CSECS 2011, pp. 000 - 000

The 7 Annual International Conference on
Computer Science and Education in Computer Science,

July 06-10 2011, Sofia, Bulgaria

A DIGITAL REPOSITORY AT UNIVERSITY DEPARTMENT LEVEL

FILIP ANDONOV, JULIANA PENEVA, STANISLAV IVANOV

***Abstract**:* The reasons for creating a digital repository of separate department at university are discussed from the perspective of intent and initial planning. An attempt to identify the broad requirements concerning the development of a departmental repository is reported. General considerations about policies and functional requirements are outlined with respect to the institution context. Technical and system issues are briefly discussed.

***Keywords**: institutional repository, functional requirements, submission policy, management and technical framework*

***ACM Classification Keywords**: H3.5 On-line information services – Data sharing; H3.7 Digital libraries-Collection*

Introduction

Nowadays companies are realizing a business advantage by managing successfully their business data. Digital resources are increasingly being recognized as a very important organizational asset on a par with finance, raw materials, etc. Resources are built of different kind of documents as images, video or audio clips, animations, presentations, online courses, web pages, to name a few. Organizations vary in types and sizes but all of them exhibit an intensive use of digital resources because these resources are stored, distributed, shared and reused without difficulty. So, building repositories to manage the digital content is a very important activity that brings value in the inventive deliverables of the overall organization. Each time a digital resource remains undiscovered or simply not used the organization waste time or staff efforts, misses opportunities or looses possibilities to gain a competitive advantage.

During the last five years different types of repositories ranging from digital libraries through various institutional collections and e-journals up to collaborative learning environments have been built. Large companies are reporting for own repository investigations as well. In addition there are many workshops and annual open repositories [1] conferences that concentrate on important issues concerning repository creation and management. Despite of the disappointments for many organizations due to the resulted greater than expected costs for set up a repository, research effort in this area appears promising. Repositories increase successfully very quickly. In this perspective, universities and scientific institutions demonstrate remarkable activity. Open access academic repositories marked a boost of 300 during the mid of 2006. Since the beginning of year 2007 the growth of such repositories listed in the *Open*DOAR Database [2] shows a constant increase of 300 repositories per year up to its present number of about 1900. The main reason for this perpetual activity is the huge diversity of purposes, deposited

resources, services and potential users. Universities need to exhibit and deploy different kinds of its intellectual asset. It is a matter not only of user's convenience, but of representativeness and prestige as well. In this plan, it is quite natural that the main share of active repositories belongs to countries with advanced higher education and science. Up to now about 1900 scholar repositories all over the world have been reported, about 20% of them United States, 28% in Europe shared among United Kingdom, Germany, Spain, France and Italy. Other 13% reside in Japan, Australia and India. In Bulgaria there are three open repositories only [3, 4, 5] registered [6] in *Open*DOAR 2011.

Trying to follow the global trend, our university launched its own scholar electronic repository. Referring to content types, the university Scholar Electronic Repository [3] generally fits the common profile of repositories as reported by [2], in which dominate articles, papers and books. However, one specific exception arises: worldwide dissertations and theses make 52% of deposited materials, while at our university repository they are not collected at all. It is due to the restrictive submission policy applied – only faculty staff can submit documents. This way, many useful products of the educational process itself like case studies, student's research projects, diploma theses etc. are left beside. Applying active learning exercises, instructors could rely on deposited successful results of previous learning activities. It appears especially helpful in the domain of computer programming where every nontrivial problem implies many equally suitable solutions. This determines our decision to develop our own, at department level, digital repository to deploy digital content not covered by the university infrastructure: LMS Moodle and Scholar electronic repository. Our development should not duplicate these efforts and will deliver digital materials not offered by these two systems. Ensuring that proper digital material is visible for the long term is very important for the department as part of its positioning strategy. The goal of this repository would be to provide added value to

the Computer Science Education community, to our students and alumni. Moreover, the university educational policy encourages the shift towards e-learning and a flexible learning process. This implies reducing the face-to-face sessions, disseminating online coursework on a wider basis and training the students any time. So, designing a new infrastructure project and applying a standards-based approach to the management, preservation and access of existing and future digital resources is essential for the department to fulfill its mission as a team of lecturers and researchers.

In the context of the above, the main goal of this paper is to present our initial work on designing an institutional repository of the Department of Informatics at New Bulgarian University. We discuss what do we need and determine the type of the material to be stored in the repository. Creating a proper digital collection that captures and preserves the department's intellectual output would increase its visibility, prestige and public value. This repository will support learning and administrative processes of our department. To build an effective repository the technical set up process is to be planned properly. In section 2 we considered the requirements with respect to our institutional context. Section 3 focuses on technical and system issues. We summarize our findings and introduce our future work in section 4.

Functional Requirements and Policy Considerations

According to the SPARC alliance [7] institutionally defined repositories are scholarly, cumulative, open and interoperable. Generally speaking a department repository can be compared to a database with a set of services used to store, index and preserve scholarly materials, research findings etc. in digital formats. The main goal is to manage and disseminate digital materials created by the department and its community members [8]. The repository will be used for electronic publishing and housing of different digitized collections (the so-called grey literature e.g. theses, dissertations, working papers, reports, etc.)

concerning the knowledge management of the department. The final goal is to offer open access to scholarly research.

At the planning phase of building the department repository we focus on service design and policies. The next section addresses technological issues. We follow the guidelines given in [9, 10, and 11] for each stage of building the repository bearing in mind the requirements of our institution about copyrights, access rights etc.

First of all we have to define the services we intend to offer as the repository is not only determined by the software and the database containing the digital materials. The service model definition follows below.

1. Service's goal.

The service's mission is to raise the visibility of the Department of Informatics at New Bulgarian University. This repository will house digitized collections not stored in the Scholar electronic repository of the university and will encourage open access. It will facilitate our students, extending their access to properly collected and organized additional learning materials.

2. Type of content.

We will accept bachelor, masters and doctoral theses, student's research materials and original learning content from the department of Informatics. The user will not be allowed to download copyright protected content.

3. Key users.

Key users of the departmental repository are going to be students and faculty.

4. Key stakeholders

Administrators, students and internal research staff are the key stakeholders of the repository.

5. Free/versus charged services.

All services will be free of charge.

6. Library responsibilities versus the content users.

There are no library responsibilities about the content except copyrights observation.

7. Type of services.

Repository services concerns the management of corpora i.e. annotated collections of digitized objects. Making visible the stored content to the user groups can be defined as a top service priority. So, we can divide the services in two main categories: administration services and user's services. Administration services include data load, data store, long-term preservation, sharing and presentation of the content, group creation. Special authorization to use these services is required. The user's services facilitate the retrieval of digitized items of interest and comprise list and search.

A policy framework is very important to determine the operational boundaries within which the repository will deliver its services. This framework contributes for an easier use of the repository, permits for it support and facilitates the decision-making processes. Some policies need legal agreements i.e. definition of a deposit license and usage license that user agree to.

Policies can be classified as strategic and operational. Strategic policies reflect the wider strategic policies of the institution. New Bulgarian University has a high- profile vision statements [19] and defined procedures concerning research, teaching and theses. Following them the repository can be easily embedded within the university. Administrators will survey the deposit of diploma theses and other research output. As learning and teaching materials are deposited within Moodle, their store in the repository is optional.

Operational policies deal with day-to-day operations. They comprise:

1. Submission policies – only administrators will be allowed to deposit submitted materials after approval.
2. Collection policies – the repository will focus on computer science and mathematics. Final versions of the artifacts after a quality review will be accepted only.
3. Preservation policies – different policies will be set for different type of materials. We will keep theses as deposited whilst teaching materials, because of the dynamics in the computer science area more likely will be updated. Regular backups at least once a week will be made.

Technical ands System Issues

Taking into account that flexibility among the different collections is a key feature the goal of the department repository is to offer a proper infrastructure with a well defined range of services. A high level archival model to act as a framework is necessary. We adopt a well established model in this area – OAIS (Reference Model for an Open Archival Information System) [12] –see Fig.1.

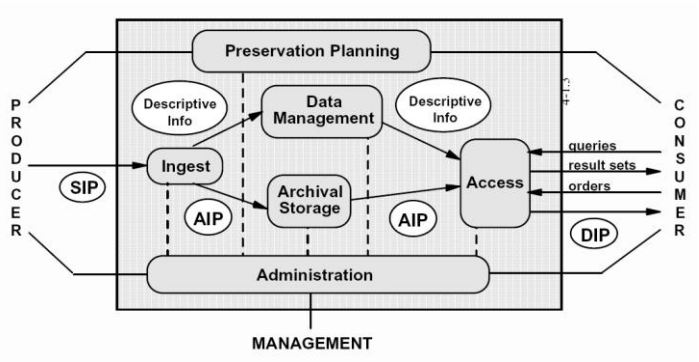

Fig.1 OAIS Functional Entities (from Reference Model for an Open Archival Information System - OAIS, 2009, Fig 4-1)

The OAIS environment is made up of the OAIS, i.e. the digital library system, the producers and consumers of its content and services, and the management and strategic input into the system. Within the OAIS are six main functions: ingest (submit), data management, storage, access, administration and planning. Common services e.g. operating system and networking services are assumed to be available. Evaluations concerning the usability of OAIS to build different kind of digital repositories are given in [13].

Taking into account the requirements we can make decisions concerning the repository infrastructure. To set up a repository three approaches can be followed [14]:

- do-it-yourself;
- use standard packages;
- outsourcing - external hosting.

With limited staff resources for long-term maintenance and support we have chosen to apply the most popular approach i.e. to use a standard package nevertheless that external hosting becomes recently more popular. Digital repository solutions consist of hardware, software and open standards. A wide variety of available software with different features and strengths exists. A functional comparison of repository software products is presented in [15]. Recently the more commonly adopted software solutions fall into two broad groups: open source and commercial software. Our investigations show that there are over 308 repositories using the EPrints software, about 711 – DSpace, 82 – Digital Commons. The rest of the software exhibits a limited (up to ten repositories) application. EPrints [16] is an open source platform for building repositories of documents like research literature, scientific data, and student theses. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets. It is applied for accessing, managing and

preserving scholarly works [17]. DSpace is used to develop the repositories at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences and the Faculty of Mathematics and Informatics, Sofia University. Digital Commons [18] offers external hosting for institutional repositories. It can include pre-prints and/or final copies of working papers, journal articles, dissertations, master's theses, conference proceedings, and a wide variety of other content types.

We have decided to run our repository on site locally, and to install it on a dedicated server. Initially the repository will be modestly populated so a quite basic server will be sufficient. Maintenance of this machine is a standard IT service. We have to make some considerations concerning possible hardware failures – a transition to other machine seems to be the best solution.

The next stage of the process of implementing a digital repository for the needs of our department was to select a system, which should be the most suitable system for our needs. We made a short-list, where the exclusionary criteria were popularity and price. Even with a short-list consisting of only three systems, Dspace, Eprints and intraLibrary, the list of criteria we considered relevant for evaluating these systems was too long. Thus in order to make a more quantifiable and formal judgment we decided to formulate the problem as a multiple criteria selection problem and to use a decision support system to solve it. The first step to take at this stage is to select the criteria. The criteria should obviously represent the qualities of the alternatives (the systems present in the short-list), there should be enough accessible information for the values of the alternatives with regard to these criteria, and a number of qualities of the criteria should be identified, the most important being whether the larger values are better or worse and what type of information the values have. The usual contradiction in such problems is between the positive qualities the alternatives have (basically what you get) and their price (what you pay). As in our case we excluded commercial systems, the

contradiction in the selection process was between the properties of the evaluated systems. We simplified the problem further by excluding from consideration those criteria. That even though important, do not differentiate our alternatives (they all have the same value). They are: Commercial Paid Support, End-user Deposition and Multi- language Support.

Having all these taken into consideration, we constructed the definition of our problem as described in Table 1:

Table 1 Problem definition

	Weight	DSpace	EPrints	intraLibrary
Number of Supported Item Types	2	8	7	8
Number of Supported Meta-data Formats	8	6	14	6
Number of Formats with Thumbnail Previews	6	4	6	6
Number of Formats to Convert from	4	6	1	0
Number of Advanced Searching features	4	3	2	3
Browse View Options	2	6	5	7
Number of supported Web 2.0 features	4	1	2	8
Number of supported Operating Systems	3	5	5	7
Number of Supported Database engines	6	2	4	2
Number of supported Scripting Languages	4	4	4	3
Machine-to-Machine Interoperability	4	9	6	3
Number of Administrators' Functions	5	3	4	3

The first column contains the names of the criteria, the second the relative weight of the criteria, the last three columns - the alternatives and the rows represent the criteria.

To solve the problem we used the Multichoice 2 - a system suitable for problems of different size and complexity. Multichoice 2 implements four methods for solving discrete multiple criteria optimisation problems, covering all three types of existing methods – weighting, outranking and interactive. The difference between these types of methods is the number of alternatives/criteria they are suitable for and the type of additional information they require from the decision maker(s). Because the interactive method is applicable only for problems with a huge number of alternatives and the implemented weighting method cannot be used for problems with more 6-8 criteria, the only choice was to select an outranking method. From the implemented two, we solved our problem with the PROMETHEE II method. The required additional information consists of the weights of the criteria, representing their relative importance, and the type of generalized criterion to be used for pair-wise comparison of the alternatives. PROMETHEE II has six predefined types, but because all of the criteria we use are quantitative and represent a number of features the evaluated systems have, we used ordinary criteria for all of them. One of the main functions of the system we are looking for is the ability to annotate the uploaded materials with different metadata, so we gave the criterion Number of Supported Meta-data Formats a bigger weight. From a usability point of view, showing the user a thumbnail preview of the content of the uploaded documents is important, so this is another criterion with big importance and thus weight. The versatility of the system with regard to software/hardware requirements is also important, because we do not have a budget for maintenance and that is the reason for the relatively big weight of the criterion Number of Supported Database Engines. All of the evaluated

systems support all popular operating systems, which made this criterion less important.

After giving the additional information the problem is ready to be solved. The final result is that EPrints is the winner, IntraLibrary is second and Dspace is last. The values of the evaluation function are shown below – see Fig.2.

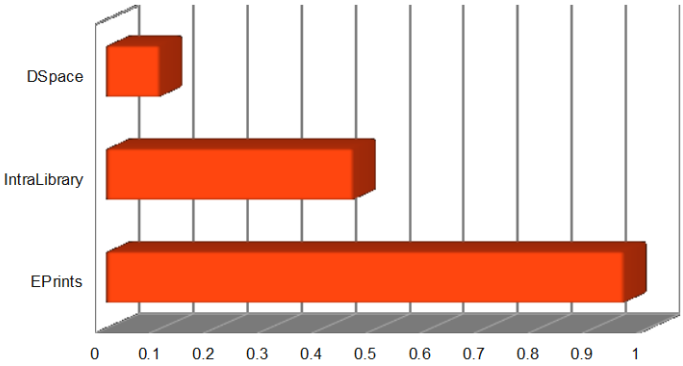

Fig.2 The values of the evaluation function

The EPrints system is written in PERL and the fact that it is one of the oldest can be seen in the code, written in old-style web programming. But in spite of that, EPrints has the right mix of features and is customizable and branding-friendly. There are several packages available for download – for the main Linux distributions (Debian, Ubuntu, Red Hat) and for Windows. The documentation available on the website [20] is extensive and useful, organized as a wiki.

From the usability point of view, the organization of EPrints is user-scenario oriented with listings of recent items, browseable views, and task bar showing the common kinds of operations to logged-in users. As a result, the most common operations and operation sequences are the most easily accessible in the UI. We find that it is a really good practice to make extensive use of RSS technology, e.g. RSS feeds for whole

repository and the fact that the results of any search can be exported as an RSS feed.

Although the fact that the other evaluated systems have more search options than EPrints, the system has powerful user-friendly features for browsing and searching the available documents. EPrints can create browse views by any complex criteria and creates and shows thumbnails of images, videos and PDFs.

One of the main strengths of the EPrints is its agile input/output interoperability. The output from any search can be exported as digital library interoperability formats (METS, Dublin Core, etc), as bibliography managers format (such as BibTeX) and even to some web services as Google Earth, Similie TimeLine and others. EPrints records can be imported from many formats or external web services e.g. PubMed and CrossRef.

So, following the local policies and practices we have decided to run an open source repository platform. This choice reflects the good will and the IT expertise of the department's staff. Running open source software appears to be the cheapest solution as the installation and the customization of the repository require a relatively short list of intensive activities. The skills required depend on specific repository platforms i.e. the programming language they are written in. There are common skills such as HTML, Web page design, SQL applicable to all choices.

In order to justify the choice of the software, pilot installations of some open source packages has been undertaken. These are used as test beds for the overall repository development. A pilot system will be used to tune the software parameters. We intend to perform users' acceptance testing as well.

Conclusions and Future Work

In this paper an attempt to identify the broad requirements concerning the development of a departmental repository is done. Specifying a

repository system imposes the definition of services and a policy framework. The choice of the proper software to build the repository implies a systems requirements analysis. Next we will proceed to determine interoperability, performance and quality requirements. Obviously digital repositories deliver value added services and offer benefits to their stakeholders and the wider world.

The decision to create one more repository to manage proper digital content is challenging. One could argue that organizational digital assets already are stored in many types of systems e.g. locally developed closed systems, virtual learning environments, portals, etc. That's why it is very important to summarize the functional requirements of a departmental repository so as to determine its inclusion in the existing institutional information architectures.

Bibliography

[1] http://openrepositories.org/

[2] www.opendoar.org

[3]http://eprints.nbu.bg/ – New Bulgarian University Scholar Electronic Repository

[4] http://research.it.fmi.uni-sofia.bg - Research at Sofia University

[5] http://sci-gems.math.bas.bg DSpace at IMI

[6] http://www.driver-support.eu/pmwiki/index.php?n=Main.Bulgaria

[7] http://www.arl.org/sparc

[8] Lynch C. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age" ARL, no. 226 (February 2003): 1-7.

[9] Barton M. Creating an Institutional Repository: LEADIRS Workbook, MIT Libraries, 2004-2005.

[10] Pederson M. et al. 100 Year Archive Requirements Survey Storage Networking Industry Association (SNIA), 2007.

[11] http://ciard.net/pathways - Develop a Repository for Digital Content, October, 2009.

[12] CCSDS 650.0-P-1.1, Reference Model for an Open Archival Information System (OAIS) (Pink Book, Issue 1.1, August 2009)

[13] Allinson J et al. OAIS as a reference model for repositories, UKOLN, University of Bath, 2006.

[14] http://www.rsp.ac.uk/start/setting-up-a-repository/technical-approaches/

[15] http://www.rsp.ac.uk/start/software-survey/results-2010/

[16] http://www.eprints.org/

[17] http://www.dspace.org/

[18] http://www.bepress.com

[19] http://www.nbu.bg

[20] http://wiki.eprints.org/w/Main_Page

Authors' Information

Philip Andonov, Ass. Prof, New Bulgarian University, 21 Montevideo St. 1618 Sofia, Bulgaria, fandonova@nbu.bg.

Major Fields of Scientific Research: *operational research, group decision making, database systems*

Juliana Peneva, PhD, Assoc. Prof,New Bulgarian University, IMI – BAS, 21 Montevideo St. 1618 Sofia, Bulgaria, jpeneva@nbu.bg.

Major Fields of Scientific Research: ***database systems, software engineering, information systems, e-learning***

Stanislav Ivanov, PhD, Assoc. Prof, New Bulgarian University, 21 Montevideo St. 1618 Sofia, Bulgaria, sivanov@nbu.bg

Major Fields of Scientific Research: *object-oriented modeling, programming*

This research is partially granted by Bulgarian National Science Fund, Ministry of Education and Sciences in the frame of the project “Automated Metadata Extraction for e-documents Specifications and Standards”, contract N D002(TK)-308/ 19.12.2008.