DAR: A Digital Assets Repository for Library Collections

Iman Saleh[1], Noha Adly [1,2], Magdy Nagi [1,2]

[1] Bibliotheca Alexandrina, El Shatby 21526,
Alexandria, Egypt
{iman.saleh, noha.adly, magdy.nagi}@bibalex.org
[2] Computer and Systems Engineering Department, Alexandria University,
Alexandria, Egypt

Abstract. The Digital Assets Repository (DAR) is a system developed at the Bibliotheca Alexandrina, the Library of Alexandria, to create and maintain the digital library collections. The system introduces a data model capable of associating the metadata of different types of resources with the content such that searching and retrieval can be done efficiently. The system automates the digitization process of library collections as well as the preservation and archiving of the digitized output and provides public access to the collection through browsing and searching capabilities. The goal of this project is building a digital resources repository by supporting the creation, use, and preservation of varieties of digital resources as well as the development of management tools. These tools help the library to preserve, manage and share digital assets. The system is based on evolving standards for easy integration with web-based interoperable digital libraries.

1 Introduction

The advent of digital technology and high speed networks are leading to widespread changes in services offered by libraries. The heightened user expectations combined with the growth of collections based on digital content makes it increasingly important for all libraries to find efficient tools to manage their digital contents and enable instant access to their digital assets. The Digital Assets Repository (DAR) of Bibliotheca Alexandrina (BA) acts as a repository for all types of digital material and provides public access to the digitized collections through web-based search and browsing facilities. DAR is also concerned with the digitization of material already available in the library or acquired from other research-related institutions. A digitization laboratory was built for this purpose at the Bibliotheca Alexandrina. DAR is built for a library institution; therefore the system adopted a data model able to describe digital objects that include books as well as images and multimedia. Another major objective of DAR is the automation of the digitization workflow and its integration with the repository.

The following goals were driving us while designing and implementing DAR:

- Integrating the actual content and metadata of varieties of objects types included in different library catalogs into one homogeneous repository.

- The automation of the digitization process such that human intervention is minimized and the outputs are integrated within the repository system.
- The preservation and archiving of digital media produced by the digital lab or acquired by the library in digital format.
- Enhancing the interoperability and seamless access to the library digital assets.

The rest of this paper is organized as follows. Section 2 presents some of the related work. Section 3 gives an overview of the system architecture. Sections 4 and 5 present the two main modules; the Digital Assets Keeper and the Digital Assets Factory, respectively. Section 6 presents the tools provided by the system. Section 7 concludes the paper and presents proposed directions for future work.

2 Related Work

There is an increasing number of digital solutions motivated by the increase in the need of preserving and maintaining digital assets.

EPrints [1,2] is a digital repository for educational material that allows authors self archiving their work. A registered user can submit a document to the EPrint archive, the document is described using a super-set of the BibTeX fields. A submitted document is indexed for searching and positioned within a subject hierarchy defined in the system. Dspace [3] is another repository system for handling educational material and depends mainly on Dublin Core records to describe an item. The system defines a workflow for the submission and supports searching, browsing and retrieval. Both EPrints and Dspace implement the OAI-PMH protocol [4]. Greenstone [5,6] is an open source software that provides out-of-the-box solution for the creation and publishing digital material. The system provides easy-to-use interface to define collections of digital objects, the metadata used to describe items within the collection and how items are displayed. According to these configurations, new collections are created and indexes are built for browsing and searching. Greenstone supports different document formats such as HTML, PDF, DJVU and Microsoft Word files. OpenDlib [7] proposes a similar system that aims at providing expandable and searchable system through customizable services.

Commercial library solutions and document management software are used by some libraries and institutions to manage their digital assets. However, most of these systems fail to address interoperability, extendibility and integration with other tools and services in the library due to their proprietary nature.

Contrary to other systems that only manage digital objects or are dedicated to educational material, DAR incorporates in one repository all types of material and formats belonging to the library collections, either born digital or digitized through the system. The DAR data model is capable of describing different metadata sets required by the heterogeneous nature of the collection while still complying with existing and evolving standards. Also, DAR integrates the digitization and OCR process with the digital repository and introduces as much automation as possible to minimize the human intervention in the process. As far as we know, this is an exclusive feature of DAR.

3 System Architecture

The architecture of DAR is depicted in Figure 1.The system core consists of two fundamental modules:

- The Digital Assets Factory (DAF) which is responsible for the automation of the digitization workflow, and
- The Digital Assets Keeper (DAK) which acts as a repository for digital assets.

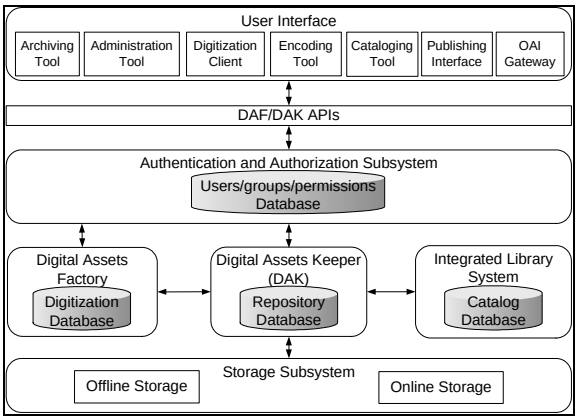

Fig. 1. Architecture of DAR

Both systems interact with the digital objects storage system. The storage system is used to store digital files either for online access and publishing purposes, or offline for long-term preservation. The system contains a set of user interfaces that interact with the system components through APIs. The user interfaces provide tools for the automation of the digitization process, the system parameterization, metadata entry, searching and browsing the repository content, and tools for the interoperability with other repositories. An authentication and authorization system controls the access to the repository contents and functionalities based on the user identity. The repository is integrated with the Integrated Library System (ILS). Plug-in modules control the metadata exchange between the repository database and the ILS database.

The system is implemented in C# using the Microsoft .Net technology. The web-based components are implemented as ASPX pages running on Microsoft IIS web server. The repository APIs are implemented as Web services. SQL sever database is used as the main repository database. The repository is integrated with the Virtua ILS [8] which uses Oracle database on UNIX platform.

4 Digital Assets Keeper - DAK

The DAK acts as a repository for digital material either produced by the digital lab or introduced directly in a digital format. All metadata related to a digital object is stored in the DAK repository database.

4.1 Data Model

One of the challenges faced by DAR is to derive a data model capable of describing all types of library assets including books, maps, slides, posters, videos and sound recordings. For this purpose, two existing standard for data representation have been studied, namely MARC 21 [9] and VRA Core Categories [10]. While the MARC standard is widely used as a data interchange standard for bibliographic data, it is designed mainly for textual materials. Therefore, MARC is seen by the visual resources community as overly elaborate and complex in ways that provide no benefit to visual resources collections, while at the same time lacking or obscuring some concepts which are important to them. On the other hand, VRA core is designed specifically for works of art and architecture that the library is likely to include in its multimedia collections. One of the imaging systems based on the VRA is the Luna Insight [11] which is a commercial imaging software that is widely used by many libraries, universities and museums as a repository for visual assets. The advantages and usefulness of the VRA are discussed in [12]. The data model used by DAR is inspired by the one proposed by the VRA. However, the VRA categories have been extended to accommodate for bibliographic data supported by the MARC standard. This resulted in a data model capable of describing both visual and textual materials in one homogeneous model that is, at the same time, compliant with both standards. The data model is depicted in Figure 2.

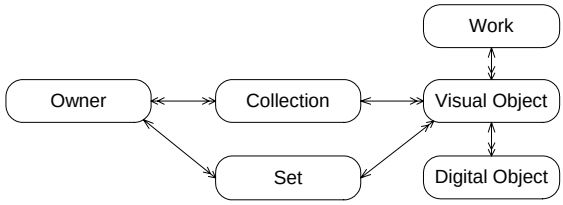

Fig. 2. DAR Basic Data Model

DAR represents a digital object by a *Work* entity related to one or more *Visual Object* entities. This is inspired by the VRA *Work* and *Image* entities. The *Work* refers to a physical entity; it might be a performance, composition, literary work, sculpture, event, or even a building, while the *Visual Object* refers to a visual representation of a *Work*. It can exist in photomechanical, photographic and digital formats. In a typical visual resources collection, a *Visual Object* is a reproduction of the *Work* that is owned by the cataloging institution and is typically a slide, photograph, or digital file. A *Visual Object* exists in one or more digital forms denoted as *Digital Objects*. A *Digital Object* might be a JPG file presenting a scanned slide, an Image-On-Text PDF for an OCR-ed book or an audio or video file.

A *Visual Object* has one *Owner*. The *Owner* is typically an institution, a department or a person. *Visual Objects* related to an *Owner* are grouped into Sets. The *Set* represents a physical grouping of *Visual Objects*; this grouping is established at the digitization phase. On the other hand, the *Collection* represents a descriptive grouping of *Visual Objects* based on a common criteria. Table 1 shows examples of values for each of the described objects.

Table 1. Example of Digital Objects

Object	**Value**
Collection	Million Book Project, OACIS Collection
Set	Box of 100 slides donated to the library
Owner	Bibliotheca Alexandrina, Yale University
Work	The building of Bibliotheca Alexandrina, The new year concert
Visual Object	A slide of the library building, a video taken in a concert
Digital Object	A JPG file of a scanned a slide, a PDF file for an OCR-ed book

4.2 Metadata

Within the DAR data model, the system holds six categories of metadata describing assets and its digital reproductions, a demonstrative example can be found in [13]:

4.2.1 Descriptive Metadata

This includes metadata common to all types of *Works* and *Visual Objects*, such as:

- Type, for a *Work* object, the type could be a painting, map, event or building. For a *Visual Object*, the type could be a slide, photograph, video, audio or book
- Title
- Creator(s), a creator could be the author, publisher, architect or artist
- Date(s), a date could represent the date of the creation, alteration or restoration

This is as well as keywords, description, dimensions, location, etc. Other metadata that is specific to a *Work* type include fields like the ISBN, language and publisher in the case of books, the technique and material in the case of a work of art.

4.2.2 Digital Content Metadata

This includes metadata describing a Digital Object. DAR supports a variety of digital objects' formats including JPG, TIFF, JPG 2000, PDF, DJVU, OCR Text and others. Metadata such as image resolution, dimensions, profile, or a video duration are extracted from digital files automatically and stored in DAK. New formats can be introduced into the system and appropriate tools can be integrated to deal with the new file formats.

4.2.3 Archiving Metadata

This includes metadata about the archiving location of a Digital Object file. The archiving metadata consists of the archiving media unique identifier. The archiving metadata can also be attached to the Visual Object, denoting the physical location where the object can be found in the owning institution.

4.2.4 Publishing Metadata

Encoded objects for publishing are stored on online storage. The publishing metadata includes the path of the published Digital Object on the server, the date of publishing, duration of publishing as well as the category of targeted users e.g. students, researchers, etc.

4.2.5 Access Right Metadata

Copyright restrictions on the repository contents are forced by defining access rights attached to each object. This consists of a copyright statement linked to the Visual Object. Also, an access right level is used by the system to indicate whether a Visual Object and its related Digital Objects are free of copyright restrictions or not. This level is used by the publishing interface to determine the display of objects; whether to display metadata only, the full digital objects or only part of it.

4.2.6 Authentication and Authorization Metadata

DAR users are identified by a username and a password. Further, user groups are defined where a user can belong to one or more groups. Permissions are given to each user or group, which are checked before accessing an application and/or digital object. User and group rights can be specified on the Visual Object level or, more practically, on the Collection level.

5 Digital Assets Factory – DAF

The DAF governs the digitization process of the library collection at the digital lab. DAF realizes one of the main goals of DAR which is the automation of the digitization process. This supports the digitization of library assets including textual material, slides, maps and others. It provides the digital lab operators with tools for entering a digitization job metadata, keeping track of digitization status, applying validation tests on digitized material, recording productions, archiving the digitized material for long term preservation and retrieving the archived material when needed. The system supports different workflows for different types of material.

After initiating a new job, the asset passes by the general phases depicted in Figure 3:

- Scanning the material.
- Processing the scanned files to enhance the quality.
- Perform Optical Character Recognition (OCR) on the textual material.
- Encoding the digitized material by generating a version suitable for publishing.
- Archiving the output of each step of the digitization. Two offline backups are taken for a file, one on CD and the other on tape. Encoded versions are moved to online storage for publishing.

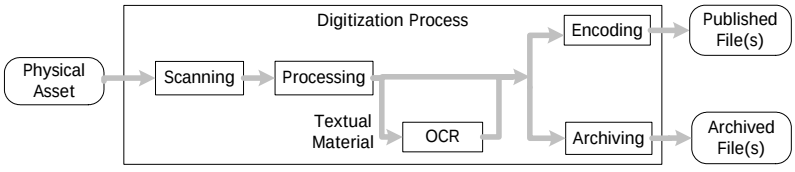

Fig. 3. Digitization Phases

The files and folders produced by each phase are stored in separate queues on a central storage server. A job folder resides in one of the four main queues: *scanned*, *processed*, *OCRed* and *ready for archiving* queue. The digital lab operator withdraws

jobs from the queues, performs the job and places the output in the next queue in the process. Alternatively, files can be introduced directly into any of the queues, for example an image that is already provided in digital form is placed directly into the *processing queue*. Table 2 shows a one year digitization statistics at the BA digital lab since the deployment of DAF in March 2004.

Table 2. Digital Lab Production Statistics

	Arabic	**Latin**
Scanned pages	4,591,463	730,141
Processed pages	4,585,833	730,141
OCRed pages	1,148,465	693,978
Scanned Slides	12,013	
Archived Data on CDs/Tapes	480 GB	

The main goals of DAF are:

- To provide a database system to keep track of the digitization process through the scanning, processing, OCR-ing, archiving and publishing.
- To keep track of digitized materials; unifying the naming conventions and exhaustively checking the produced folders and files for consistency.
- To provide timely reports to various levels of management describing the workflow on a daily, weekly or longer basis and to allow online queries about the current status of a certain asset at the digital lab.
- To apply necessary encodings on the scanned materials to be suitable for electronic publishing.
- To manage the archiving and retrieval of the digitized material.

Digitization Metadata

For objects that are digitized using the DAF applications, the digitization metadata is gathered during the different digitization stages, such as the scanning date(s) and scanning operator(s), the processing date(s) and processing operator(s), the OCR font data, the accuracy achieved by the OCR before and after learning, etc...

6 TOOLS

The DAR system deals with three types of users; digitization operators, librarians - which are divided into catalogers and reviewers - and the end users. Each type of users is provided with tools to make use of the system functionalities.

6.1 Administration Tool

The *Administration Tool* is one of the DAF Web-based tools used by the operator in the digital lab. The tool is used to initiate a new job by entering minimal descriptive metadata for the material to be digitized. If the material is cataloged in the library catalog, the ILS id - a book barcode, for example - is used to retrieve the metadata

from the library catalog. This id is also used to link the record in DAR to the one in the library catalog for future synchronization. If the material is not previously cataloged, the operator enters the minimal metadata that can be deduced from the physical item in hand. The tool uses this metadata to derive a unique folder name for the scanned files. The tool is also used for the system parameterization and to generate reports on production rates and jobs in different digitization queues in the lab.

6.2 Digitization Client – DLClient

The *DLClient* is a DAF application used by the operator in the digital lab. The tool creates structured folders for new digitization jobs and after the completion of each digitization phase, the *DLClient* tool is used to perform the following:

- Validate the files; generate warnings if any inconsistencies are detected.
- Update the job status in the database by setting the username for the operator who performed the job, the job completion date and the count of produced files.
- Move the folders and files to the queue of the next digitization phase on a storage server. Before moving any folder, a lock is acquired on the folder and sub files to avoid concurrent access to the folder while moving.

The *DLClient* is used by the operator through the three main digitization phases; scanning, processing and OCR.

Scanning

Physical assets submitted to the lab for digitization are placed in a *scanning queue*. The operator retrieves a job from the queue and uses the *DLClient* to create the folder structure where scanned files are to be stored. Mainly, a digitization folder contains three subfolders for three types of files: the original scanned files, the processed files and the encoded output. The encoded output, the folder structure and the scanning resolution varies according to the material type; text, image, audio or video. When the scanning is done, the *DLClient* places the produced files in the *processing queue*.

Processing

The operators use the *DLClient* to retrieve a job from the *processing queue*. A combination of manual and automated image processing tools is used to enhance the quality of the scanned images. After the job is done, the *DLClient* places the job at the *OCR queue* for textual material and directly to the *archiving queue* for other types of material.

Optical Character Recognition

Using the *DLClient*, a processed textual material is retrieved from the *processing queue* to be OCR-ed extracting text from the scanned images. OCR is used to enable full text searching. Currently, the system supports Latin OCR using Fine Reader 6.0 from ABBYY [14] and Arabic OCR using Sakhr Automatic Reader [15]. To enhance the recognition quality of Arabic text, BA has built a library of fonts using learning samples taken from different books. Before starting the recognition, the OCR operator matches the book font with the nearest font library.

Reprocessing

The system supports a special workflow for reprocessing a digitized material. Reprocessing may be needed to enhance the OCR quality, to apply new image processing procedure or simply to generate new publishing format of the digitized material.

Reprocessing begins by searching and retrieving the files to be reprocessed from the archive. The files are then placed in the appropriate digitization queue. The reprocessed files go through the normal digitization steps described before until they reach the archiving phase. Only altered files are re-archived, changes in files are detected using checksums that are calculated before and after the reprocessing. The archiving information of a new file version is recorded in the repository database and a link is made to the parent version archiving location so that file versions may be tracked in the database from the most recent to the base version.

6.3 Archiving Tool

In the current version, a digital object is represented by one or more files with different formats and/or resolutions, these files are stored for online access on RAID storage system or on offline storage for long-term preservation. Typically, the preserved material is the scanned originals and the processed version with high resolution. Lower versions derived for publishing purposes are saved on online storage for ease of access; this includes low resolution JPG, PDF, and DJVU. Files stored offline are archived on two medias; CDs and tapes. Unique labels are generated, printed and attached to the media for future retrieval. The system keeps track of different versions of a file by linking a newer version to its older one. More sophisticated content versioning and object representation is to be applied in future versions of DAR, this could build on the architecture proposed by the Fedora repository [16].

The *Archiving Tool* is one of the DAF Windows-based applications used by the lab operators and offers the following functionalities:

- Checking files and folders consistency.
- Preparing the folders for archiving by compressing the subfolders and files, grouping them into bundles that fit into the media capacity (CD or tape), generating the media label, printing the label.
- The tool generates checksums for the archived files to detect changes in case of downloading and reprocessing a file.
- A search facility enables the user to retrieve an archived folder by locating the folder, uncompressing the subfolders and files and copying the uncompressed files and folders to a destination specified by the user.
- Managing the space on the storage server hard drives, the tool generates warnings when storage level exceeds a predefined value for each drive.
- The tool updates the DAK database by recording the archiving information related to a digital file.

6.4 Encoding Tool

In the encoding step, a final product is generated for publishing. For images, slides and maps, different JPG resolutions are generated. For audio and video, different qualities are generated to accommodate for different network connections' speed. For textual material like books, special developed tools are used to generate the image-on-

text equivalent of the text; this is done on an encoding server built on Linux platform. The Encoding Server encodes digital books into light-weight image-on-text documents in DjVu and PDF. Support for DjVu is built around DjVu Libre, an open source implementation of a DjVu environment, or, alternatively, Document Express, LizardTech's commercial DjVu product. Support for PDF is implemented based on iText, an open source API for composing and manipulating PDF documents. The Encoding Server supports multilingual content through integration with Sakhr Automatic Reader [13]. The Encoding Server allows for the integration of any OCR engine through writing OCR converters, which transforms the native OCR format into a common OCR format that the Encoding Server is capable of processing along with page images in TIFF or JFIF format to compose image-on-text documents. A generated file is copied to a publishing server, the encoding tool updates the DAK database by inserting the corresponding *Digital Object* record. The record is populated with metadata extracted from the digital files and with the publishing information; publishing server and URL.

6.5 Cataloging Tool

The *Cataloging Tool* is a Web-based application used by the librarian to add and edit metadata in the DAK subsystem. Using the Cataloging Tool, the librarian enriches the digital repository records – created in the digitization phase - by adding metadata. The librarian can also create new records for digital objects and upload their corresponding files. The repository is preloaded with controlled vocabularies lists. The tool allows defining configurable templates, importing metadata from external sources and automatic extraction of digital content metadata.

6.6 Publishing Interface

The *Publishing Interface* is a Web-based interface related to the DAK that provides access to the repository of digital objects through search and browsing facilities.
The repository Publishing Interface offers the following functions:

- Browse the repository contents by *Collection*, *Work* Type, *Visual Object* Type, Subject, Creator and Title.
- Search the content by an indexed metadata field; Creator, Title, Subject, …
- For textual material, a search in the full text can be conducted. The user can choose whether exact or morphological matching is applied.
- For images, different levels of zooming are available.
- Display brief and full record information with links to the digital objects.
- Display the records in MARC or DC in XML formats.
- Hyperlinked data fields that can invoke searches e.g. by Keywords, Subjects and Creator.

6.7 Integration with the ILS

DAR can be easily integrated and synchronized with external sources – e.g. bibliographic catalog, external repository, imaging systems - by implementing appropriate integration modules. An integration module is a plug-in component designed to export records from DAR to an external repository, or to import records from an external repository into DAR, or both.
The integration module is fully configured based on the following:

1. A record unique identification: This identifier is used as a link between the record in DAR and the one in the external repository.
2. Metadata mapping table: The mapping table defines how data fields are mapped from DAR to the external repository and vice versa including the concept of *Work* and *Visual Object*.
3. Synchronization schedule: This schedule defines how often the two repositories are synchronized. The synchronization process considers only the newly created and modified records.

In the current version, a module is implemented for integration with the Virtua ILS [8] which is deployed at BA.

6.8 Authentication and Authorization

In DAR, a *User* is a member of one or more *Group*. Each *Group* is assigned access permissions on the repository contents and functionalities. A basic username and password scheme is used to identify the user. Anonymous access to the repository is also allowed, the access right of an anonymous user are defined by the permissions assigned to a special group *Guest*. This simple authorization scheme will be augmented in future versions to accommodate for special groups' permissions.

6.9 OAI Gateway

DAR OAI Gateway implements the Open Archive Protocol for Metadata Harvesting developed by the Open Archives Initiative [4] to provide access to the repository contents across an organization's architecture. The Gateway receives XML requests and translates them to the equivalent database queries. When the request result sets are retrieved, the gateway translates them into XML and responds to the requesting application.
The gateway implements the six types of requests required for OAI-PMH compliance: Identify - ListMetadataFormats - ListSets - ListIdentifiers - ListRecords - GetRecord

7 CONCLUSIONS AND FUTURE WORK

We have presented in this paper the DAR system implemented at the Bibliotheca Alexandrina. The system acts as a repository for digital assets owned by the library and associates the metadata with the content to provide efficient search and retrieval.

DAR addresses the main challenges faced by digital repositories including supporting different digital formats, digitization workflows, preservation of digital material and content dissemination. The DAF subsystem has been fully implemented and deployed since March 2004. The DAK cataloging and publishing modules are being implemented. It is foreseen that these modules will need several iterations in the implementation based on user feedback and requirements. The beta version will include features presented in this paper. Future enhancements include:

- Building a more sophisticated security system based as on existing and emerging standards that are appropriate for the web services environment.
- Implementing a generic digital assets viewer. The viewer should support different file formats (PDF, DJVU, Images, Video and Audio).
- Joining the Open Source community by making the system source code publicly available and using free-of-charge development tools and database engine.
- Providing query translation tools to enable cross-language information retrieval.
- Using XML format for encoding objects metadata. This will facilitate exchange of objects among repositories.

8 REFERENCES

1. GNU EPrints. http://software.eprints.org/
2. L. Carr, G. Wills, G. Power, C. Bailey, W. Hall and S. Grange: Extending the Role of the Digital Library: Computer Support for Creating Articles. Proceedings of Hypertext 2004 (Santa Cruz, California, August 2004).
3. R. Tansley, M. Bass, D. Stuve, M. Branschofsky, D. Chudnov, G. McClellan and M. Smith: The DSpace Institutional Digital Repository System: Current Functionality. Proceedings of JCDL '03 (Houston, Texas, May 2003).
4. The Open Archives Initiatives. http://www.openarchives.org/
5. D. Bainbridge, J. Thompson and I. H. Witten: Assembling and Enriching Digital Library Collections. Proceedings of JCDL '03 (Houston, Texas, May 2003).
6. I. H. Witten, S. J. Boddie, D. Bainbridge and R. J. McNab: Greenstone: a comprehensive open-source digital library software system. Proceedings of the fifth ACM conference on Digital libraries (June 2000).
7. D. Castelli and P. Pagano: A System for Building Expandable Digital Libraries. Proceedings of JCDL '03 (Houston, Texas, May 2003).
8. Virtua Integrated Library System. http://www.vtls.com/
9. MARC 21 Standard. http://www.loc.gov/marc/
10. VRA Core Categories, Version 3.0. http://www.vraweb.org/vracore3.htm
11. Luna Imaging Software. http://www.luna-imaging.com/
12. P. Caplan: International Metadata Initiatives: Lessons in Bibliographic Control. Available at http://www.loc.gov/catdir/bibcontrol/caplan_paper.html (2001).
13. I. Saleh, N. Adly and M. Nagi: DAR: A Digital Assets Repository for Library Collections – An Extended Overview. Internal Report available at http://www.bibalex.org/English/researchers/isis/TR-DAR.pdf
14. ABBYY Fine Reader OCR software. http://www.abbyy.com/
15. Sakhr Automatic Reader OCR software. http://www.sakhr.com/
16. S. Payette and C. Lagoze: Flexible and Extensible Digital Object and Repository Architecture. Proceedings of ECDL'98 (Greece, September, 1998).