DOI 10.15622/sp.18.2.444-470

D. Paneva-Marinova, J. Stoikov, L. Pavlova, D. Luchev

SYSTEM ARCHITECTURE AND INTELLIGENT DATA CURATION OF VIRTUAL MUSEUM FOR ANCIENT HISTORY

Paneva-Marinova D., Stoikov J., Pavlova L., Luchev D. **System Architecture and Intelligent Data Curation of Virtual Museum for Ancient History**

Abstract. Preserving the cultural and historical heritage of various world nations, and their thorough presentation is a long-term commitment of scholars and researchers working in many areas. From centuries every generation is aimed at keeping record about its labor, so that it could be revised and studied by the next generations. New information and multimedia technologies have been developed during the past couple of years, which introduced new methods of preservation, maintenance and distribution of the huge amounts of collected material. This article aims to present the virtual museum, an advanced system managing diverse collections of digital objects that are organized in various ways by a complex specialized functionality. The management of digital content requires a well-designed architecture that embeds services for content presentation, management, and administration. All elements of the system architecture are interrelated, thus the accuracy of each element is of great importance. These systems suffer from the lack of tools for intelligent data curation with the capacity to validate data from different sources and to add value to data. This paper proposes a solution for intelligent data curation that can be implemented in a virtual museum in order to provide opportunity to observe the valuable historical specimens in a proper way. The solution is focused on the process of validation and verification to prevent the duplication of records for digital objects, in order to guarantee the integrity of data and more accurate retrieval of knowledge.

Keywords: virtual museum, system architecture, functionality, data integrity, knowledge retrieval, data validation, record de-duplication, cultural heritage.

1. Introduction. For a long time, cultural heritage has been maintained in museums, galleries, libraries and research laboratories, where not everyone was able to access this wealth. Digital technologies that have been developed during the past couple of years introduced new solutions of documentation, maintenance and distribution of the huge amounts of collected material. Among these new technologies are virtual museums, which have already proven their worth as a contemporary conceptual solution for access to and attractive presentation of cultural archives. Virtual museums contain diverse collections of digital objects (such as text, images, and media objects) that are organized in various ways and are managed by complex specialized services such as content structuring and grouping, attractive visualization, advanced search (semantic-based search, multilayer and personalized search, context-based search), resources and collection management, indexing, semantic description, knowledge retrieval, metadata management, personalization and content adaptability, content protection and preservation, tracking services, etc. Thus, the valuable cultural heritage wealth is accessible anytime and anywhere, in a friendly, multi-modal, efficient, and affective way.

However, these systems suffer from the lack of tools and services for intelligent data curation with the capacity to add value to data. In this paper, a solution for intelligent data curation in a virtual museum is proposed in order to provide opportunity to observe and analyze valuable ancient history specimens in a proper way and in their historic context, so that some yet undiscovered treasured of the human civilizations be manifested [31]. This solution more specifically is focused on the validation and prevention of duplication of newly added or existing records for digital objects.

Section 2 of this paper discusses some challenges raised during the design and the development of virtual museums. Section 3 and 4 present current concepts for virtual museum system architecture, tracking main functionality and services supporting users' needs. Section 5 includes a discussion on intelligent data curation issues. In section 6, a model of intelligent data curation service is described. The paper ends with some conclusions and further development plans.

2. Virtual museum design issues. The development of the technologies during the last years provides new functionalities and advanced services to contemporary virtual museum (VM) transforming their static complex structures to environment with a dynamic federation of functional units. This change resulted from the needs of the market, the emergence of new technologies, and especially from the request for stricter use of the existing resources and adapting VMs content and services to the needs of different user groups.

Some key research questions, raised during the design and the development of these systems, are:

– How to present the selected resources in a given context and to determine the conditions and use cases – cognitive or educational goals, analysis, creative use, etc.?

– How to help the user not just to view, but to also gain knowledge?

– How to provide knowledge in the most suitable way and form?

– How to adapt the offered information content for each individual user or group in order to achieve their goals and tasks? [29]

– How to choose the most suitable resources for a specific situation and the method of introduction to the domain, which is subject to research, etc.?

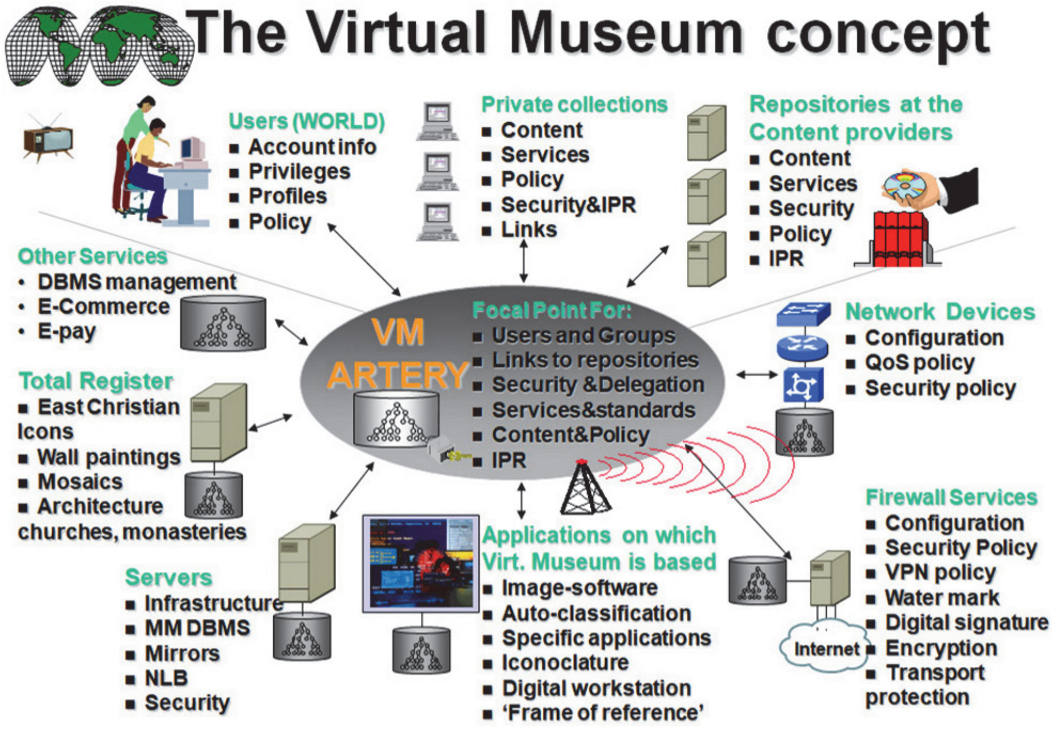
The difficulties in solving these research issues are related to the lack of common model and working solutions regarding the basic and the extended functionality, and synchronizing the solutions with the existing standards and regulations in the area; analysis, understanding and better interpretation of digital cultural content; context-dependent use of digital cultural resources; increase and generalisation of visitor experience, contextual techniques for personalising visitor experience, etc. [30].

A considerable interest in this area in recent years is demonstrated by Bulgarian scientists. The main efforts are concentrated in applied aspects, especially for increasing the presence of digital artefacts and collections of the Bulgarian cultural and historical heritage in the global information space. Besides, work is done towards developing ICT tools and systems for digital presentation and preservation of cultural heritage artefacts. There is also intensified interest in fundamental research (priority areas of Informatics, ICT and Cultural Heritage of the Strategy for the Development of Science in Bulgaria till 2020, Innovation Strategy for Intelligent Specialisation, Horizon 2020, etc.) in search of innovations especially in areas/subareas relevant to data processing, access control, intelligent supervision, security, semantics, etc. The current research activities include the study and applications of new methods and tools for the creation, integration and development of innovative systems, managing digital cultural assets [1, 3, 4, 8]. The focus is in researching and exploitation of new or emerging technologies for the development of innovative products, tools, applications and services for the creative digital content production, usage and management. The aim is to transform cultural heritage into digital units, which integration and reuse through research-led methods will have high commercial potential for cultural institutions, tourism, and creative and media industries.

The innovation principles for VM require visual rules, that characterize the different kinds of visual symbols; data rules, that specify the characteristics of the data model, the database schema, and the database instances; -mapping rules, that specify the link between data and visual elements; methodologies and tools for the support of cross-language retrieval from the nodes of the federated architecture must be developed [32].

They include device independent access to a world-wide digital repositories, including interconnected publications and reports; Service's availability 24h/day, 7d/week; Reduced cost and difficulty of content distribution ; Possibility to merge knowledge thematically in one personalized book by making a selection among the large offer of digital content (articles, chapters, extracts, etc.); Optimized management of perishable contents thanks to the maintenance service; Reusability, archiving and availability of the digital content in the "memory of knowledge"; Possibility to create and expand links between small and specialized museum communities; Socio-economic Innovation principles [28].

VM deal with issues of co-operation within the context of an information society comprised of independent organisations with different rules, traditions, organisation structures, motivations for profit, nationality, laws, culture and languages [26, 30] (see Figure 1).



The Virtual Museum concept

Users (WORLD)
- Account info
- Privileges
- Profiles
- Policy

Private collections
- Content
- Services
- Policy
- Security&IPR
- Links

Repositories at the Content providers
- Content
- Services
- Security
- Policy
- IPR

Other Services
- DBMS management
- E-Commerce
- E-pay

VM ARTERY

Focal Point For:
- Users and Groups
- Links to repositories
- Security &Delegation
- Services&standards
- Content&Policy
- IPR

Network Devices
- Configuration
- QoS policy
- Security policy

Total Register
- East Christian Icons
- Wall paintings
- Mosaics
- Architecture churches, monasteries

Servers
- Infrastructure
- MM DBMS
- Mirrors
- NLB
- Security

Applications on which Virt. Museum is based
- Image-software
- Auto-classification
- Specific applications
- Iconoclature
- Digital workstation
- ‘Frame of reference’

Firewall Services
- Configuration
- Security Policy
- VPN policy
- Water mark
- Digital signature
- Encryption
- Transport protection

Fig. 1. Virtual Museum example concept

There are many challenges while working on the given task:

– A necessity for clear definition of the user's needs of some specific functionality;

– The presentation of information content in the most suitable way for the chosen user types, the content's ability to be easily found and reached;

– Assuring reusability of the resources in specific context and situation; adapting resources;

– Searching for flexible conceptual solutions, which are easily transferable and implementable via new technological means;

– Synchronization with established standards and specifications, etc.

These questions suggest deep research and analysis of the different components of the system – content, user needs, offered services and its applications. The following are of great importance:

– Building of a straightforward model/specification of the activities that the system will serve;

– Developing and introducing proper functionalities for ensuring flexible access to the resources;

– Analysis of the context in which the resources will be used (including educational one) and searching for methodological approaches and techniques for improving the access to the resources to meet the user needs to the highest extent.

3. Virtual museum system architecture. The virtual museum mainly contains service panels for *Museum content management*, *Museum content presentation*, *Administrative services* (see Figure 2), jointed to *a Media repository* and a *User data repository*.

The *Museum content management* module refers to the activities related to basic content creation: add (annotate and semantic indexing), store, edit, preview, delete, group, and manage multimedia digital objects; manage metadata; search, select (filter), access and browse digital objects.

The Museum content presentation module supports objects and collections display. It also provides collections creation (incl. search, select/browse and group multimedia digital objects according to different criteria and/or context of usage), their metadata/semantic descriptions and attractive visualization, status of collection display.

Content presentation module aims to provide access to all virtual museum services through wide range of contemporary technologies and devices – not only desktop PCs, but mobile phones, tablets, TVs, VR devices, etc. Interactive media technologies are used to provide best user experience within the content of the virtual museum.

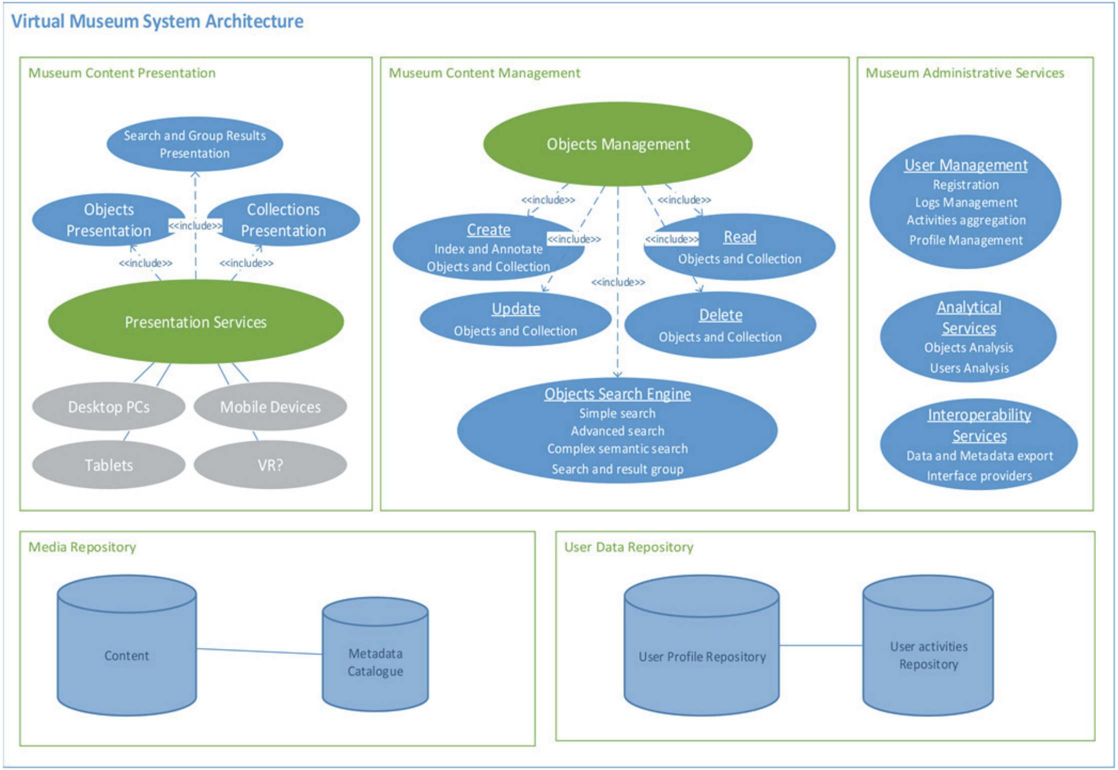


Fig. 2. Virtual Museum System Architecture

The *Administrative services panel* mainly provides user data management, data export, tracking and analysis services.

For every object all semantic and technical metadata are saved in the Media repository. These metadata are represented in catalogue records that point to the original media file/s associated to every object.

The User profile repository manages all user data and their changes.

4. Virtual museum functionalities in details.

4.1. *Museum content management*. The main part of the content creation process is the annotation and semantic indexing of digital objects in order to add them to the museum media repositories. The entering of technical and semantic metadata for the digital objects is implemented through different automated annotation and indexing services.

The technical metadata, that could be expressed in Dublin Core or other standards, are attached to every multimedia object automatically. They cover the general technical information, such as file type and format, identifier, date, provider, publisher, contributor, language, rights, etc.

An annotation template needs to be implemented for the semantic description of digital objects. The template provides several options for easy and fast entering of metadata:

– Autocomplete services (All used (already entered) field values are available in a panel for reuse.);

– Automated appearance of dependencies coming from the relations of the defined classes' (concepts) in an ontological descriptive structure valid for the museum object domain. All main relations and rules expressed in the ontological structure are incorporated during the development of the annotation template;

– Bilingual data entering with automated relation between the relevant values in different languages (if it is applicable or necessary);

– Automated appearance of the number of the used field value, providing regular data tracking (if it is applicable or necessary);

– A tree-based structure of the annotation template. Only checked fields are displayed for entering metadata;

– Possibility for adding more than one media for one metadata description in order to create rich heterogeneous multimedia digital objects tracking (if it is applicable or necessary);

– Reuse of an already created annotation for new objects: the new media object has to replace the older one, the annotation is kept and the new object appears after saving;

– Automated watermarking of the image and video objects;

– Automated resizing and compression of the media objects (image or video);

– Automated identification of file formats;

– Automated conversion of the media object (audio, video, text) in a format suitable for Web-preview;

– Automated terms explanation (if it is applicable or necessary): After saving a new digital object in the museum media repository, a special machine traces for the appearance of dictionary terms in the object metadata description. If some terms are available, the machine adds links to their explanations. In the case of entering a new dictionary term, its presence in the available objects is discovered automatically and a link is added.

– Digital object duplication checks (similarities calculation): In order to avoid duplicate image objects a service that checks the similarity between images is provided. It uses an algorithm that caching images for optimizing their compare (see [1]).

The virtual museum provides a wide range of search services, such as keyword search, extended keyword search, semantic-based search, complex search, search with grouping results, etc. Their realization is based on querying action to the metadata knowledge base. Moreover, five types of conditions for the results set are meant:

– "objects having or NOT characteristic c";

– "objects having value =, ≠, ≤, ≥, < or > for characteristic c". In the search templates, the user could search digital objects with précised criteria.

The search services support content request and delivery via index-based search and browse of managed content and its description.

4.2. *Museum content presentation*. Content presentation is a key activity in the virtual museum. The proper design and intelligent implementation of this service provide a stable base for overall VM functionality.

During the design of the content presentation services a profound analysis was made of content selection and preview possibilities in order to satisfy the user's needs. First, it is necessary to determine the preview possibilities of a separate digital object and its components and after that the preview of grouped objects (collection preview).

The visualization of the rich semantic description of the separate digital object is determined through hidden parts appearing in a new window after link selection. This possibility is used mainly for the long descriptions and for the dictionary terms. Parts of the descriptive data field are hidden, but their values are available for searching in special forms.

During the design of object grouping services, the main ontology classes of the object descriptive structure (*viz.* museum domain ontology) could be selected as object grouping criteria.

Every user can create his private collection of selected objects after search activity. Rich search possibilities (mentioned above) are available in order to assist collection creation. The user can write the collection's title and short description. He can also select its status: private or shared with other users. New objects for a collection appear automatically after their entering.

Custom collections in virtual museum are dynamic objects. They are criteria based, not list based. Every new object in the virtual museum will be automatically added to a collection, if it meets the criteria defined for the collection. The owner/creator/follower of a collection can be notified when a new object is added and becomes part of the collection. This service uses the following rule:

Let $P = \{p_1, p_2, ..., p_n\}$ be the set of all iconographical objects.

Let $A_m = \{a_{m1}, a_{m2}, \ldots, a_{mk}\}$ be the set (also called a collection) with k iconographical objects with a selected characteristic m**,** $A_m \subseteq P$.

Let p_{ti} be a new iconographical object, added to the library, $P = P \cup \{p_{ti}\}$.

IF $t \equiv m$ THEN $A_m = A_m \cup \{p_{ti}\}$.

Let $M = \{m_1, m_2, ..., m_r\}$ be a set of characteristics for a collection $A_M = \{a_{M1}, a_{M2}, ..., a_{Mk}\}$ with k iconographical objects.

Let p_{ti} be a new iconographical object, added to the library, $P = P \cup \{p_{ti}\}$, and $M' = \{m_1', m_2', ..., m_r'\}$ be its set of characteristics.

IF $M \subseteq M'$ THEN $A_M = A_M \cup \{p_{it}\}$.

The home page of the library contains a panel with last visited objects, aiding the user's observation of the content. This service uses the following algorithm:

Let p_i be an iconographical object, and $P = \{p_1, p_2, ..., p_n\}$ be the set of all iconographical objects.

Let t_j be the time an object was visited $T = \{t_1, t_2, ..., t_m\}$

$Q = P \times T$

(p_i, t_j) means that the object p_i was visited at the time t_j .

Steps of the algorithm:

1. Create series $Q' = \{(p_{i_1}, t_{j1}), (p_{i_2}, t_{j2}), ..., (p_{i_d}, t_{jd})\}$, where $t_{j1} > t_{j2} > .... > t_{jd}$.

2. Remove all (p_{i_k}, t_{jk}) , where $\exists t_{jl} : l < k \; \& \; p_{i_k} \equiv p_{jl}$.

3. Select first $\{q_1, ..., q_v\} \in Q'$.

Every object and collection can be presented in interactive way using any browser compatible device – PC, phone, tablet, or TV. Content presentation services are based on responsive web technologies in order to satisfy the majority of the modern devices diversity and provide great user experience no matter if user consumes the virtual museum services through their phone, tablet, PC, or other smart device. Moreover, the virtual museum should provide technological solutions for people with disabilities [22, 23, 24].

4.3. *Administrative Services*. The Administrative services panel mainly provides user data management, data export, tracking services, and analysis services. The user data management covers the activities related to registration, data changes, level set, and tracking activities of the user. The tracking services have two main branches: tracking of objects, tracking of user' activities. The tracking of objects spies on the activities of add, edit, preview, search, delete, selection, export to XML, and group of objects/collections in order to provide a wide range of statistic data (for frequency of service use, failed requests, etc.) for internal use and generation of inferences about the stable work (stability) and the flexibility of the work and the reliability of the environment. The tracking of users' activities monitors user logs, personal data changes, access level changes and user behaviour in the museum environment. The QlickTech® QlinView® Business Intelligence software could be used as an analysis provider. Therefore, it needs to be connected to the museum tracking services and objects data base by preliminary created data warehouse. It will provide fast, powerful and visual in-memory analysis based on online analytical processing and quickly answered multi-dimensional analytical queries [2]. This information can be used for making conclusions about people's interest in objects, collections and the museum content, in order to further fill the repository of the museum.

The export data services provide the transfer of information packages (for example, packages with digital objects/collections, user profiles, etc.) compatible with other data base systems. For example, with these services a package with objects could be transported in an XML-based structure for new external use in e-learning or e-commerce applications. Moreover, VMs power will increase significantly if they use mechanisms for ubiquitous sharing of their e-artefacts and they distribute attractive content in the social networks, reflecting community demands and needs [27].

During the VM system design the interoperability services and protocols need to be have in mind, *viz.*:

– Services concerning to interoperability and integration - describe the ways in which repositories work with other systems using common standards and protocols. Sometimes these interfaces are used directly by people (e.g. web user interfaces or RSS feeds) and sometimes they are used by machines (e.g. OAI-PMH and SWORD). Interfaces used by machines are sometimes referred to as m2m (machine-to-machine) interfaces;

– Services supporting linking mechanism - for effective use of distributed electronic resources in libraries. Some examples are Open URL linking, Link resolvers, Electronic resource integration, DOI, CrossRef, Handle.Net. Linking mechanism makes possible to build global museum services and portals, because it provides unique item identifiers, persistent identifiers are used for citation management, etc.;

4.4. *Storage and long term preservation of digital information.* Storage Management gives many benefits:

Digital content relies on common standards for metadata, storage data formats, indexing, etc. with necessity for provision of support for the whole live cycle. The term storage management encompasses the technologies and processes organizations use to maximize or improve the performance of their data storage resources. It's a broad category that includes virtualization, replication, mirroring, security, compression, traffic analysis, process automation, storage provisioning and related techniques. Storage management techniques can be applied to primary, backup or archived storage. Deployment and implementation procedures will vary widely depending on the type of storage management selected and the vendor. In addition, the skills and training of storage administrators and other personnel add another level to an organization's storage management capabilities. There are numerous storage technologies that impact the storage and long term preservation of digital information design, following principles of consolidation of storage into a central location, removal of the storage burden from host OSes, and so on.

Technologies, like storage virtualization, deduplication and compression, allow better utilization of existing storage, resulting in lower costs for operating and maintain storage devices. They simplify the management of storage networks and devices, reducing overall storage operating costs.

The appropriate storage management improves digital data reliability, performance, availability, agility and resilience.

Technologies like replication, mirroring and security are often particularly important for backup and archive digital information. Capacity optimization technologies like parity RAID, delta snapshots, thin provisioning etc. are applicable to storage of both structured and unstructured data.

Over them polices according to systems and software management, physical security, data security, data backups, disaster recovery, redundancy of data (multiple data duplication, digital archives, global web portals, providing content aggregation from various sources distributed over the Internet), are applied.

The emerging new generation of information technologies is a synergy of business intelligence analytics, from personalization towards making data-driven decisions and forecasts providing an integrated business solution gradually alienated from the software toward services and functionalities offered to the users.

The important services in the contemporary virtual museum are: content creation and presentation, crawling, storage, browse, measurement, retrieval, classification/ categorization, filtering, clustering, summarization, mining, preservation, decision support, user modelling/ personalization, etc. A main task for the developers is the proper design and intelligent implementation of these services. The design and the implementation of the described services result from a long-term observation of the users' preferences, cognitive goals, needs, object observation style, and interests, made during the testing processes in several digital content management systems. The main goal was the satisfaction of users' preferences and needs with appropriate navigation, visualization and content presentation techniques.

The authors actively try to develop workable solutions in the field of cultural heritage database management and presentation. The above described solution for virtual museum architecture follows previous developments in the digital content management systems (*viz.* digital libraries, digital repositories, galleries, etc.) for Bulgarian artworks and treasuries (see [3-9]). These systems are successfully implemented to presents the valuable Bulgarian cultural heritage: Bulgarian iconographic art, Bulgarian ethnographic and folklore artefacts, medieval and early modern Bulgarian texts for saints in combination with ethnological data and visual sources, church bells and plates, etc.

4.5. *Security of the Virtual Museum.* The Virtual Museum hosts digital data containing core assets, including user's information, intellectual property, and other critical content. With emerging trends such as Big Data, bring-your-own-device (BYOD) mobility, and global online collaboration sparking an explosion of data, the VM will only become more important and will extending be the target of advanced malware and other cyber attacks [25]. What is needed to be done to secure the Virtual Museum is:

– To shield the VM from advanced persistent threats (APTs) and sophisticated malware found in content stores, web and application servers, and common file shares.

– To stop attacks entering organizations via mobile devices and portable storage.

– To receive on-target analysis to pinpoint possible gaps that need addressing.

– To protect key assets and prevent attacks with products and services that work together and share and threat intelligence.

– To prevent attacks with a nimble, adaptive cyber security strategy.

– To safeguard VMs from attacks that use web servers and other data center infrastructure to host malware.

– To detect threats quickly to reduce lag time before resolution.

– To get reliable, fast malware analysis with agentless network-based threat detection and protection engine.

– To provide continuous, dynamic, non-disruptive resolution to incidents.

Recommended for Virtual Museum architecture is the Adaptive Defense approach to cyber security, which delivers technology, expertise, and intelligence in a unified, nimble framework, which demands the adaptation of the security architecture to prevent today's cyber attacks and avert their worst effects.

5. Intelligent data curation. In the modern era of big data, the curation of data has become more prominent, particularly for software processing high volume and complex data systems [10]. The term is also used in historical uses and cultural heritage digital assets and content management solutions, where increasing cultural and scholarly data from digital projects requires the expertise and analytical practices of data curation. In broad terms, curation means a range of activities and processes done to create, manage, maintain, and validate a component. Specifically, data curation is the attempt to determine what information is worth saving and for how long [11]. The essential elements of a powerful data curation tool are annotations, metadata, standards, models, databases, etc.

Moreover, data curation is intellectually intensive activity that is time consuming and requires a lot of dedicated resources. Taking into account the increasing role and amount of data, curation risks to be a bottleneck for any digital asset management or content management project in the long term. One of the challenges for the automation of data curation is difficulty in completing missing data and the level of granularity. Such a solution, however, looks practical because the data curation process is one of many iterations, consistency and includes complex data evaluation. The human and machine aspect need to be combined in order to solve the two most crucial data-integration problems: linkage of records (which often refers to linking records across disparate sources, referring to the same real-world entity) and schema mapping (mapping columns and attributes of different datasets).

One approach is to use a record to modularize curation processes. Splendiani [12] considers curation activities as functions in a "curation space” that is exemplified via a “curation record”. The curation process is broken down to the following classes of operations:

– *Schema mapping*: Machine-assisted process to identify and map the similar attributes from different data sources together in one, unified data set. The same entities (*e.g.,* events, studies, places) might be described by data sources of different origin in separate ways and in this case the usage of different schemas and vocabularies (a dataset schema is generally an official description of the main attributes and the values that can be taken by them). For instance, one source may refer to a person’s credentials by the means of two attributes (Name and Title), another source may use the terms Pers. Name and Royal Title, and a third might use PN and Rank, in order to address the same thing. The major activity in schema mapping is to set a mapping among those attributes. The problem may occur to be more challenging and may involve different conceptualizations especially in the cases when relationships in one source are represented as entities in another. Most often, in the ETL suites are used the most common schema mapping solutions that focus traditionally on the mapping of a small number of such schemas (usually less than ten) that deliver to users a suggested mapping that considers some similarity among column’s name and the content of them. With the maturity of the big data stack, however, the enterprises have the power to easily acquire a huge number of different data sources and have at their service applications that can ingest data sources as they are generated. An example from the pharmaceutical industry and the conducted clinical studies can be used, where tens of thousands of studies and assays are conducted by scientists across the world, often using separate technologies and a combination of local schemas and standards. It is essential for the companies’ businesses and is often required as mandatory by regulations and laws to use standardized and cross-mapping collected data. This approach has changed the main assumption of most solutions for schema mapping that the suggestions curated by users should be part of a manual process. In such a case the main encountered challenges are: (1) providing of automated solution that requires reasonable interaction with the user, meanwhile being able to map numerous schemas; and (2) designing of matching algorithms that are robust enough to accommodate different languages, formats, reference master data, and data units and granularity [13].

– *Standard setting*: Building a probabilistic machine learning model specific to the organization’s domain and stakeholders based on answering a series of yes/no questions to whether two records are the same. Given enough feedback, a pattern is captured that is required to build and maintain logic in order to generate de-duplicated, master data.

– *Validation*: Throughout the human-guided process to build a machine learning model for mastering data, the user is able to see measurable outputs for each item of yes/no feedback provided. The feedback directly corrects the model. This calculation is culminated in the ‘confusion matrix’, indicating the precision, recall, accuracy, and F score of the model based on human feedback [14].

Further defining the data curation process, it is based on the organization and integration of data collected from various sources. Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data" [15]. For example, in science, data curation may indicate the process of extraction of important information from scientific texts, such as research articles by experts, to be converted into an electronic format [16]. Using an efficient master data management solution can greatly facilitate the above-described process. However, a step further in the mastering process is, providing the ability to select a representative record, or golden record, for each set of duplicate, or grouped, data records derived from all data sources. For example, if a cluster of records is identified across systems for the same artist, but each record has a variation on the artist’s name the human-guided machine learning approach merges those records and generate a single golden record using the most common values for each attribute - assuming that aligns with how the business perceives their data. The goal of the golden record is to consolidate and generate a single record of truth. This approach is especially applicable when update of existing record is required from several different data sources.

Summing it all up, the curation process can be expressed in terms of rules that embed “atomic operations” like extractors, transformations, etc. The rules can rely on abstraction/inferences for higher genericity and can also be used to produce meta-information [12].

6. Model for intelligent data curation in virtual museum. *Record de-duplication issues.* The linkage of records, the resolution of entity and the deduplication of records are only a few of the terms that describe the need for unification of multiple mentions or database records that describe the same real-world entity. Concerning the example in Table 1 (showing a single schema for simplicity), it is obvious that the records are about Alexander, but they look quite different [17, 18, 19, 20].

Actually, all these records are correct or were correct at some point in time. It is easy for a well-qualified human to determine if such a cluster refers to the same entity, but it is hard for a machine to conduct this judgement. Therefore, more robust algorithms should be utilized to find such matches in the presence of errors, different styles of presentation and mismatches of granularity and references of time.

Table 1. Data unification at scale

Name	Attribute	Title	Year
Alexander			356 – 323 BC
Alexander	the Great		356 – 323 BC
Alexander	of Macedon		356 – 323 BC
Alexander	III		356 – 323 BC
Alexander	III the Great		356 – 323 BC
Alexander	III of Macedon		356 – 323 BC
Alexander	(all atributes)	King of Macedonia	336 – 323 BC
Alexander	(all atributes)	Basileus of Macedonia	336 – 323 BC
Alexander	(all atributes)	Hegemon of Hellenic League	336 - 323 BC
Alexander	(all atributes)	Pharaoh of Egypt	332 – 323 BC
Alexander	(all atributes)	King of Persia	330 – 323 BC
Alexander	(all atributes)	Lord of Asia	331 – 323 BC

The issue is an old one. In the recent decades, the community that conducts researches has come up with many similarity functions, supervised classifiers in order to differentiate matches from non-matches, and clustering algorithms for collecting matching pairs in the same group. Current algorithms can deal with thousands of records (or millions of records partitioned in disjointed groups of thousands of records), similar to schema mapping. Taking into account the massive amount of collected dirty data – and in the context of the abovementioned schema-mapping problem, a number of challenges are faced:

Challenge One: How to scale the quadratic problem (comparing every record to all other records, so computational complexity is quadratic in the number of records).

Challenge Two: How to train and build machine learning classifiers that handle the subtle similarities as in Table 1.

Challenge Three: How to engage humans and domain experts in providing training data, given the nature of the matches, which are rare in most cases.

Challenge Four: How to leverage the knowledge of all domains and previously developed rules and matchers in one integrated tool.

Talking about similarity, both the problems of schema mapping and deduplication occur after finding matching pairs (attributes in the case of schema mapping and records in the deduplication case).

Most of the building blocks can be reused and leveraged for both problems. Regarding correlation, most record matchers depend on some known schema for the two compared records; however, unifying schemas requires some type of schema mapping, even if incomplete. For this reason and many other, the solution at hand is for consolidating these activities and devising core matching and clustering building blocks for the unification of data that could: (1) be leveraged for different activities for unification (in order to avoid piecemeal solutions); (2) scale to a massive number of sources and data; and (3) have human in the loop as a guiding driver of the machine in building classifiers and applying the unification at large scale, in a trusted and explainable way. The idea is to use a human in the loop to resolve ambiguities when the algorithm's confidence on a match falls below a threshold [13].

When extracting data from different sources in cases of initial data upload or record updates, with large masses of data exists the risk of accumulating a lot of duplicate records. In this section will be presented a solution approach for deduplication. The first step is to look for mechanisms to enrich the data. In this way, extra fields can be added to each record which can assist in the deduplication process.

As a result of this step, each record has K attributes of information. The next step depends on the availability of training data. This consists of a collection of pairs of records which a human specifies as matches (*i.e.* duplicates) and a collection of pairs of records that are non-matches.

To this step fits a decision tree model in the following manner. For each attribute is requires a distance function, D (a1, a2), which specifies how far apart are any two values a1 and a2. In general, a distance function can be user-specified. However, for each character string attributes, Jacard and cosine similarity distance are popular metrics, and a human is asked to choose between these two. For numeric data are used arithmetic distance. For each attribute, is chosen a collection of split points based on dividing the training data into L equal sized buckets. Then, for each attribute it tries these L "split points ", and avidly chooses the attribute and the split point that most accurately classifies the training data. In effect each of the L* K cases is a predicate of the form:

Attribute-I < split point => non-match
Attribute-I >= split point => match
And
Attribute-I >= split point => non-match
Attribute-I < split point => match.

After that is selected the predicate that best fits the data at hand. With this "root node" chosen, continues the fit of the two second level nodes. It continues in this fashion until the benefit of additional levels is marginal or until a user-defined maximum depth, Max, is reached. In effect a decision tree model is fitted to the training data, with parameters D, L and Max.

If there is not enough training data active learning is used to get more. A "cluster review" process can also be employed. This step allows a human to review suggested matches and to correct ones that are in error. Hence, cluster review produces additional training data to refine the model used, and can be thought of as an active learning scheme.

So far are identified collections of records that it thinks represent the same entity, *i.e.* are duplicates. Consider one particular collection and resources that represents Alexander (356 – 323 BC) – a king (basileus) of the ancient kingdom of Macedon, as shown below.

Name:

Alexander the Great (Greek: *Αλέξανδρος ὁ Μέγας*, Bulgarian: *Александър Велики*)

Alexander of Macedon (Greek: *Αλέξανδρος ὁ Μακεδών*, Bulgarian: *Александър Македонски*)

Alexander III (Greek: *Αλέξανδρος Γʹ*, Bulgarian: *Александър III*)

Title with period of reign:

(Alexander) King of Macedonia (336 – 323 BC)

(Alexander) Basileus of Macedonia (336 – 323 BC)

(Alexander) Hegemon of Hellenic League (336 BC)

(Alexander) Pharaoh of Egypt (332 – 323 BC)

(Alexander) King of Persia (330–323 BC), etc.

Apparently, a canonical form for name, a resolution for several values for the title of the ruler that are attached to the name, and the recognition that have several different periods of reign, are requested.

First, user-specified column rules which define how to aggregate the column values in a cluster into a "golden value", are used. Also supported are the options to "choose the most frequent value", "majority consensus", "keep all values" and "choose average value". Based on applying these rules, each cluster of data is reduced to a simpler one with less multi-valued attributes.

Then, is examined each column, looking for patterns of values. For example, in the Alexander cluster, it removes the duplicate value "Alexander" and is left with:

III (The Third)

The Great

Of Macedon.

Then, it assumes that longer strings are better than shorter ones, and forms candidate substitution rules, as follows:

III of Macedon (*Γʹὁ Μακεδών*, *III Македонски*)

III The Great (*Γʹὁ Μέγας*, *III Велики*)

The above described example is based on content units and their descriptive metadata, for which is in process the development of a virtual museum of ancient history and civilization.

Similar cases can be often observed when documenting historic facts and events in the middle ages. There are situations with substantial number of versions for the name and title of historic figures like the medieval Bulgarian ruler Asparuh, named as Asparuh/Asparukh (Bulg. *Аспарух*), Isperih (Bulg. *Исперих*), Esperih (Bulg. *Есперих*), Ispor (Bulg. *Испор*), Asparhruk, Batiy, etc. with several versions of title "han", "khan", "knyaz", and "tsar".

Then are performed analysis for each multi-valued field in any column that does not have the "keep all" designator. The net result is a collection of possible rules and a count of the number of times each occurs.

Finally, the rules are sorted into frequency order and presents the first one to a human along with a sample of the clusters to which it applies. The human is asked to respond "yes", "no" or "maybe". The rule is automatically applied or discarded in the first two cases. In the third case, it asks a human to start tagging values as "correct" or "not correct". Based on this training data, is formed a decision tree model for the collection of clusters. This process of examining the most frequent possible rules continues until a human decides that the point of diminishing returns has occurred [21].

7. Conclusion. In the last decade the cultural heritage has been radically transformed, with broad acceptance of digital technologies facilitating the exhibition and sharing of its richness, the affirmation of national identity and the development of culture in society. The digital technologies provided the means for active inclusion, engaging and participation of broad user community in the processes of access, exploration and study of this wealth. The investigation of these essential transformations is of crucial value and covers studies of new means and methods for improved attainment, understanding, analysis and interpretation of cultural content, context-dependent use and sharing of the huge databases of cultural and historical objects, new forms of interaction with digital cultural and historical content through the modern digital data transfer channels, modern forms of communication, civic engagement, etc., towards national identity and development of culture in the society.

This paper is focused on the modern digital content management systems that have to cover a complex set of functionalities in order to operate as complete solutions. The management of digital objects in a virtual museum for cultural heritage requires a well-designed architecture that embeds services for content presentation, content management, administration of user data and analysis. This set of services is interdependent and demands a high level of data integrity, which is hard to achieve when digital objects originate from disparate data sources with specific and non-standardized data formats and elements. In such a scenario, intelligent data curation that leverages machine learning to clean and unify data, is a sound approach that increases efficiency and eliminates errors of duplicate and inaccurate data. In this process are employed logistic regression, decision trees and ad-hoc models that use training data to fit a model, often assisted by human feedback and active learning. These models undergo continuous evolution and get improved by additional techniques with each new use case.

Further investigations in the above-discussed domain point to a wide variety of directions:

– Creation of workable methods and tools, aiming to increase and generalize the visitors' experience in the virtual museum. Moreover, creative user experiences will support the effective on-line learning through virtual museums.

– Design and development of contextual techniques for personalizing user's work in these platforms.

– Design and development of multimodal interfaces and intelligent visualisation of complex and heterogeneous media objects relying on enhanced usability (incl. user-centric visualisation and analytics, real-time adaptable and interactive visualisation, real-time and collaborative 3D visualisation, dynamic clustering of information, etc.), etc.

The field has great potential for innovations, especially in present world of active imposition of new e-devices. The focus will be also in the research and exploitation of new or emerging technologies (e.g. 3D, augmented and virtual reality, visual computing, smart world, media convergence, social media, etc.) for the development of innovative products, tools, applications, and services for creative digital content production, usage and management. The aim is to transform and customize the valuable parts of mankind's cultural and historical ancestry into digital assets, whose integration and reuse through research-lead methods has high commercial and non-commercial potential for learning and cultural institutions, tourism, creative and media industries.

References

1. Pavlov R., Paneva-Marinova D., Goynov M., Pavlova-Draganova L. Services for content creation and presentation in an iconographical digital library. *International Journal "Serdica Journal of Computing"*. 2010. vol. 4. pp. 279–292.
2. Codd E., Codd S., Salley C. Providing OLAP to user-analysts. 1993. Available at: http://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf (accessed: 10.03.2019.).

3. Paneva-Marinova D., Goynov M., Luchev D. Multimedia digital library: Constructive block in ecosystems for digital cultural assets. Basic functionality and services. LAP LAMBERT Academic Publishing. 2017. 117 p.
4. Luchev D., Paneva-Marinova D., Pavlova-Draganova L., Pavlov R. New digital fashion world. 14th International Conference on Computer Systems and Technologies (CompSysTech'13). 2013. vol. 767. pp. 270–275.
5. Rangochev K., Goinov M., Dimitrova M., Hristova-Shomova I.. Enciclopaedia Slavica Sanctorum: activity, users, statistics. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2013. vol. 3. pp. 81–90.
6. Rangochev K., Dimitrova M.. [Two models for presenting the Balkan folklore heritage in digital libraries]. *Dobre doshli v Kiberiya: zapiski ot digitalniya teren – Welcome to Cyberbia: notes from the digital terrain*. 2014. pp. 397–411. (In Bulg.).
7. Bogdanova G., Todorov T.Y., Noev N. Using graph databases to represent knowledge base in the field of cultural heritage. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2016. vol. 6. pp. 199–206.
8. Pavlova-Draganova L., Paneva-Marinova D., Pavlov R., Goynov G. On the wider accessibility of the valuable phenomena of Orthodox iconography through digital library. Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010). 2010. pp. 173–178.
9. Bogdanova G., Todorov T.Y., Kancheva S. Virtual museum of Russian bells in Bulgaria. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2017. vol. 7. pp. 215–222.
10. Furht B., Escalante A. Handbook of data intensive computing. Springer-Verlag. 2011. 793 p.
11. Borgman C. Big data, little data, no data: scholarship in the networked world. MIT Press. 2015. 416 p.
12. Splendiani A. AI for data curation. Yes, can we? 2017. Available at: https://www.slideshare.net/sergentpepper/artificial-intelligence-in-data-curation (accessed: 10.03.2019.).
13. Ilyas I. Data unification at scale: data tamer. Making Databases Work. Association for Computing Machinery and Morgan & Claypool. 2018. pp. 269–277.
14. Tamr. Agile data mastering raising expectations for master data management (MDM). 2019. Available at: http://www.tamr.com: http://www.tamr.com/wp-content/uploads/2019/01/Tamr_WP_Agile-Data-Mastering-_01-14-19.pdf (accessed: 10.02.2019.).
15. Miller R. Big data curation. Proceedings of the 20th International Conference on Management of Data (COMAD). Computer Society of India. 2014. pp. 4.
16. Blank G. Studyguide for the sage handbook of Internet and online research methods. 2012. Cram101. 80 p.
17. Walbank F. Alexander the Great: King of Macedonia. 2019. Available at: https://www.britannica.com/biography/Alexander-the-Great (accessed: 01.02.2019.).
18. O’ Brien J. Alexander the Great: The Invisible Enemy: A Biography. Routledge. 2005. 360 p.
19. Bosworth A.B., Baynham E.J. Alexander the Great in fact and fiction. Oxford University Press. 2000. 384 p.
20. Green P. Alexander of Macedon, 356—323 B.C.: A historical biography. University of California Press. 1991. 617 p.
21. Stonebraker M. Machine learning for data unification practical applications in Tamr’s software platform. 2017. Available at: https://www.tamr.com/wp-

content/uploads/2017/07/Machine_Learning_For_Data_Unification_072117_2.pdf (accessed: 10.03.2019.).

22. Georgieva-Tsaneva G., Subev N. Technologies, Standarts and Approaches to Ensure Web Accessibility for Visually Impaired People. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2018. vol. 8. pp. 143–150.
23. Bogdanova G., Noev N. Digitization and preservation of digital resources and their accessibility for blind people. *Cyber-physical systems for social applications*. 2019. pp. 184–206.
24. Karpov A., Ronzhin A. A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces. Proceedings of the 8th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2014). 2014. pp. 369–378.
25. Yoshinov R., Kotseva, M, Pavlova D. Specifications for Centralized DataCenter serving the educational cloud for Bulgaria. Proceedings of XII International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2015. pp. 1–6.
26. Yoshinov R. Bringing up-to-date the principles of the I.DB.I. Artery. Proceedings of IX International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2009. pp. I3–1.
27. Yoshinov R., Iliev O. "Controlled self-study" in thematic educational community environment. The 47th Spring Conference of the Union of Bulgarian Mathematicians. 2018. pp. 200–213.
28. Yoshinov R., Iliev O. Content reuse – a major problem with modern content storage systems. Eleventh National Conference with International Participation "Education and Research in the Information Society". 2018.
29. Yoshinov R., Kotseva M. The steps for elaboration of the "Rosetta stone" demonstrator. Proceedings of International Conference Inspiring Science Education. 2016. pp. 91–96.
30. Yoshinov R., Arapi P., Kotseva M., Christodoulakis S. Supporting Personalized Learning Experiences on top of Multimedia Digital Libraries. *International journal of education and information technologies*. 2016. vol. 10. pp. 152–158.
31. Trifonov R., Yoshinov R., Jekov B., Pavlova G. Methodology for Assessment of Open Data. *International Journal of Computers*. 2017. vol. 2. pp. 28–37.
32. Yoshinov R.D., Iliev O.P. (2018) The Structural Way for Binding a Learning Material with Personal Preferences of Learners. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2018. vol. 5(60). pp. 189–215.

Paneva-Marinova Desislava Ivanova — Ph.D., associate professor, head of the Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, technologies for knowledge presentation and processing, data management and processing, data analytics, intelligent data curation, Semantic web, digital content management systems, web services. The number of publications — 112. dessi@cc.bas.bg; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +359888894814.

Stoikov Jordan Stoikov — Ph.D. student at the Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, business management systems and services for intelligent data curation, data integrity, knowledge retrieval, data validation. The number of publications — 2. jstoikov@shieldui.com; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +35929792874.

Pavlova Lilia Radoslavova — Ph.D., assistant professor at the Laboratory of Telematics, Bulgarian Academy of Sciences (BAS). Research interests: computer science, technologies for knowledge presentation and processing, Semantic web, e-learning. The number of publications — 34. pavlova.lilia@gmail.com; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +35929793831.

Luchev Detelin Mihailov — Ph.D., associate professor, Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, digital content management systems, technologies for knowledge presentation and processing, data management and processing, web services, mobile applications, history, ethnology. The number of publications — 75. dml@math.bas.bg; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +359885978788

Acknowledgements. This research is partially supported by the Bulgarian Ministry of Education and Science under the National Research Programme "Cultural heritage, national memory and development of society" approved by DCM №577/17.08.2018 and by the Bulgarian Scientific Fund under the research project № DN02/06/15.12.2016 "Concepts and Models for Innovation Ecosystems of Digital Cultural Assets".

УДК 006.72 **DOI** 10.15622/sp.18.2.444-470

Д.И. Панева-Маринова, Й.С. Стойков, Л.Р. Павлова, Д.М. Лучев

АРХИТЕКТУРА СИСТЕМЫ И ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ДАННЫХ ВИРТУАЛЬНОГО МУЗЕЯ ДРЕВНЕЙ ИСТОРИИ

Йошинов Р.Д., Панева-Маринова Д.И., Стойков Й.С., Павлова Л.Р., Лучев Д.М. **Архитектура системы и интеллектуальная обработка данных виртуального музея древней истории.**

Аннотация. Сохранение культурного и исторического наследия разных народов мира и их тщательное изложение – это долгосрочное обязательство ученых и исследователей, работающих во многих областях. На протяжении веков каждое поколение стремится вести учет своего труда, чтобы его могли пересмотреть и изучить следующие поколения. За последние пару лет были разработаны новые информационные и мультимедийные технологии, которые представили новые методы сохранения, обслуживания и распространения огромного количества собранного материала. Эта статья призвана представить виртуальный музей, передовую систему, управляющую разнообразными коллекциями цифровых объектов, которые по-разному организованы с помощью сложной специализированной функциональности. Управление цифровым содержанием требует хорошо продуманной архитектуры, которая включает в себя сервисы для представления, управления и администрирования содержания. Все элементы архитектуры системы взаимосвязаны, поэтому точность каждого элемента имеет большое значение. Эти системы страдают от недостатка инструментов для интеллектуального курирования данных с возможностью проверки данных из разных источников и повышения ценности данных. В этой статье предлагается решение для интеллектуального курирования данных, которое может быть реализовано в виртуальном музее, чтобы предоставить возможность надлежащим образом наблюдать ценные исторические образцы. Решение сфокусировано на процессах валидации и верификации, чтобы предотвратить дублирование записей цифровых объектов, чтобы гарантировать целость данных и более точный поиск знаний.

Ключевые слова: виртуальный музей, архитектура системы, функциональность, целостность данных, поиск знаний, проверка данных, дедупликация записей, культурное наследие.

Панева-Маринова Десислава Иванова — Ph.D., доцент, заведующая кафедрой математической лингвистики Института математики и информатики Болгарской академии наук (БАН). Научные интересы: информатика, технологии представления и обработки знаний, управление и обработка данных, аналитика данных, интеллектуальное курирование данных, семантическая сеть, системы управления цифровым контентом, веб-сервисы.Число научных публикаций — 112. dessi@cc.bas.bg; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +359888894814.

Стойков Йордан Стойков — аспирант кафедрой математической лингвистики Института математики и информатики Болгарской академии наук (БАН). Сфера научных интересов: компьютерные науки, системы и услуги управления бизнесом для интеллектуального хранения данных, целостности данных, поиска знаний, проверки данных.Число научных публикаций — 2. jstoikov@shieldui.com; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +35929792874.

Павлова Лилия Радославова — Ph.D., научный сотрудник Лаборатории телематики Болгарской академии наук (БАН). Сфера научных интересов: информатика, технологии представления и обработки знаний, семантическая паутина, электронное обучение.Число научных публикаций — 34. pavlova.lilia@gmail.com; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +35929793831.

Лучев Детелин Михайлов — Ph.D., доцент кафедры математической лингвистики Института математики и информатики Болгарской академии наук (БАН). Сфера научных интересов: информатика, системы управления цифровым контентом, технологии представления и обработки знаний, управление и обработка данных, веб-сервисы, мобильные приложения, история, этнология. Число научных публикаций — 75. dml@math.bas.bg; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +359885978788

Поддержка исследований. Исследование выполнено при частичной финансовой поддержке Министерством образования и науки Болгарии в рамках Национальной исследовательской программы «Культурное наследие, национальная память и развитие общества», DCM № 577/17.08.2018, и Болгарским научным фондом в рамках исследовательского проекта № DN02/06/15.12.2016 «Концепции и модели инновационных экосистем цифровых культурных ценностей».

Литература

1. *Pavlov R., Paneva-Marinova D., Goynov M., Pavlova-Draganova L.* Services for content creation and presentation in an iconographical digital library // International Journal "Serdica Journal of Computing". 2010. vol. 4. pp. 279–292.
2. *Codd E., Codd S., Salley C.* Providing OLAP to user-analysts. 1993. URL: http://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf (дата обращения: 10.03.2019.).
3. *Paneva-Marinova D., Goynov M., Luchev D.* Multimedia digital library: Constructive block in ecosystems for digital cultural assets. Basic functionality and services // LAP LAMBERT Academic Publishing. 2017. 117 p.
4. *Luchev D., Paneva-Marinova D., Pavlova-Draganova L., Pavlov R.* New digital fashion world // 14th International Conference on Computer Systems and Technologies (CompSysTech'13). 2013. vol. 767. pp. 270–275.
5. *Rangochev K., Goinov M., Dimitrova M., Hristova-Shomova I.*. Enciclopaedia Slavica Sanctorum: activity, users, statistics // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2013. vol. 3. pp. 81–90.
6. *Рангочев К., Димитрова М.*. Два модела за представяне на българското фолклорно наследство в цифрови библиотеки // Добре дошли в Киберия: записки от дигиталния терен. 2014. С. 397–411.
7. *Bogdanova G., Todorov T. Y., Noev N.* Using graph databases to represent knowledge base in the field of cultural heritage // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2016. vol. 6. pp. 199–206.
8. *Pavlova-Draganova L., Paneva-Marinova D., Pavlov R., Goynov G.* On the wider accessibility of the valuable phenomena of Orthodox iconography through digital library // Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010). 2010. pp. 173–178.
9. *Bogdanova G., Todorov T.Y., Kancheva S.* Virtual museum of Russian bells in Bulgaria // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2017. vol. 7. pp. 215–222.

10. *Furht B., Escalante A.* Handbook of data intensive computing (1 ed.) // Springer-Verlag. 2011. 793 p.
11. *Borgman C.* Big data, little data, no data: scholarship in the networked world // MIT Press. 2015. 416 p.
12. *Splendiani A.* AI for data curation. Yes, can we? 2017. URL: https://www.slideshare.net/sergentpepper/artificial-intelligence-in-data-curation (дата обращения: 10.03.2019.).
13. *Ilyas I.* Data unification at scale: data tamer. Making Databases Work // Association for Computing Machinery and Morgan & Claypool. 2018. pp. 269–277.
14. Tamr. Agile data mastering raising expectations for master data management (MDM). 2019. URL: http://www.tamr.com: http://www.tamr.com/wp-content/uploads/2019/01/Tamr_WP_Agile-Data-Mastering-_01-14-19.pdf (дата обращения: 10.02.2019.).
15. *Miller R.* Big data curation // Proceedings of the 20th International Conference on Management of Data (COMAD). Computer Society of India. 2014. pp. 4.
16. *Blank G.* Studyguide for the sage handbook of Internet and online research methods. 2012. Cram101. 80 p.
17. *Walbank F.* Alexander the Great: King of Macedonia. 2019. URL: https://www.britannica.com/biography/Alexander-the-Great (дата обращения: 01.02.2019.).
18. *O' Brien J.* Alexander the Great: The Invisible Enemy: A Biography // Routledge. 2005. 360 p.
19. *Bosworth A.B., Baynham E.J.* Alexander the Great in fact and fiction // Oxford University Press. 2000. 384 p.
20. *Green P.* Alexander of Macedon, 356–323 B.C.: A historical biography // University of California Press. 1991. 617 p.
21. *Stonebraker M.* Machine learning for data unification practical applications in Tamr's software platform. 2017. URL: https://www.tamr.com/wp-content/uploads/2017/07/Machine_Learning_For_Data_Unification_072117_2.pdf (дата обращения: 10.03.2019.).
22. *Georgieva-Tsaneva G., Subev N.* Technologies, Standarts and Approaches to Ensure Web Accessibility for Visually Impaired People // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2018. vol. 8. pp. 143–150.
23. *Bogdanova G., Noev N.* Digitization and preservation of digital resources and their accessibility for blind people // Cyber-physical systems for social applications. 2019. pp. 184–206.
24. *Karpov A., Ronzhin A.* A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces // Proceedings of the 8th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2014). 2014. pp. 369–378.
25. *Yoshinov R., Kotseva, M, Pavlova D.* Specifications for Centralized DataCenter serving the educational cloud for Bulgaria // Proceedings of XII International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2015. pp. 1–6.
26. *Yoshinov R.* Bringing up-to-date the principles of the I.DB.I. Artery // Proceedings of IX International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2009. pp. I3–1.
27. *Yoshinov R., Iliev O.* "Controlled self-study" in thematic educational community environment // The 47th Spring Conference of the Union of Bulgarian Mathematicians. 2018. pp. 200–213.
28. *Yoshinov R., Iliev O.* Content reuse – a major problem with modern content storage systems // Eleventh National Conference with International Participation "Education and Research in the Information Society". 2018.

29. *Yoshinov R., Kotseva M.* The steps for elaboration of the “Rosetta stone” demonstrator // Proceedings of International Conference Inspiring Science Education. 2016. pp. 91–96.

30. *Yoshinov R., Arapi P., Kotseva M., Christodoulakis S.* Supporting Personalized Learning Experiences on top of Multimedia Digital Libraries // International journal of education and information technologies. 2016. vol. 10. pp. 152–158.

31. *Trifonov R., Yoshinov R., Jekov B., Pavlova G.* Methodology for Assessment of Open Data // International Journal of Computers. 2017. vol. 2. pp. 28–37.

32. *Yoshinov R.D., Iliev O.P.* (2018) The Structural Way for Binding a Learning Material with Personal Preferences of Learners // Труды СПИИРАН. 2018. Вып. 5(60). pp. 189–215.