CONTENT ANALYZING AND SYNTHESIZING SERVICES IN A DIGITAL LIBRARY

Desislava Paneva-Marinova, Maxim Goynov, Radoslav Pavlov

***Abstract**: Current research on digital libraries (DL) is mostly focused on the generation of large collections of multimedia resources and regular tools for their indexing and retrieval. However, digital libraries should provide more than advanced content maintenance and retrieval services. They should aid the users in their content observation, knowledge acquisition and better satisfying their needs, interests and wishes. This paper presents an extension of the current DL functionality with content analyzing services. The main goal is to reach implicit and hidden data, content, rules and facts, dependences and tendencies, valid for the content in the DL repository, to synthesize and summarize the collected data in order to use them in various investigations and learning. These services also observe the DL tracking services' output and provide different inferences for the frequency of service usage, failed requests, user logs and activities, etc., assisting the DL environment maintenance through the generation of inferences about its stability, flexibility, and reliability. This interpretation of DL analyzing services is not proposed and analyzed until now. We try to push up a new research point, aiming to aid user's work in the DL environment.*

***Keywords**: Digital Libraries, Services, Content Analysis, East-Christian Iconographical Art.*

***ACM Classification Keywords**: H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues*

Introduction

Digital libraries "should enable any citizen to access all human knowledge anytime and anywhere, in a friendly, multi-modal, efficient, and effective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices" [Bertino et al., 2001]. The key for such an environment and its efficiency is the provisioning of strictly designed functionalities, which are powered by the observation of the users' preferences, cognitive goals, and needs in order to find optimal functionality solutions for the end users. Current DL releases mainly provide content management services such as: content creation, i.e., adding (annotating and semantic indexing), storing, editing, previewing, browsing, deleting, grouping, and managing multimedia digital objects (images, text, sound, video), collections and their descriptions; metadata management; simple and extended keyword search, complex semantic and context-based search, selection and grouping of objects; data export, etc. A natural extension of these services could be the DL content analyzing, synthesizing and summarizing services, providing content and functionality observation, mining, inference, evaluation and tracking.

Current work in DL content analysis mainly concerns the improvement of the DL content by identifying areas of knowledge that are lacking content and using external information sources to augment the existing knowledge [Carmel et al., 2008]. The DL knowledge management systems provide solutions for acquiring new knowledge by either pulling potential data from external sources or by having the data pushed directly from external content providers. Topic-driven and user-driven focused crawling are the mainly used techniques for finding missing content [Chakrabarti et al., 1999][Pant et al., 2004][Zhuang et al., 2005].

In this paper we will present a different idea for DL content analysis. Our work focuses first on the search of implicit data/content, context, rules, facts, dependences, tendencies, etc., valid for the content in the DL repository, and the synthesizing and summarizing of the collected data in order to use them in various investigations and learning. We also analyze the DL tracking services' output in order to provide different inferences for the frequency of service usage, failed requests, user logs and activities, etc., aiding the DL environment maintenance through the generation of inferences about its stability, flexibility, and reliability.

The experiments are performed in the "Virtual Encyclopedia of the East-Christian Art" multimedia digital library (also called Bulgarian Iconography Digital Library, BIDL), whose specification is included in section 4 of the paper. Section 2 tracks current work in digital library content analysis. Section 3 formulates investigations and content analysis that are typically done in the target iconographical domain. The overview, design and implementation of the analyzing services in Bulgarian iconography digital library are presented in Section 5 and Section 6.

Current Work in Digital Library Content Analysis

The idea of digital library content analysis appeared in order to answer the question how the content of a digital library can be enhanced to better satisfy users' needs, interests, wishes. Missing content is identified by finding missing content topics in the system's query log or in a pre-defined taxonomy of required knowledge. The collection is then enhanced with new relevant knowledge, which is extracted from external sources that satisfy those missing content topics. Experiments of Carmel's team measured the precision of the system before and after content enhancement [Carmel et al., 2008]. The results demonstrate a significant improvement in the system effectiveness as a result of content enhancement and the superiority of the missing content enhancement policy over several other possible policies.

Other solutions are provided by the DL knowledge management systems that acquire new knowledge by either pulling potential data from external sources or by having the data pushed directly from external content providers. Topic-driven and user-driven focused crawling are the mainly used techniques for finding missing content (However, digital libraries that are based on active crawling methods such as CiteSeer often have missing documents in collections of archived publications, such as ACM and IEEE). The goal of the focused crawler is to selectively seek out pages that are relevant to a pre-defined set of *topics*. The topics are specified not using keywords, but using exemplary documents. Rather than collecting and indexing all accessible Web documents to be able to answer all possible ad-hoc queries, a focused crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the Web. The basic concept of a focused crawler (topical crawlers) [De Bra et al., 1994], is based on a crawling strategy that relevant Web pages contain more relevant links, and these relevant links should be explored first. Initially, the measure of relevancy was based on keywords matching; connectivity-based metrics were later introduced [Cho et al., 1998]. In [Chakrabarti et al., 1999] the concept of a focused crawler was formally introduced as a crawler that seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the Web. This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date.

Pant’s teams developed a topical crawler [Pant et al., 2004]. Its crawler follows hyperlinks to automatically retrieve pages from the Web while biasing its search towards topically relevant portions of the Web. A trained classifier provides the crawler with the needed bias. Once a collection of Web pages has been downloaded by the crawler, the system analyzed them to find more structured information such as potential Web communities and their descriptions. The analysis process includes both lexical as well as link (graph) based analysis. The final result of the analysis is then shown as an interactive graphical report that describes various clusters (potential communities) found through the crawl, their examples, as well as authorities and hubs within each cluster.

Today, focused crawling techniques have become more important for building specialty and niche (vertical) search engines. While both the sheer volume of the Web and its highly dynamic content increasingly challenge the task of document collection, digital libraries based on crawling benefit from focused crawlers since they can quickly harvest a high-quality subset of the relevant online documents.

Most of the current focused crawling approaches perform syntactic matching, that is, they retrieve documents that contain particular keywords from the user's query. Unfortunately, this often leads to poor discovery results, because the keywords in the query can be semantically similar but syntactically different, or vice-versa. Moreover, the query matching score is calculated taking into account only the keywords from the user's query. Thus, regardless of the context, the same list of results is returned in response to a particular query. In [Pahal et al., 2007] it is offered an approach for document discovery building on a comprehensive framework for context-ontology driven focused crawling of Web documents.

Su’s team presented an intelligent focused crawler algorithm [Su et al., 2005] in which they embedded ontology to evaluate the page’s relevance to the topic. Compared with other algorithms using domain knowledge, this algorithm can evolve the ontology automatically during crawl process. Considering the instinct characteristics of the ontology, propagation has also been imported to accelerate the evolution of the ontology. This approach is applied in several tasks and provided an empirical evaluation which has shown promising results.

The possible interpretations of the DL content analysis beyond the crawling techniques and solutions. In our work we focused on the search of implicit data/content, context, rules, facts, dependences, tendencies, etc., valid for the content in the DL repository, and then we synthesize and summarize the collected data in order to use them in various investigations and training situations. We also analyze the DL tracking services’ output in order to provide different inferences for the frequency of service usage, failed requests, user logs and activities, etc., aiding the DL environment maintenance through the generation of inferences about its stability, flexibility, and reliability. This interpretation of DL content analysis is not proposed and analyzed until now. We try to push up a new research point for the content analysis, aiding user’s work in the DL environment.

Needs of Content Analysis in the Iconographical Art Domain

Analyzing iconographical artifacts is an activity performed mostly by art experts, theologians, restoration specialists, and art researchers. It subsumes, inter alia, analysis of the theological meaning of iconographical images, art analysis of the tendencies in iconography, the development in time of characters and scenes, the occurrence and activities in iconographical schools, style similarities between objects, periodizing iconographical tendencies, tracing the iconographical technologies in different time periods, iconographical schools, and authors, technological analysis of pieces of art (researching the base, ground, painting layer, polish, etc.), researching the donors' and authors' writings, authenticating the object, researching the objects' origin, current condition, state, restoration traces, overpaintings, etc. Simple activities helping the analysis are: building (selecting) a collection of samples having certain characteristics (properties), certain values of the properties, having restrictions/rules for the property values; determining the strength of the chosen object set, the internal order and grouping of the objects, displaying the collection, choice evaluation, etc. At present this work is done by hand, which takes much time and effort.

An example of a simple task for iconographical arts critics is to make an art critical analysis of the development in time of the iconographic image of Jesus Christ in the various iconographical schools in Bulgaria. The researchers have to perform the following steps:

- Select a certain number of iconographic objects containing the image of Jesus Christ in a one-figure composition. (Note: The right choice requires selecting iconographic objects with the character or Jesus Christ Pantocrator, or Blessing Christ, or Jesus Christ enthroned, or St. Veronica, etc.)
- Arrange the iconographic objects in groups by school of iconography.
- If a school of iconography's group contains objects by an eminent author and founder of the school, place these high on the list. Among the objects designated for art critical analysis there should be at least one by a prominent author/school founder, if available.
- Ensure that the iconographic objects designated for art critical analysis are currently in good condition.
- Ensure that at least one primitive iconographic object and at least one Renaissance iconographic object are included in the iconographic objects designated for art critical analysis.
- In writing the art critical analysis compare the selected iconographic objects by contrasting clothing, gesture/s, the character proportions, object/s, the presence of other character/s and/or symbol/s, backgrounds, other element/s (e.g., clouds, etc.) in the iconography of the image of Christ. Look for changes in the iconography of these components, for example, appearance or lack of components (objects, symbols, characters, etc.), changes in the background, clothing, etc., in the selected set of samples.

Another example is the sample task for the art technique team. It has to find iconographic artifacts/objects containing the image of Jesus Christ in order to compare their specifics from a technological point of view.

Steps to be performed:

- Find all the iconographic scenes with Jesus Christ.
- Choose one iconographic scene with a Lord's Day (Holy Cross, Nativity, Epiphany, Palm Sunday, Ascension, Pentecost and Transfiguration), with the most samples (iconographic objects), minimum 6.
- Ensure the selected iconographic objects are on solid base (wood, stone and metal, bone, glass).
- Ensure only iconographic techniques (tempera, oil, mixed) are used in the painting of the iconographic objects.
- Ensure the iconographic objects contain gilding.
- Ensure the Iconographic objects are arranged by temporal characteristics, for example, century.
- In writing the analysis compare iconographic objects in one or more iconographic techniques and evaluate the quality of their execution. Look for periodisations of the employed iconographic techniques

in the selected set of samples. Examine the type and technology of the gilding and the structure of the base.

In our work we try to execute these tasks in a DL environment in order to simplify the specialists' work. These examples of analysis constitute a real case for learning-by-authoring in a scenario for technology-enhanced learning process [Pavlova-Draganova et al., 2009] in the frames of the SINUS project "Semantic Technologies for Web Services and Technology Enhanced Learning"[1]. The main goal is to demonstrate creative learning-by-doing through active authoring of specific learning materials on East-Christian iconography by learners, using multimedia and information resources delivered through the "Virtual Encyclopedia of the East-Christian Art" multimedia digital library [Paneva-Marinova et al., 2010][Pavlov et al., 2010][Pavlova-Draganova et al., 2010]. SINUS's learning analysis solutions are oriented to semantic-based grouping of iconographic objects using semantic descriptors, representing an extension of the descriptive scheme of BIDL iconographical art content.

For example, in SINUS project the subtasks of the art critics' analysis show steps (sub-goals) to be executed. These steps are presented as a formula combining one of the "Bloom's Taxonomy" verbs [Bloom et al., 1956] with a term (concept) from the ontology of the East-Christian iconographical art [Pavlova-Draganova et al., 2007] [Paneva et al., 2007]. In the SINUS learning platform the **Student** "will execute" the Bloom's verb action on the concept(s) from the ontology of the East-Christian iconographical art. For example, in step 1 the **Student** *collects* iconographical objects presenting *Iconographical character* = *Jesus Christ* in a composition type = one-figure. In step 2 the **Student** *classifies* (i.e. arranges the iconographical objects in groups) iconographical objects by a certain iconographic school. In step 3 the **Student** has *to discover-select-show* iconographical objects by a certain author type, etc. Tracking all the sub-goals clearly shows the place of the taxonomy terms of the East-Christian iconographical art ontology needed for the learning analysis.

"Virtual Encyclopedia of the Bulgarian Iconography" Multimedia Digital Library

East-Christian (Orthodox) iconographical art is recognized as one of the most significant areas of painting. Until recently it was neglected in the digital documentation and the registry of this art. But the accessibility of this valuable part of mankind's cultural and historical heritage was enhanced greatly with the appearance of the "Virtual Encyclopedia of the Bulgarian Iconography" multimedia digital library in the global virtual space (see http://bidl.cc.bas.bg/). This Internet-based environment becomes a place where iconographical objects of different kinds and origins were documented, classified, and "exhibited" in order to be widely accessible to both professional researchers and the wide public. Rare specimens, private collections, icons from difficult-to-access storages, distant churches, chapels, and monasteries, objects in a risk environment or unstable conditions, etc., are appearing for new e-exposition. The library provides services for registration, documentation, access and exploration of a practically unlimited number of East-Christian iconographical artifacts and knowledge and the end users could use this rich knowledge base through its interactive preview, objects complex search, selection, and grouping. The first release of the BIDL was developed five years ago within the national project "Digital Libraries with Multimedia Content and its Application in Bulgarian Cultural Heritage" (contract 8/21.07.2005 between the

[1] Research project № D-002-189 with the National Science Foundation of the Ministry of Education and Science. Project executors: consortium of two science organizations – the Institute of information Technologies at the Bulgarian Academy of Sciences and the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences – and one high technology software company – "Active solutions" Ltd.

Institute of Mathematics and Informatics, BAS, and the State Agency for Information Technologies and Communications). As of now, the library has been used in several cross-media, ubiquitous and technology-enhanced learning applications [Paneva-Marinova et al., 2008][Paneva-Marinova et al., 2009][Pavlov et al., 2007].

BIDL Functionality: The key for the current release of BIDL is the efficiency and the provision of strictly designed functionalities. Special attention was paid to content creation, preview, content grouping, content search (semantic, context-based, etc.), metadata management, administrative services, adaptive and personalized access to the content, multilinguality support, etc. presented in [Paneva-Marinova et al., 2010][Pavlov et al., 2010][Pavlova-Draganova et al., 2010]. Moreover, the BIDL semantic content description orders the specification of a unique descriptive scheme for iconographical art content, covering the rich semantic identification and technical features of the iconographical objects [Pavlova-Draganova et al., 2007][Paneva et al., 2007].

BIDL Content: BIDL includes several hundred specimens of Bulgarian iconographical objects from different artists, historical periods, and schools. There are also incorporated information objects, presenting iconographic techniques, authors' biographies, schools' history, terms vocabulary, etc. Several users created and driven collections are shown (for example, the unique collection of pencil-drawings of Zacharya Tsanyuv or the rich collection of icons from "Saint Trinity Church" in Bansko, etc.). The BIDL specimens are in the possession of the Bulgarian Orthodox church and the originals are currently exposed and freely accessible in acting Bulgarian churches and monastery.

The future extensions of the library are related to the content enrichment and the inclusion of wide range of artefacts of the Balkan countries. In BIDL will also be included services for aggregating iconographical content for the European digital library EUROPEANA, thus providing possibilities for pan-European access to rich digitalised collections of East Christian iconographical heritage.

Analyzing and Synthesizing Services in BIDL

An extension of the BIDL functionality is the analyzing, synthesizing and summarizing of content, maintaining content and functionality observation, mining, inference, evaluation and tracking. In BIDL these services are performed through the QlickTech® QlinView® Business Intelligence software[1]. As an analysis services provider, it is connected to the BIDL objects repository and tracking services by a preliminary created data warehouse. The QlickTech® QlinView® Business Intelligence software provides fast, powerful and visual in-memory analysis and synthesis of the data, analytical processing (OLAP), quick answering of multi-dimensional analytical queries, etc. The ETL (Extract, Transform, Load)[2] is a completely automatic process and is performed by administrator request.

[1] Business Intelligence is an architecture and a collection of integrated operational as well as decision-support applications and databases that provide easy access to a large amount of (business) data.

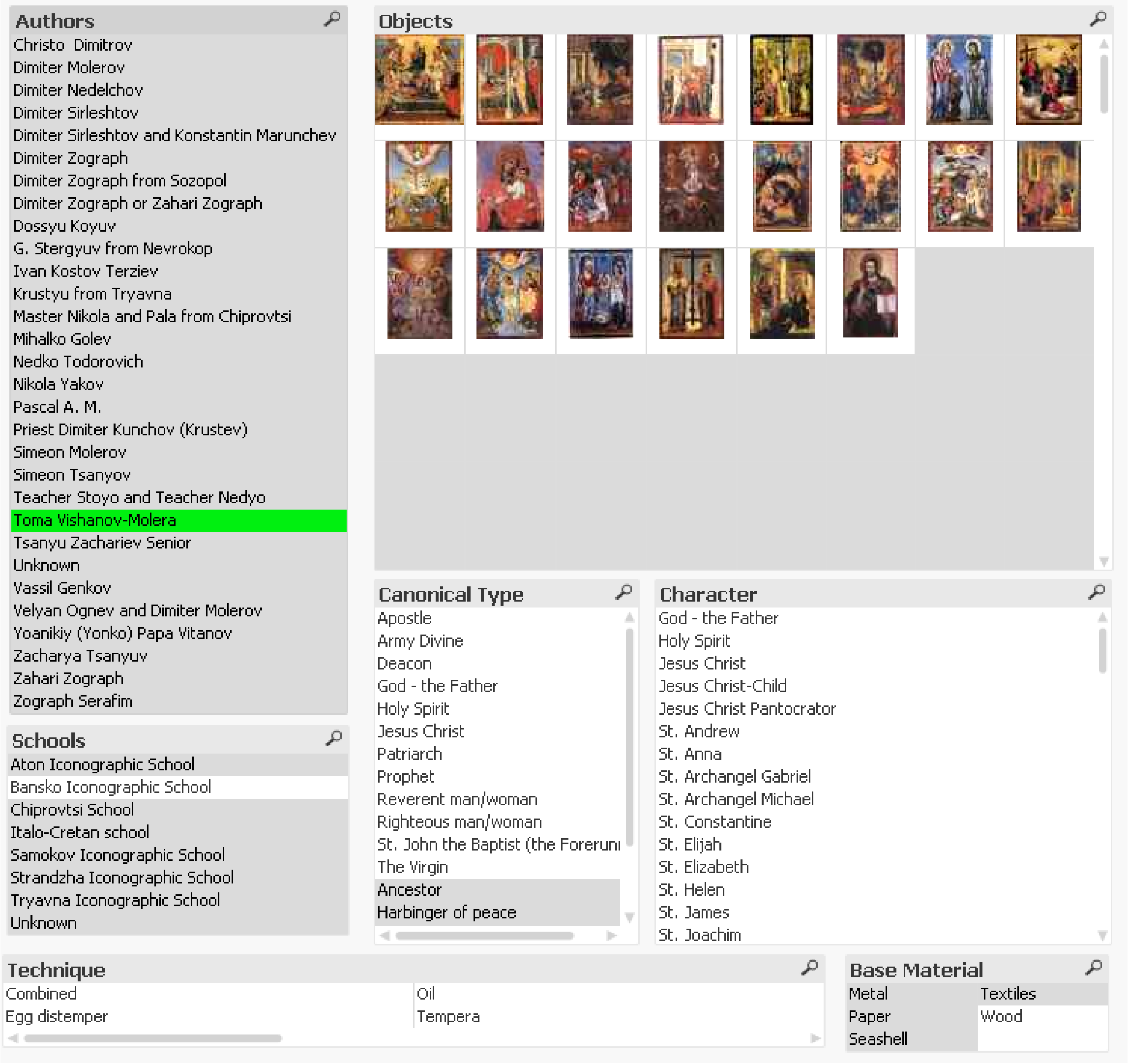
[2] Extract, transform, and load (ETL) is a process in database usage and especially in data warehousing that involves: extracting data from outside sources, transforming it to fit operational needs (which can include quality levels), and loading it into the end target (database or data warehouse).

The variety of generated statistical information about BIDL data extends the available visualization services, enabling the user to analyze the iconography domain as well as the library repository at the most granular level of detail required, providing unparalleled insight into the actual states and data dependencies.

For example, figure 1 depicts the synthesis of the available icons from Toma Vishanov, indicating the author's iconographic school (viz., Bansko Iconographic School), the canonical types and characters painted (i.e., Holy Spirit, Martyr, Jesus Christ, etc.), iconographic techniques and base materials used.

This information snapshot could be used for an analytical research of an author's work, for an art analysis of the emphasis, trends, and areas it covers, the priorities in their work. There is an opportunity to know their art in more detail.

There is another type of diagrams, related to tracing the integrity, status and ratio of the content distribution in the repository of the digital library. Such an example is figure 2 where a PIE diagram is depicted making a canonical sub-types analysis about the Apostle canonical type.



Authors
Christo Dimitrov
Dimiter Molerov
Dimiter Nedelchov
Dimiter Sirleshtov
Dimiter Sirleshtov and Konstantin Marunchev
Dimiter Zograph
Dimiter Zograph from Sozopol
Dimiter Zograph or Zahari Zograph
Dossyu Koyuv
G. Stergyuv from Nevrokop
Ivan Kostov Terziev
Krustyu from Tryavna
Master Nikola and Pala from Chiprovtsi
Mihalko Golev
Nedko Todorovich
Nikola Yakov
Pascal A. M.
Priest Dimiter Kunchov (Krustev)
Simeon Molerov
Simeon Tsanyov
Teacher Stoyo and Teacher Nedyo
Toma Vishanov-Molera
Tsanyu Zachariev Senior
Unknown
Vassil Genkov
Velyan Ognev and Dimiter Molerov
Yoanikiy (Yonko) Papa Vitanov
Zacharya Tsanyuv
Zahari Zograph
Zograph Serafim

Schools
Aton Iconographic School
Bansko Iconographic School
Chiprovtsi School
Italo-Cretan school
Samokov Iconographic School
Strandzha Iconographic School
Tryavna Iconographic School
Unknown

Objects

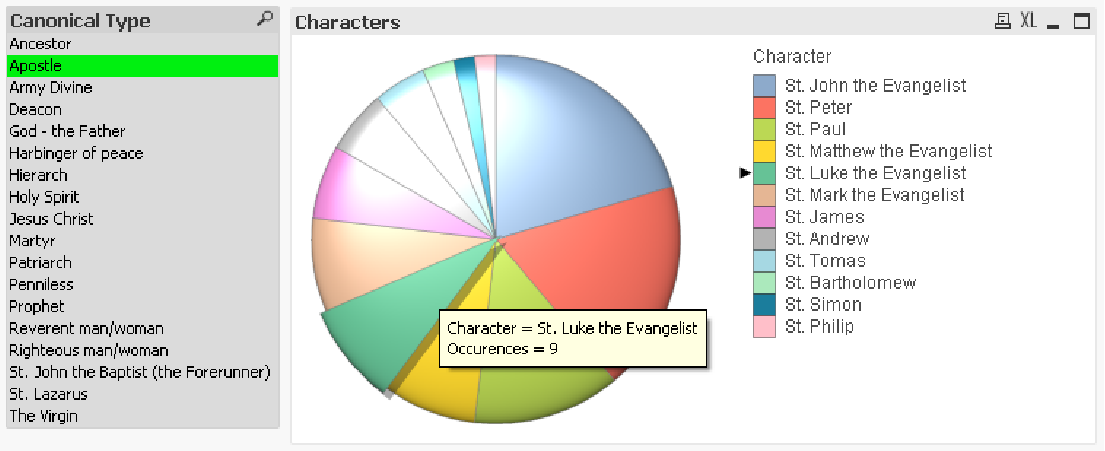
Canonical Type
Apostle
Army Divine
Deacon
God - the Father
Holy Spirit
Jesus Christ
Patriarch
Prophet
Reverent man/woman
Righteous man/woman
St. John the Baptist (the Foreruni
The Virgin
Ancestor
Harbinger of peace

Character
God - the Father
Holy Spirit
Jesus Christ
Jesus Christ-Child
Jesus Christ Pantocrator
St. Andrew
St. Anna
St. Archangel Gabriel
St. Archangel Michael
St. Constantine
St. Elijah
St. Elizabeth
St. Helen
St. James
St. Joachim

Technique
Combined
Egg distemper
Oil
Tempera

Base Material
Metal
Paper
Seashell
Textiles
Wood

Figure 1: An overview of Toma Vishanov-Molera's specimens



Canonical Type

Ancestor
Apostle
Army Divine
Deacon
God - the Father
Harbinger of peace
Hierarch
Holy Spirit
Jesus Christ
Martyr
Patriarch
Penniless
Prophet
Reverent man/woman
Righteous man/woman
St. John the Baptist (the Forerunner)
St. Lazarus
The Virgin

Characters

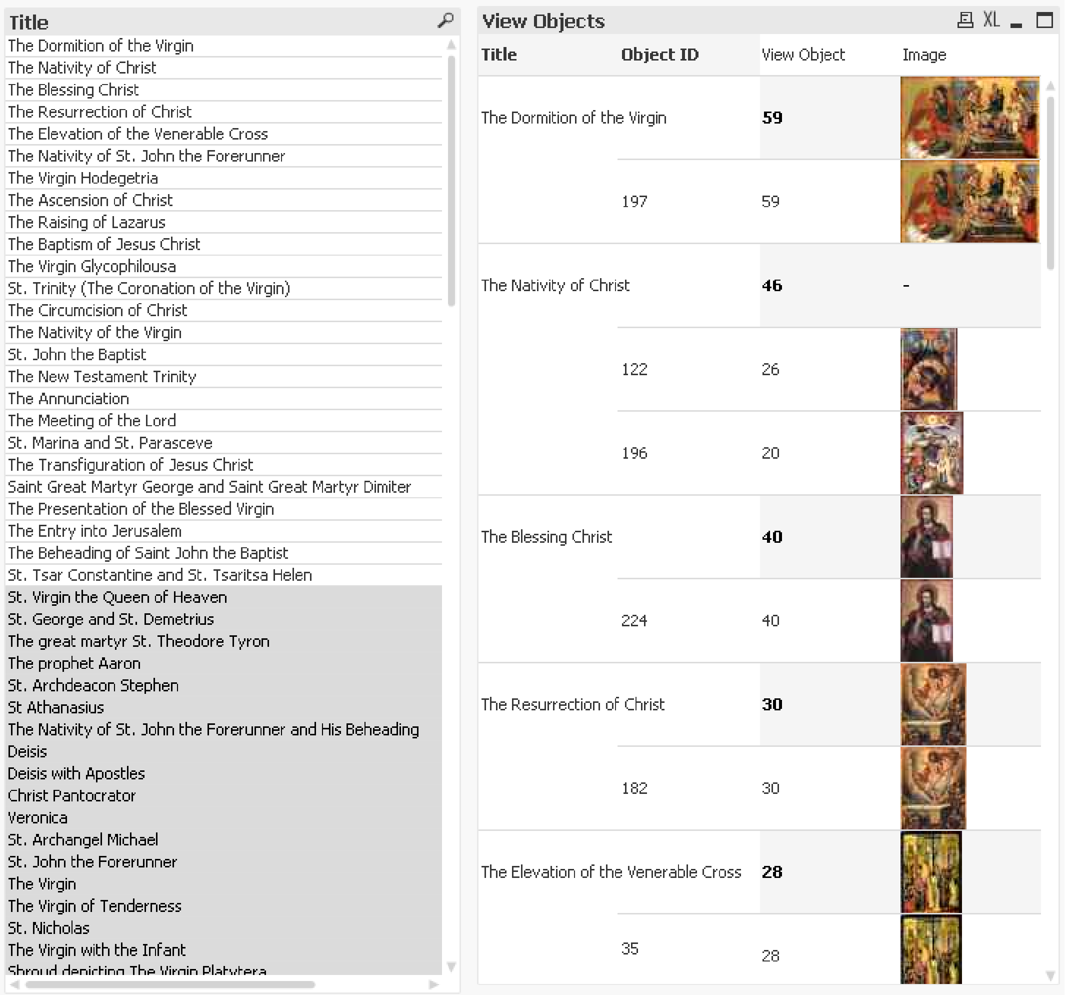
Character = St. Luke the Evangelist
Occurences = 9

Character

St. John the Evangelist
St. Peter
St. Paul
St. Matthew the Evangelist
St. Luke the Evangelist
St. Mark the Evangelist
St. James
St. Andrew
St. Tomas
St. Bartholomew
St. Simon
St. Philip

Figure 2: PIE diagram of canonical sub-types for Apostle canonical type

Figure 3 depicts the frequency of objects' preview, showing the individual objects.



Title

The Dormition of the Virgin
The Nativity of Christ
The Blessing Christ
The Resurrection of Christ
The Elevation of the Venerable Cross
The Nativity of St. John the Forerunner
The Virgin Hodegetria
The Ascension of Christ
The Raising of Lazarus
The Baptism of Jesus Christ
The Virgin Glycophilousa
St. Trinity (The Coronation of the Virgin)
The Circumcision of Christ
The Nativity of the Virgin
St. John the Baptist
The New Testament Trinity
The Annunciation
The Meeting of the Lord
St. Marina and St. Parasceve
The Transfiguration of Jesus Christ
Saint Great Martyr George and Saint Great Martyr Dimiter
The Presentation of the Blessed Virgin
The Entry into Jerusalem
The Beheading of Saint John the Baptist
St. Tsar Constantine and St. Tsaritsa Helen
St. Virgin the Queen of Heaven
St. George and St. Demetrius
The great martyr St. Theodore Tyron
The prophet Aaron
St. Archdeacon Stephen
St Athanasius
The Nativity of St. John the Forerunner and His Beheading
Deisis
Deisis with Apostles
Christ Pantocrator
Veronica
St. Archangel Michael
St. John the Forerunner
The Virgin
The Virgin of Tenderness
St. Nicholas
The Virgin with the Infant
Shroud depicting The Virgin Platytera

View Objects

Title	Object ID	View Object	Image
The Dormition of the Virgin		59	
	197	59	
The Nativity of Christ		46	-
	122	26	
	196	20	
The Blessing Christ		40	
	224	40	
The Resurrection of Christ		30	
	182	30	
The Elevation of the Venerable Cross		28	
	35	28	

Figure 3: Frequency of objects' preview

This information can be used for making conclusions about people's interest in objects, collections and the library content, in order to further fill the repository of the library.

With the QlickTech® QlinView® Business Intelligence software we also perform a paralleled insight of the tracking services' output (BIDL objects tracking and BIDL users' activities tracking). The tracking services "spy on" the activities of add, edit, preview, search, delete, selection, export to XML, and group of MDL objects/collections, user logs, personal data changes, access level changes and user behavior, etc., in order to provide a wide range of statistic data for frequency of service usage, failed requests, etc. for internal usage and generation of inferences about stable work (stability), the flexibility, and the reliability of the environment.

Figure 4 depicts a diagram for the user's activities during a fixed period generated by QlickTech® QlinView® Business Intelligence software.

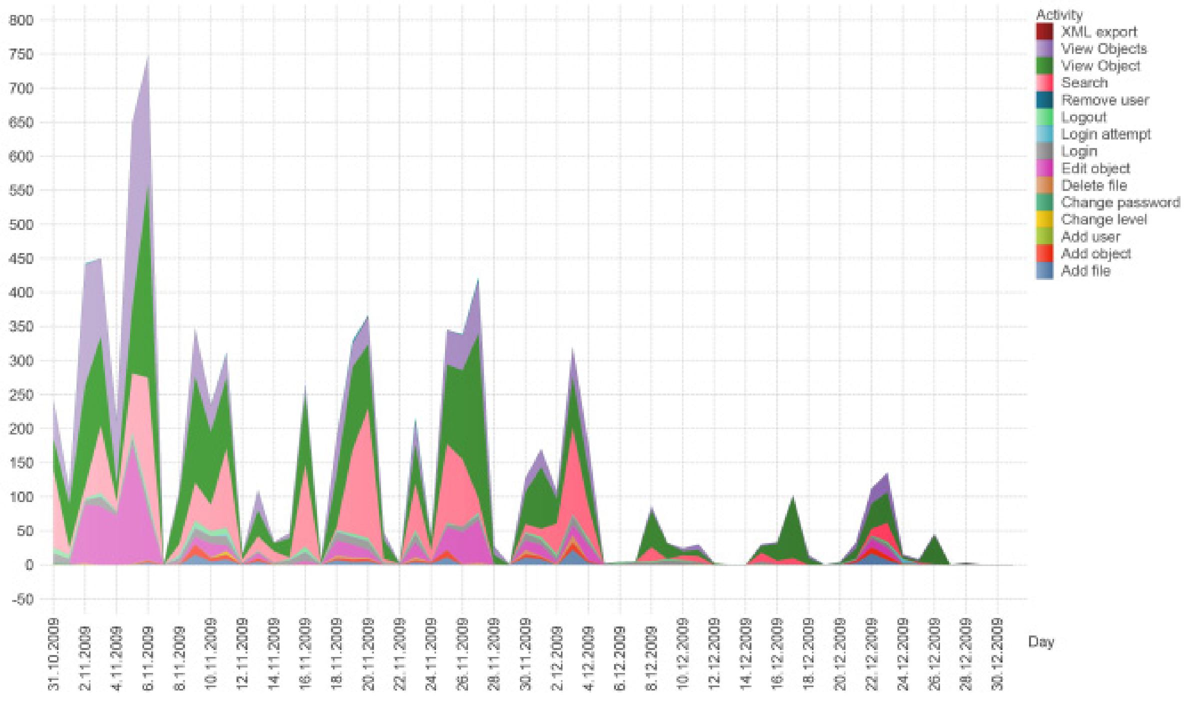
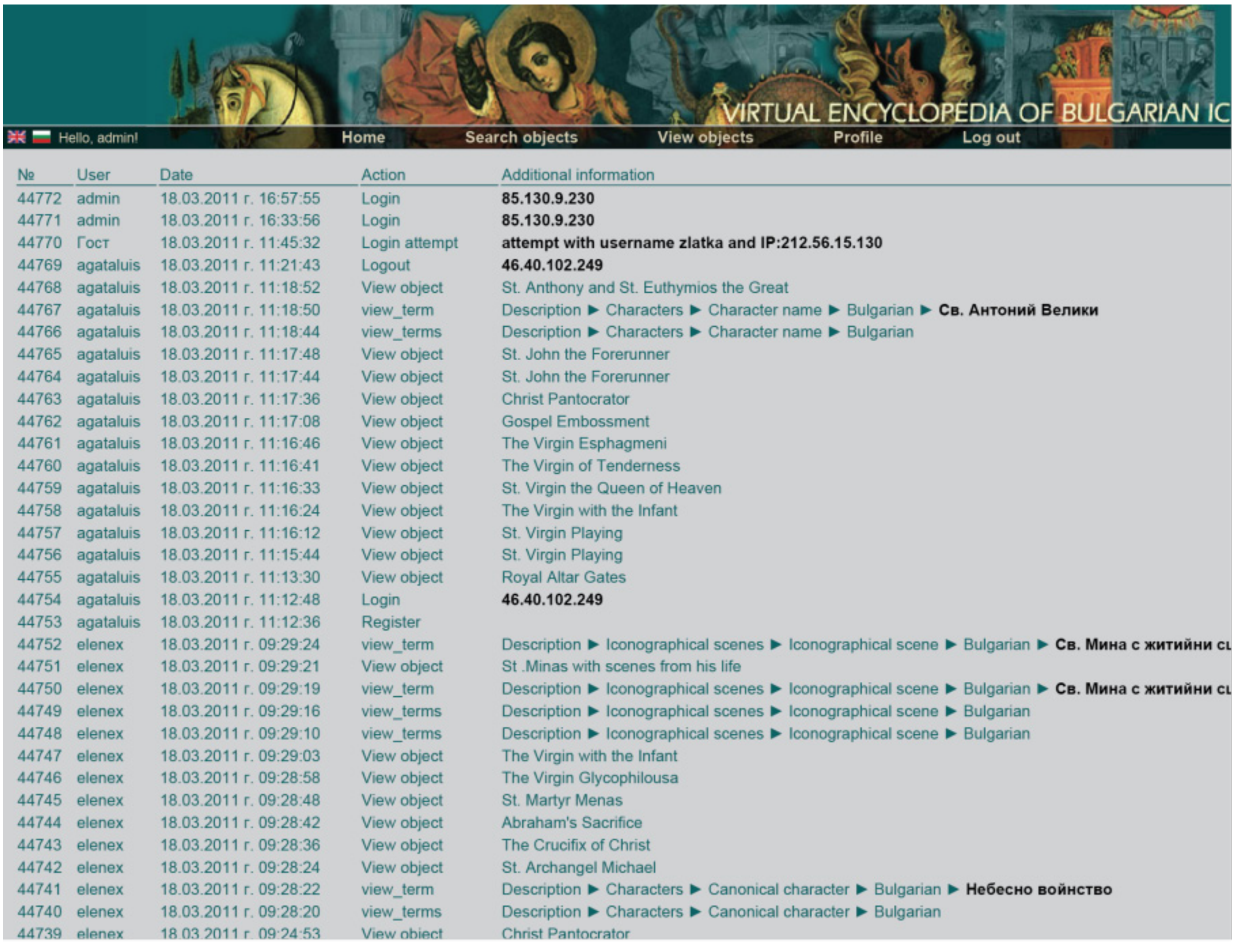


Figure 4: Users' activities during a defined period

In comparison, as depicted in figure 5, the BIDL tracking services return only data result rows, which say what occurred and when. It lacks "high-level" information about services utilization, user engagement, etc., from this data, necessary for real profound analysis.



№	User	Date	Action	Additional information
44772	admin	18.03.2011 г. 16:57:55	Login	**85.130.9.230**
44771	admin	18.03.2011 г. 16:33:56	Login	**85.130.9.230**
44770	Гост	18.03.2011 г. 11:45:32	Login attempt	**attempt with username zlatka and IP:212.56.15.130**
44769	agataluis	18.03.2011 г. 11:21:43	Logout	**46.40.102.249**
44768	agataluis	18.03.2011 г. 11:18:52	View object	St. Anthony and St. Euthymios the Great
44767	agataluis	18.03.2011 г. 11:18:50	view_term	Description ▶ Characters ▶ Character name ▶ Bulgarian ▶ **Св. Антоний Велики**
44766	agataluis	18.03.2011 г. 11:18:44	view_terms	Description ▶ Characters ▶ Character name ▶ Bulgarian
44765	agataluis	18.03.2011 г. 11:17:48	View object	St. John the Forerunner
44764	agataluis	18.03.2011 г. 11:17:44	View object	St. John the Forerunner
44763	agataluis	18.03.2011 г. 11:17:36	View object	Christ Pantocrator
44762	agataluis	18.03.2011 г. 11:17:08	View object	Gospel Embossment
44761	agataluis	18.03.2011 г. 11:16:46	View object	The Virgin Esphagmeni
44760	agataluis	18.03.2011 г. 11:16:41	View object	The Virgin of Tenderness
44759	agataluis	18.03.2011 г. 11:16:33	View object	St. Virgin the Queen of Heaven
44758	agataluis	18.03.2011 г. 11:16:24	View object	The Virgin with the Infant
44757	agataluis	18.03.2011 г. 11:16:12	View object	St. Virgin Playing
44756	agataluis	18.03.2011 г. 11:15:44	View object	St. Virgin Playing
44755	agataluis	18.03.2011 г. 11:13:30	View object	Royal Altar Gates
44754	agataluis	18.03.2011 г. 11:12:48	Login	**46.40.102.249**
44753	agataluis	18.03.2011 г. 11:12:36	Register	
44752	elenex	18.03.2011 г. 09:29:24	view_term	Description ▶ Iconographical scenes ▶ Iconographical scene ▶ Bulgarian ▶ **Св. Мина с житийни сц**
44751	elenex	18.03.2011 г. 09:29:21	View object	St .Minas with scenes from his life
44750	elenex	18.03.2011 г. 09:29:19	view_term	Description ▶ Iconographical scenes ▶ Iconographical scene ▶ Bulgarian ▶ **Св. Мина с житийни сц**
44749	elenex	18.03.2011 г. 09:29:16	view_terms	Description ▶ Iconographical scenes ▶ Iconographical scene ▶ Bulgarian
44748	elenex	18.03.2011 г. 09:29:10	view_terms	Description ▶ Iconographical scenes ▶ Iconographical scene ▶ Bulgarian
44747	elenex	18.03.2011 г. 09:29:03	View object	The Virgin with the Infant
44746	elenex	18.03.2011 г. 09:28:58	View object	The Virgin Glycophilousa
44745	elenex	18.03.2011 г. 09:28:48	View object	St. Martyr Menas
44744	elenex	18.03.2011 г. 09:28:42	View object	Abraham's Sacrifice
44743	elenex	18.03.2011 г. 09:28:36	View object	The Crucifix of Christ
44742	elenex	18.03.2011 г. 09:28:24	View object	St. Archangel Michael
44741	elenex	18.03.2011 г. 09:28:22	view_term	Description ▶ Characters ▶ Canonical character ▶ Bulgarian ▶ **Небесно войнство**
44740	elenex	18.03.2011 г. 09:28:20	view_terms	Description ▶ Characters ▶ Canonical character ▶ Bulgarian
44739	elenex	18.03.2011 г. 09:24:53	View object	Christ Pantocrator

Figure 5: Screen of the BIDL tracking services results

Implementation of the Analyzing and Synthesizing Services in BIDL

The implementation of the analyzing services in BIDL passes over the building of a special logging service, the design of a fast performing data warehouse and the defining of the ETL process.

Building the logging service for the digital library

The initial BIDL database had the structure depicted in figure 6.

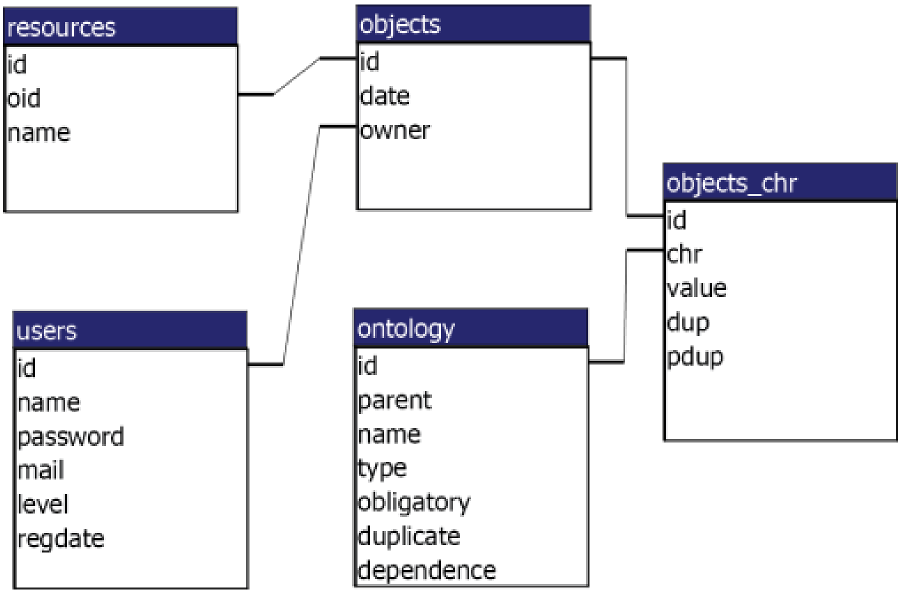


Figure 6: BIDL Database structure

There are five main tables used for storing the user data and the content data. For creating the logging service it was necessary to design another table for that database. The new table has to store all of the user activities which we are interested in. So with the new table we have a database like the one depicted in figure 7.

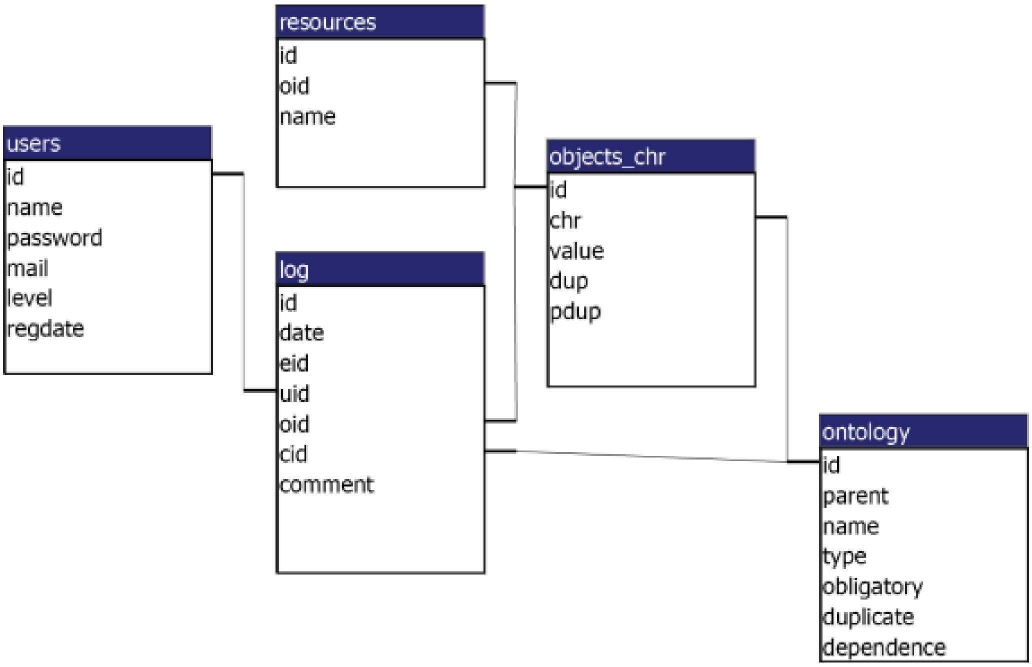


Figure 7: BIDL Database structure Updated

We have added the log table. Each row of this table represents one user activity. Each activity has:

- A unique identification number;
- A timestamp (the exact time of its execution);
- Event identification (eid – specifies the type of activity carried out by the user);
- User ID – unique user identification number (from the users table);
- Object ID – identifier of the object on which an action is performed (if any)
- Characteristic ID – identifies any concrete object characteristic that takes part in the action which the user has performed;
- Comment – provides additional information about the event.

After analyzing our needs for tracking user activities, we decided to track the following types of events (event identification – eid):

- Add file – when a new resource is added;
- Add object – when a new object is created;
- Add user – when an user registers;
- Change password – when an user changes their password;
- Delete file – a file is deleted;
- Delete object – an object is deleted;
- Edit object – an object is edited (modified);
- XML export – XML export of all objects has been performed;
- Group – the group objects service is performed (started, run);
- Login – an user has logged in;
- Login attempt – bad login attempt;
- Logout – user has logged out;
- Remove user – user has been deleted by administrator;
- Search – search action has been executed;
- Change level – user level has been changed;
- View Map – the map service has been executed;
- View Object – view object service;
- View Objects – view a list of objects;
- View Term – view the meaning of a term;
- View Terms – view a list of terms;

These types of events will help us make the various analyses of user behavior in order to improve the quality of our services according to the DL objects' interests.

Data Warehouse Design

To implement the analyzing tool for our DL, we need to design and build a fast performing data warehouse. We choose the snowflake schema for the data warehouse (see figure 8).

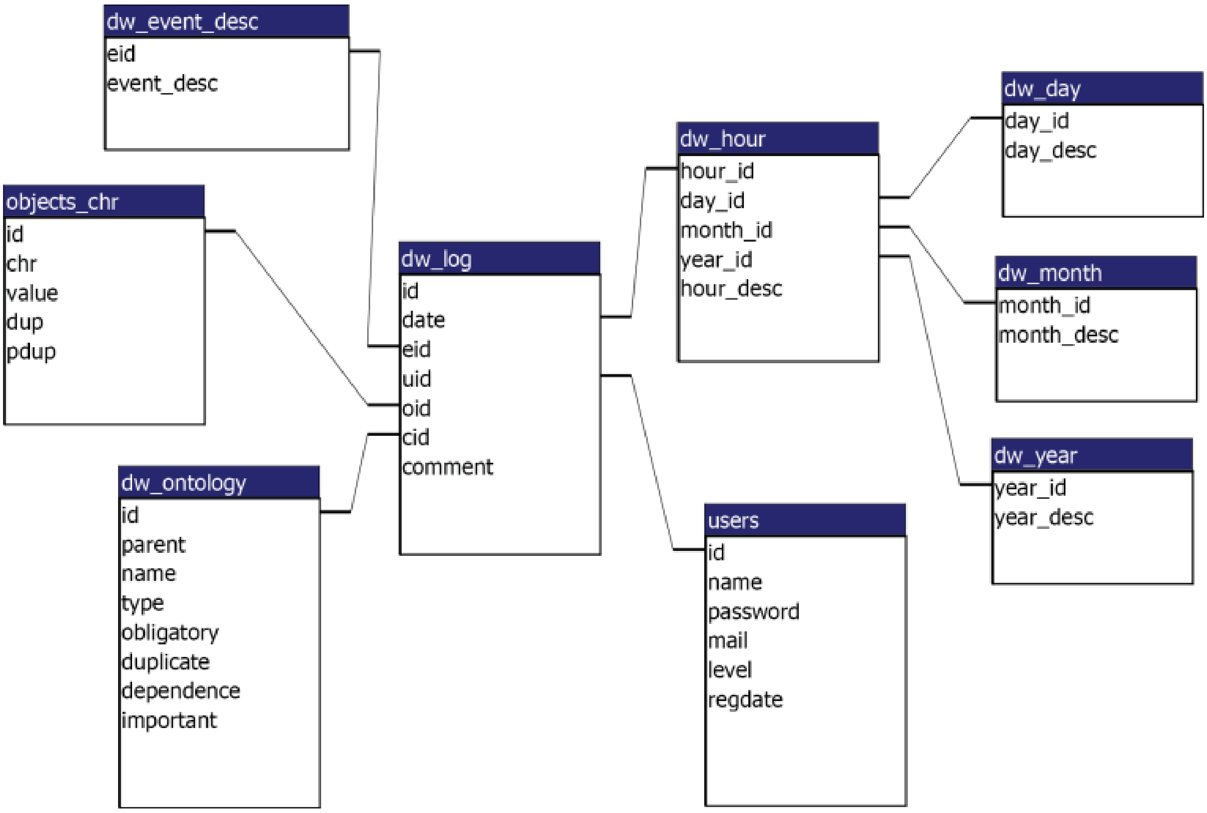


Figure 8: Snowflake schema for the BIDL data warehouse

We use the log table for the fact table of our data warehouse, and the objects, characteristics, and users as dimensional tables. Also, there are additional tables for the time separation.

We have to note that the table dw_log is different from the table log (and dw_ontology differs from ontology), regardless of the fact that they have the same attributes. The reason is that our data warehouse aims at fast performance, and that is because another process is needed before we start creating our analysis.

The ETL process (or how to build and update our data warehouse)

The purpose of this process is to make our data warehouse compatible to our database, so the data transfer from database to data warehouse becomes easy and flawless.

For example, in our case we needed to transfer the timestamp (which contained Year, Month, Day, Hour, Minute, and Second) of the log table to individual entities like: hour, month, year, day for the tables dw_log and dw_hour. We also had to extend the dw_hour table to contain not only values which are connected with one activity, but all values between the times of the first and last activity. This is a common action when building a data warehouse.

So we achieved the desired result for the current example using the PHP language:

```
        while ($begin<$s3)
        {
          $m1 = localtime($begin,1);
          $m2 = localtime($begin-3600,1);
          $hour_id = $begin;
          $day_id = mktime(0, 0, 0, $m1['tm_mon']+1,  $m1['tm_mday'], $m1['tm_year']+1900);
          $month_id = mktime(0, 0, 0, $m1['tm_mon']+1, 1, $m1['tm_year']+1900);
          $year_id = mktime(0, 0, 0, 1, 1, $m1['tm_year']+1900);
          $hour_desc = $m1['tm_hour'];
          $year_desc = $m1['tm_year']+1900;
          $month_desc = ($m1['tm_mon']+1) . ".$year_desc";
          $day_desc = $m1['tm_mday']."."$month_desc";
          msq("REPLACE dw_hour (hour_id, day_id, month_id, year_id, hour_desc) VALUES ($hour_id, $day_id, $month_id,
$year_id, $hour_desc)");
          if ($m1['tm_mon']!=$m2['tm_mon'] OR $first) msq("REPLACE dw_month (month_id, month_desc) VALUES
($month_id, '$month_desc')");
          if ($m1['tm_mday']!=$m2['tm_mday'] OR $first) msq("REPLACE dw_day (day_id, day_desc) VALUES ($day_id,
'$day_desc')");
          if ($m1['tm_year']!=$m2['tm_year'] OR $first) msq("REPLACE dw_year (year_id, year_desc) VALUES ($year_id,
$year_desc)");
          $begin+=3600;
          $first=0;
        }
```

As seen above, we modify data and insert it into the data warehouse.

When the data warehouse is built and the ETL process is defined, we are ready to start creating our analysis through the QlickTech® QlinView® Business Intelligence software.

Acknowledgements

This work is partly funded by Bulgarian NSF under the project D-002-189 SINUS "Semantic Technologies for Web Services and Technology Enhanced Learning".

Bibliography

[Bertino et al., 2001] Bertino, E., Casarosa, V., Crane, G., Croft, B., Del Bimbo, A., Fellner, D., Fiander, P., Fox, E.: Digital Libraries: The Future Directions for European Research Programme. Brainstorming Report (2001)

[Bloom et al., 1956] Bloom B., Krathwohl D., editors. Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I, Cognitive Domain, Longman, New York (1956).

[Carmel et al., 2008] Carmel, D., Yom-Tov, E., Roitman H.: Enhancing Digital Libraries Using Missing Content Analysis. In: Joint Conference on Digital Libraries (JCDL), pp. 1–10. Pittsburgh, PA, USA (2008)

[Chakrabarti et al., 1999] Chakrabarti, S., Van den Berg, M., Dom, B.: Focused crawling: A new approach to topic-specific web resource discovery. Computer Networks, 31(11–16), 1623–1640 (1999)

[Cho et al., 1998] Cho J., Garcia-Molina, H., Page, L.: Efficient Crawling Through URL Ordering. In Proceedings of the 7th World Wide Web Conference, Brisbane, Australia, pp. 161-172. April 1998.

[De Bra et al., 1994] De Bra, P., Houben, G., Kornatzky, Y., Post, R.: Information Retrieval in Distributed Hypertexts. In the Proceedings of the 4th RIAO (Computer-Assisted Information Retrieval) Conference, pp. 481-491 (1994).

[Pahal et al., 2007] Pahal, N., Chauhan, N., Sharma, A.K.: Context-Ontology Driven Focused Crawling of Web Documents, In the Proceedings of Third International Conference on Wireless Communication and Sensor Networks, WCSN '07, p. 121-124 (2007).

[Paneva et al., 2007] Paneva, D., Pavlova-Draganova, L., Draganov, L.: Towards Content-sensitive Access to the Artefacts of the Bulgarian Iconography. In: 5th International Conference on Information Research and Applications (i.Tech 2007), vol. 1, pp. 33–38. Varna, Bulgaria (2007)

[Paneva-Marinova et al., 2008] Paneva-Marinova, D., Pavlova-Draganova, L., Pavlov, R., Sendova M.: Cross-media and Ubiquitous Learning Applications on Top of Iconographic Digital Library. In: 14th International Conference on Virtual Systems and Multimedia, pp. 367–371. ARCHAEOLINGUA, Limassol, Cyprus (2008)

[Paneva-Marinova et al., 2009] Paneva-Marinova, D., Pavlova-Draganova, L., Draganov, L., Pavlov, R., Sendova M.: Development of a Courseware on Bulgarian Iconography for Ubiquitous On-demand Study. In: Open Conference “New Technology Platforms for Learning – Revisited”, pp. 37–46. Budapest, Hungary (2009)

[Paneva-Marinova et al., 2010] Paneva-Marinova, D., Pavlov, R., Goynov, M., Pavlova-Draganova, L., Draganov L.: Search and Administrative Services in Iconographical Digital Library, In: International Conference on Information Research and Applications (i.Tech 2010), pp. 177–187. Varna, Bulgaria (2010)

[Pant et al., 2004] Pant, G., Tsioutsiouliklis, K., Johnson, J., Giles, C. L.: Panorama: extending digital libraries with topical crawlers. In JCDL ’04: Proceedings of the 4th ACM/IEEE–CS joint conference on Digital libraries, pp. 142–150. ACM Press (2004)

[Pavlov et al., 2007] Pavlov, R., Paneva, D.: Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content. Cybernetics and Information Technologies, 6(3), 51–62 (2007)

[Pavlov et al., 2010] Pavlov, P., Paneva-Marinova, D., Goynov, M., Pavlova-Draganova L.: Services for Multimedia Resource Annotation and Presentation in Iconographical Digital Library. Serdica Journal of Computing, 4(1), 101–114 (2010)

[Pavlova-Draganova et al., 2007] Pavlova-Draganova, L., Paneva, D., Draganov L.: Knowledge Technologies for Description of the Semantics of the Bulgarian Iconographical Artefacts, In: Open Workshop on Knowledge Technologies and Applications, pp. 41–46. Kosice, Slovakia (2007)

[Pavlova-Draganova et al., 2009] Pavlova-Draganova, L., Paneva-Marinova, D., Draganov, L.: A Use Case Scenario for Technology-enhanced Learning through Semantic Web Services. Information Technologies & Knowledge, 3(3), 257–268 (2009)

[Pavlova-Draganova et al., 2010] Pavlova-Draganova, L., Paneva-Marinova, D., Pavlov, R., Goynov M.: On the Wider Accessibility of the Valuable Phenomena of Orthodox Iconography through Digital Library. In: International Conference dedicated on Digital Heritage (EuroMed 2010), pp. 173–178. ARCHAEOLINGUA, Lymassol, Cyprus (2010)

[Su et al., 2005] Chang Su, Yang Gao, Jianmei Yang, Bin Luo: An Efficient Adaptive Focused Crawler Based on Ontology Learning", In the Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05), pp.73-78 (2005).

[Zhuang et al., 2005] Zhuang, Z., Wagle, R., Giles, C. L.: What's there and what's not?: focused crawling for missing documents in digital libraries. In JCDL '05: Proceedings of the 5th ACM/IEEE–CS joint conference on Digital libraries, pp. 301–310. ACM Press (2005)

Authors' Information

Desislava Paneva-Marinova *– PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: dessi@cc.bas.bg*

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.

Radoslav Pavlov *– PhD in Mathematics, Associated Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: radko@cc.bas.bg*

Major Fields of Scientific Research: Multimedia and Language Technologies, Digital Libraries, Information Society Technologies, e-Learning, Theoretical Computer Science, Computational Linguistics, Algorithmic, Artificial Intelligence and Knowledge Technologies.

Maxim Goynov *– Programmer, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: maxfm@abv.bg*

Major Fields of Scientific Research: Multimedia Digital Libraries and Applications.