (project code: T1EDK-03874).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyzes, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Kenteris, M.; Vafopoulos, M.N.; Gavalas, D. Cultural Informatics in Web Science: A Case of Exploiting Local Cultural Content. In Proceedings of the 12th Pan-Hellenic Conference on Informatics, Samos Island, Greece, 28–30 August 2008.
2. Salvatore, C.L. (Ed.) *Cultural Heritage Care and Management: Theory and Practice*; Rowman & Littlefield: London, UK, 2018.
3. Antoniou, A.; Katifori, A.; Roussou, M.; Vayanou, M.; Karvounis, M.; Kyriakidi, M.; Pujol-Tost, L. Capturing the Visitor Profile for a Personalized Mobile Museum Experience: An Indirect Approach. In Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalisation (UMAP 2016), Halifax, NS, Canada, 13–17 July 2016.
4. Deladiennee, L.; Naudet, Y. A graph-based semantic recommender system for a reflective and personalised museum visit. In Proceedings of the 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Bratislava, Slovakia, 9–10 July 2017.
5. Bourlakos, I.; Wallace, M.; Antoniou, A.; Vassilakis, C.; Lepouras, G.; Karapanagiotou, A.V. Formalization and Visualization of the Narrative for Museum Guides. In Proceedings of the Semantic Keyword-Based Search on Structured Data Sources Conference (IKC), Gdansk, Poland, 11–12 September 2018; Springer: Cham, Switzerland, 2018; pp. 3–13.
6. Vassilakis, C.; Pouloupoulos, V.; Antoniou, A.; Wallace, M.; Lepouras, G.; Nores, M.L. exhiSTORY: Smart exhibits that tell their own stories. In *Future Generation Computer Systems*; Elsevier: Amsterdam, The Netherlands, 2018; Volume 81, pp. 542–556.
7. Meghini, C.; Bartalesi, V.; Metilli, D.; Benedetti, F. A Software Architecture for Narratives. In Proceedings of the Italian Research Conference on Digital Libraries, Udine, Italy, 25–26 January 2018.
8. Bampatzia, S.; Bravo-Quezada, O.G.; Antoniou, A.; Nores, M.L.; Wallace, M.; Lepouras, G.; Vassilakis, C. The Use of Semantics in the CrossCult H2020 Project. In Proceedings of the Semantic Keyword-Based Search on Structured Data Sources Conference (IKC), Cluj-Napoca, Romania, 8–9 September 2016; Springer: Cham, Switzerland, 2016; Volume 10151, pp. 190–195.
9. Kyvernitou, I.; Bikakis, A. An Ontology for Gendered Content Representation of Cultural Heritage Artefacts. *Digit. Humanit. Q.* **2017**, *11*. Available online: <https://discovery.ucl.ac.uk/id/eprint/10041951/> (accessed on 23 April 2020).
10. Vlachidis, A.; Bikakis, A.; Kyriaki-Manessi, D.; Triantafyllou, I.; Antoniou, A. The CrossCult Knowledge Base: A Co-inhabitant of Cultural Heritage Ontology and Vocabulary Classification. In Proceedings of the European Conference on Advances in Databases and Information Systems, Nicosia, Cyprus, 24–27 September 2017.
11. Bartalesi, V.; Meghini, C. Using an ontology for representing the knowledge on literary texts: The Dante Alighieri case study. *Semant. Web* **2017**, *8*, 385–394. [[CrossRef](#)]
12. Antoniou, A.; Lepouras, G. Modeling visitors’ profiles: A study to investigate adaptation aspects for museum learning technologies. *J. Comput. Cult. Herit. (JOCCH)* **2010**, *3*, 1–19. [[CrossRef](#)]
13. Martin, J.; Trummer, C. Personalized Multimedia Information System for Museums and Exhibitions. In *Lecture Notes in Computer Science, Proceedings of the 1st International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN), Madonna di Campiglio, Italy, 30 November-2 December 2005*; Springer: Cham, Switzerland 2005; Volume 3814, pp. 332–335.

14. Rowe, J.P.; Lobene, E.V.; Mott, B.W.; Lester, J.C. Serious Games Go Informal: A Museum-Centric Perspective on Intelligent Game-Based Learning. In *Lecture Notes in Computer Science, Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS), Honolulu, HI, USA, 5–9 June 2014*; Springer: Cham, Switzerland, 2014; Volume 8474, pp. 410–415.
15. Tavcar, A.; Antonya, C.; Butila, E. Recommender System for Virtual Assistant Supported Museum Tours. *Inform. (Slovenia)* **2016**, *40*, 279–284.
16. Vassilakis, C.; Antoniou, A.; Lepouras, G.; Pouloupoulos, V.; Wallace, M.; Bampatzia, S.; Bourlacos, I. Stimulation of reflection and discussion in museum visits through the use of social media. *Soc. Netw. Anal. Min.* **2017**, *7*, 40. [CrossRef]
17. Bampatzia, S.; Antoniou, A.; Lepouras, G.; Vassilakis, C.; Wallace, M. Using social media to stimulate history reflection in cultural heritage. In *Proceedings of the 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Thessaloniki, Greece, 20–21 October 2016*; pp. 89–92.
18. Fontanella, F.; Molinara, M.; Gallozzi, A.; Cigola, M.; Senatore, L.J.; Florio, R.; Clini, P.; D’Amico, F.C. HeritageGO (HeGO): A Social Media Based Project for Cultural Heritage Valorization. In *Proceedings of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP), Larnaca, Cyprus, 4–17 July 2019*; pp. 377–382.
19. Nguyen, T.T.; Camacho, D.; Jung, J.E. Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services. *Pers. Ubiquitous Comput.* **2017**, *21*, 267–279. [CrossRef]
20. Monti, L.; Delnevo, G.; Mirri, S.; Salomoni, P.; Callegati, F. Digital Invasions Within Cultural Heritage: Social Media and Crowdsourcing. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Proceedings of the 3rd International Conference on Smart Objects and Technologies for Social Good (GOODTECHS), Pisa, Italy, 29–30 November 2017*; Springer: Cham, Switzerland, 2017; Volume 233, pp. 102–111.
21. Nguyen, T.T.; Hwang, D.; Jung, J.J. Using Geotagged Resources on Social Media for Cultural Tourism: A Case Study on Cultural Heritage Tourism. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Proceedings of the 7th International Conference on Big Data Technologies and Applications (BDTA), Seoul, Korea, 17–18 November 2016*; Springer: Cham, Switzerland, 2016; Volume 194, pp. 64–72.
22. Liew, C.L. Participatory Cultural Heritage: A Tale of Two Institutions’ Use of Social Media. *D-Lib Mag.* **2014**, *20*. Available online: <http://www.dlib.org/dlib/march14/liew/03liew.html> (accessed on 23 April 2020). [CrossRef]
23. Jensen, B. Instagram as cultural heritage: User participation, historical documentation, and curating in Museums and archives through social media. In *Proceedings of the Digital Heritage International Congress, Marseille, France, 28 October–1 November 2013*; pp. 311–314.
24. 7th International Euro-Mediterranean Conference (EuroMed), LNCS, Nicosia, Cyprus, 29 October–3 November 2018; Springer: Cham, Switzerland, 2018. Available online: <https://wbc-rti.info/object/event/17918> (accessed on 23 April 2020).
25. 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH), ACL, Minneapolis, MN, USA, June 2019. Available online: <https://www.aclweb.org/anthology/volumes/W19-25/> (accessed on 23 April 2020).
26. 10th International Workshop on Human-Computer Interaction, Tourism and Cultural Heritage (HCITOCH), LNCS, Florence, Italy, 5–7 September 2019; Springer: Cham, Switzerland, 2019.
27. *Communications in Computer and Information Science, 1st International Conference on VR Technologies in Cultural Heritage (VRTCH), Brasov, Romania, 29–30 May 2018*; Springer: Cham, Switzerland, 2018; Volume 904. Available online: <http://library.oapen.org/handle/20.500.12657/23304> (accessed on 23 April 2020).
28. Bai, D.; Messenger, D.W.; Howell, D. A pigment analysis tool for hyperspectral images of cultural heritage artifacts. In *Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIII, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Anaheim, CA, USA, 9–13 April 2017*; Volume 10198.
29. Hoonjong.; Stoykova, E.; Berberova, N.; Park, J.; Nazarova, D.; Park, J.S.; Kim, Y.; Hong, S.; Ivanov, B.; Malinowski, N. Three-dimensional imaging of cultural heritage artifacts with holographic printers. In *Proceedings of the 19th International Conference and School on Quantum Electronics: Laser Physics and Applications (ICSQE), Sozopol, Bulgaria, 26–30 September 2016*; Volume 10226.

30. Themistocleous, K. Debate and Considerations on Using Videos for Cultural Heritage from Social Media for 3D Modelling. In Proceedings of the 6th International Conference on Progress in Cultural Heritage: Documentation, Preservation, and Protection (EuroMed), Nicosia, Cyprus, 31 October–5 November 2016; Volume 10058, pp. 513–520.
31. Torres, J.C.; López, L.; Romo, C.; Arroyo, G.; Cano, P.; Lamolda, F.; del Mar Villafranca, M. Using a Cultural Heritage Information System for the documentation of the restoration process. In Proceedings of the Digital Heritage International Congress, Marseille, France, 28 October–1 November 2013; pp. 249–256.
32. Nurminen, M.; Heimburger, A. Representation and Retrieval of Uncertain Temporal Information in Museum Databases. In Proceedings of the 21st European—Japanese Conference on Information Modelling and Knowledge Bases (EJC), Tallinn, Estonia, 6–10 June 2011; Volume 1. Available online: <http://ebooks.iospress.nl/publication/6781> (accessed on 23 April 2020).
33. Chias, P.; Abad, T. Visualising Ancient Maps as Cultural Heritage: A Relational Database of the Spanish Ancient Cartography. In Proceedings of the 12th International Conference on Information Visualisation, London, UK, 9–11 July 2008; Volume 1.
34. Meyer, E.; Grussenmeyer, P.; Perrin, J.P.; Durand, A.; Drap, P. A web information system for the management and the dissemination of Cultural Heritage data. *J. Cult. Herit.* **2007**, *8*, 396–411. [[CrossRef](#)]
35. Jancsó, A.L.; Jonlet, B.; Hoffsummer, P.; Delye, E.; Billen, R. An Analytical Framework for Classifying Software Tools and Systems Dealing with Cultural Heritage Spatio-Temporal Information. In *Lecture Notes in Geoinformation and Cartography, Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT); L'Aquila, Italy, 4–8 September 2017*; Springer: Cham, Switzerland, 2017; pp. 325–337.
36. Colace, F.; Santo, M.D.; Greco, L.; Chianese, A.; Moscato, V.; Picariello, A. CHIS: Cultural Heritage Information System. *IJKSR* **2013**, *4*, 18–26. [[CrossRef](#)]
37. Chinnov, A.; Kerschke, P.; Meske, C.; Stieglitz, S.; Trautmann, H. *An Overview of Topic Discovery in Twitter Communication through Social Media Analytics*; AMCIS: Morristown, NJ, USA, 2015.
38. Korzun, D.G. Designing Smart Space Based Information Systems: The Case Study of Services for IoT-Enabled Collaborative Work and Cultural Heritage Environments. In *Frontiers in Artificial Intelligence and Applications, Proceedings of the 12th International Baltic Conference on Databases and Information Systems (DB&IS), Riga, Latvia, 4–6 July 2016*; Arnicans, G., Arnican, V., Borzovs, J., Niedrite, L., Eds.; IOS Press: Clifton, VA, USA, 2016; Volume 291, pp. 183–196.
39. Pouloupoulos, V.; Vassilakis, C.; Antoniou, A.; Wallace, M.; Lepouras, G.; Nores, M.L. ExhiSTORY: IoT in the service of Cultural Heritage. In Proceedings of the IEEE Global Information Infrastructure and Networking Symposium (GIIS), Thessaloniki, Greece, 23–25 October 2018; pp. 1–4.
40. Su, X.; Sperli, G.; Moscato, V.; Picariello, A.; Esposito, C.; Choi, C. An Edge Intelligence Empowered Recommender System Enabling Cultural Heritage Applications. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4266–4275. [[CrossRef](#)]
41. Pandolfo, L.; Pulina, L.; Grosso, E. A User Model Ontology for Adaptive Systems in Cultural Tourism Domain. In *Frontiers in Artificial Intelligence and Applications, Proceedings of the 1st International Conference on Applications of Intelligent Systems (APPIS), Las Palmas de Gran Canaria, Spain, 10–12 January 2018*; Petkov, N., Strisciuglio, N., Travieso-González, C.M., Eds.; IOS Press: Clifton, VA, USA, 2018; Volume 310, pp. 212–219.
42. Díaz-Corona, D.; Lacasta, J.; Latre, M.Á.; Zarazaga-Soria, F.J.; Noguerras-Iso, J. Profiling of knowledge organisation systems for the annotation of Linked Data cultural resources. *Inf. Syst.* **2019**, *84*, 17–28. [[CrossRef](#)]
43. Larosiliere, G.D.; Carter, L.D.; Meske, C. How does the world connect? Exploring the global diffusion of social network sites. *J. Assoc. Inform. Sci. Technol. (JASIST)* **2017**, *68*, 1875–1885. [[CrossRef](#)]
44. Miloslavskaya, N.; Tolstoy, A. Big Data, Fast Data and Data Lake Concepts. *Procedia Comput. Sci.* **2016**, *88*, 300–305. [[CrossRef](#)]
45. TripMentor Project. Available online: <https://www.researchgate.net/project/TripMentor> (accessed on 1 April 2020).
46. Chianese, A.; Marulli, F.; Piccialli, F. Cultural Heritage and Social Pulse: A Semantic Approach for CH Sensitivity Discovery in Social Media Data. In Proceedings of the 10th International Conference on Semantic Computing (ICSC), IEEE Computer Society, Laguna Hills, CA, USA, 4–6 February 2016; pp. 459–464.

47. Moscato, V.; Picariello, A.; Subrahmanian, V.S. Multimedia Social Networks for Cultural Heritage Applications: The GIVAS Project. In *Data Management in Pervasive Systems; Data-Centric Systems and Applications*; Springer: Cham, Switzerland, 2015; pp. 169–182.
48. Colace, F.; Santo, M.D.; Moscato, V.; Picariello, A.; Schreiber, F.A.; Tanca, L. PATCH: A Portable Context-Aware ATLAS for Browsing Cultural Heritage. In *Data Management in Pervasive Systems; Data-Centric Systems and Applications*; Springer: Cham, Switzerland, 2015; pp. 345–361.
49. Vodopivec, B.; Eppich, R.; Zarnic, R. Cultural Heritage Information Systems State of the Art and Perspectives. In *Lecture Notes in Computer Science, Proceedings of the 5th International Conference on Progress in Cultural Heritage: Documentation, Preservation, and Protection (EuroMed), Limassol, Cyprus, 3–8 November 2014*; Springer: Cham, Switzerland, 2014; Volume 8740, pp. 146–155.
50. Alkhafaji, A.S.A.; Fallahkhair, S. Smart Ambient: Development of Mobile Location Based System to Support Informal Learning in the Cultural Heritage Domain. In *Proceedings of the 14th International Conference on Advanced Learning Technologies (ICALT), IEEE Computer Society, Athens, Greece, 7–10 July 2014*; pp. 774–776.
51. Cassatella, C.; Volpiano, M.; Seardo, B.M. Interpreting historic and cultural landscapes: Potentials and risks in Geographical Information Systems building for knowledge and management. In *Proceedings of the Digital Heritage International Congress, IEEE, Marseille, France, 28 October–1 November 2013*; pp. 107–110.
52. Torres, J.C.; López, L.; Romo, C.; Soler, F. An Information System to Analyze Cultural Heritage Information. In *Lecture Notes in Computer Science, Proceedings of the 4th International Conference on Progress in Cultural Heritage: Documentation, Preservation, and Protection (EuroMed), Limassol, Cyprus, 29 October–3 November 2012*; Springer: Cham, Switzerland, 2012; Volume 7616, pp. 809–816.
53. Ploszajski, G. Technical Metadata and Standards for Digitisation of Cultural Heritage in Poland. In *New Trends in Multimedia and Network Information Systems; Frontiers in Artificial Intelligence and Applications*; IOS Press: Amsterdam, The Netherlands, 2008; Volume 181, pp. 155–170.
54. Smirnov, A.V.; Kashevnik, A.M.; Ponomarev, A. Context-based infomobility system for cultural heritage recommendation: Tourist Assistant—TAIS. *Pers. Ubiquitous Comput.* **2017**, *21*, 297–311. [[CrossRef](#)]
55. Narumi, T.; Hayashi, O.; Kasada, K.; Yamazaki, M.; Tanikawa, T.; Hirose, M. Digital Diorama: AR Exhibition System to Convey Background Information for Museums. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Virtual and Mixed Reality—New Trends, Orlando, FL, USA, 9–4 July 2011*; Shumaker, R., Ed.; Springer: Cham, Switzerland, 2011; Volume 6773, pp. 76–86.
56. Gentile, A.; Andolina, S.; Massara, A.; Pirrone, D.; Russo, G.; Santangelo, A.; Trumello, E.; Sorce, S. A Multichannel Information System to Build and Deliver Rich User-Experiences in Exhibits and Museums. In *Proceedings of the International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA), IEEE Computer Society, Barcelona, Spain, 26–28 October 2011*; pp. 57–64.
57. Chen, S.; Pan, Z.; Zhang, M. A Virtual Informal Learning System for Cultural Heritage. *Trans. Edutainment* **2012**, *7*, 180–187.
58. Wu, S. Systems integration of heterogeneous cultural heritage information systems in museums: A case study of the National Palace Museum. *Int. J. Digit. Libr.* **2016**, *17*, 287–304. [[CrossRef](#)]
59. Chen, C.; Chang, B.R.; Huang, P. Multimedia augmented reality information system for museum guidance. *Pers. Ubiquitous Comput.* **2014**, *18*, 315–322. [[CrossRef](#)]
60. Chanhom, W.; Anutariya, C. TOMS: A Linked Open Data System for Collaboration and Distribution of Cultural Heritage Artifact Collections of National Museums in Thailand. *New Gener. Comput.* **2019**, *37*, 479–498. [[CrossRef](#)]
61. Naudet, Y.; Antoniou, A.; Lykourantzou, I.; Tobias, E.; Rompa, J.; Lepouras, G. Museum personalization based on gaming and cognitive styles: The BLUE experiment. *Int. J. Virtual Communities Soc. Netw. (IJVCSN)* **2015**, *7*, 1–30. [[CrossRef](#)]
62. Bampatzia, S.; Bourlakos, I.; Antoniou, A.; Vassilakis, C.; Lepouras, G.; Wallace, M. Serious games: Valuable tools for cultural heritage. In *Proceedings of the International Conference on Games and Learning Alliance, Utrecht, The Netherlands, 5–7 December 2016*; Springer: Cham, Switzerland, 2016; pp. 331–341.
63. Dorter, G.; Davis, L. Bringing geographic information systems (GIS) into the museum world. In *Proceedings of the Digital Heritage International Congress, IEEE, Marseille, France, 28 October–1 November 2013*.
64. Soler, F.; Torres, J.C.; León, A.J.; Luzón, M.V. Design of cultural heritage information systems based on information layers. *JOCCH* **2013**, *6*, 1–17. [[CrossRef](#)]

65. Soler, F.; Torres, J.C.; León, A.J.; Luzón, M.V. Design of an Information System for Cultural Heritage. In Proceedings of the Spanish Computer Graphics Conference (CEIG), Eurographics Association, Jaén, Spain, 12–14 September 2012; pp. 113–122.
66. Wikipedia The Free Encyclopedia. Available online: <https://www.wikipedia.org/> (accessed on 1 April 2020).
67. Europeana. Available online: <https://www.europeana.eu/portal/en> (accessed on 1 April 2020).
68. DBLP: Computer Science Bibliography. Available online: <https://dblp.org/> (accessed on 1 April 2020).
69. Odysseus Ministry of Culture and Sports. Available online: http://odysseus.culture.gr/index_en.html (accessed on 1 April 2020).
70. WikiCFP A wiki for Calls For Papers. Available online: <http://www.wikicfp.com/cfp/> (accessed on 1 April 2020).
71. Meske, C.; Junglas, I.A.; Schneider, J.; Jaakonmaki, R. How Social is Your Social Network? Toward A Measurement Model. In Proceedings of the 40th International Conference on Information Systems (ICIS), Munich, Germany, 15–18 December 2019.
72. Stieglitz, S.; Meske, C.; Ross, B.; Mirbabaie, M. Going Back in Time to Predict the Future—The Complex Role of the Data Collection Period in Social Media Analytics. *Inf. Syst. Fronti.* **2018**. [CrossRef]
73. von der Putten, A.M.R.; Hastall, M.; Köcher, S.; Meske, C.; Heinrich, T.; Labrenz, F.; Ocklenburg, S. “Likes” as social rewards: Their role in online social comparison and decisions to like other People’s selfies. *Comput. Hum. Behav.* **2019**, *92*, 76–86. [CrossRef]
74. Myers, D.; McGuffee, J.W. Choosing scrapy. *J. Comput. Sci. Coll.* **2015**, *31*, 83–89.
75. Scrapy at a Glance. Available online: <https://docs.scrapy.org/en/latest/intro/overview.html> (accessed on 10 March 2020).
76. Chaulagain, R.S.; Pandey, S.; Basnet, S.R.; Shakya, S. Cloud based web scraping for big data applications. In Proceedings of the IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, 3–5 November 2017; pp. 138–143.
77. Santos, A.; Pham, K. GitHub—VIDA-NYU/ache. Available online: <https://github.com/VIDA-NYU/ache> (accessed on 1 April 2020).
78. Barbosa, L.; Freire, J. An adaptive crawler for locating hidden-web entry points. In Proceedings of the 16th International Conference on World Wide Web (WWW), Banff, AL, Canada, 8–12 May, 2007; pp. 441–450.
79. Vieira, K.; da Silva, L.B.A.S.; Freire, J.; Moura, E. Finding seeds to bootstrap focused crawlers. *World Wide Web (WWW)* **2016**, *19*, 449–474. Available online: <https://link.springer.com/article/10.1007/s11280-015-0331-7> (accessed on 23 April 2020). [CrossRef]
80. Bonzanini, M. *Mastering Social Media Mining with Python*; Packt Publishing Ltd.: Birmingham, UK, 2016.
81. Stauffer, M. *Laravel: Up & Running: A Framework for Building Modern PHP Apps*; O’Reilly Media: Sebastopol, CA, USA, 2019.
82. TripAdvisor: Read Reviews, Compare Prices & Book. Available online: <https://www.tripadvisor.com/> (accessed on 1 April 2020).
83. Facebook. Available online: <https://www.facebook.com/> (accessed on 1 April 2020).
84. Twitter. Available online: <https://twitter.com/> (accessed on 1 April 2020).
85. DBpedia Homepage. Available online: <https://wiki.dbpedia.org/> (accessed on 1 April 2020).
86. Art & Architecture Thesaurus Online. Available online: <https://www.getty.edu/research/tools/vocabularies/aat/index.html> (accessed on 1 April 2020).
87. Marketakis, Y.; Minadakis, N.; Kondylakis, H.; Konsolaki, K.; Samaritakis, G.; Theodoridou, M.; Flouris, G.; Doerr, M. X3ML mapping framework for information integration in cultural heritage and beyond. *IJDL* **2017**, *18*, 301–319. [CrossRef]
88. Stavropoulos, T.G.; Kontopoulos, E.; Meroño-Peñuela, A.; Tachos, S.; Andreadis, S.; Kompatsiaris, Y. Cross-domain Semantic Drift Measurement in Ontologies Using the SemaDrift Tool and Metrics. In Proceedings of the MEPDaW & LDQ @ ESWC, Bologna, Italy, 29 May 2017.
89. Initiative, G.F. FAIR Principles. 2019. Available online: <https://www.go-fair.org/fair-principles/> (accessed on 6 April 2020).
90. Bozzon, A.; Brambilla, M.; Ceri, S.; Silvestri, M.; Vesce, G. Choosing the Right Crowd: Expert Finding in Social Networks. In Proceedings of the 16th International Conference on Extending Database Technology (EDBT), Genoa, Italy, 18–22 March 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 637–648. [CrossRef]

91. Lin, S.; Hong, W.; Wang, D.; Li, T. A survey on expert finding techniques. *J. Intell. Inf. Syst.* **2017**, *49*, 255–279. [[CrossRef](#)]
92. Nikzad-Khasmakhi, N.; Balafar, M.; Reza Feizi-Derakhshi, M. The state-of-the-art in expert recommendation systems. *Eng. Appl. Artif. Intell.* **2019**, *82*, 126–147. [[CrossRef](#)]
93. Lykourantzou, I.; Khan, V.J.; Papangelis, K.; Markopoulos, P. Macrotask Crowdsourcing: An Integrated Definition. In *Human—Computer Interaction Series*; Lykourantzou, I., Khan, V.J., Papangelis, K., Markopoulos, P., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 1–13. [[CrossRef](#)]
94. Schmitz, H.; Lykourantzou, I. Online Sequencing of Non-Decomposable Macrotasks in Expert Crowdsourcing. *Trans. Soc. Comput.* **2018**, *1*, 1–33. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).