Professional paper

Implementation of Digital Repository at the Ruđer Bošković Institute: Organizational and Technical Issues

Alen Vodopijevec
Ruđer Bošković Institute, Library
Bijenička cesta 54, Zagreb, Croatia
alen.vodopijevec@irb.hr

Bojan Macan
Ruđer Bošković Institute, Library
Bijenička cesta 54, Zagreb, Croatia
bojan.macan@irb.hr

Summary

This paper will focus on implementation of digital repository at the Ruđer Bošković Institute (RBI). Based on the RBI needs and considering consolidation of IT services, it was decided that RBI digital repository will include three different types of archived content: scientific output, documentary and press clipping materials. This is the main difference against other commonly implemented institutional repositories which archive mostly only scientific articles and thesis. Choosing of appropriate software, its implementation and customization will be discussed, as well as the methodology of gathering content for the repository and the method of its archiving. Archived material will be organized in collections based on the RBI needs and divisional structure. During the process of creation of metadata schemes for different types of archived content, especially those of scientific character, special attention was paid to OAI compliance with other digital/institutional repositories, as well as with Croatian Scientific Bibliography (CROSBI). The goal is that researchers should archive scientific materials primarily in RBI digital repository, while metadata will be automatically exported to CROSBI. Successful implementation of this data-exchange mechanism could also be useful to other institutions which are considering creating their own repository.

Key words: institutional repository, digital repository, IR, CDS Invenio, Ruđer Bošković Institute, open access, OA

Introduction

The Ruđer Bošković Institute is the largest Croatian scientific research institution in the fields of natural sciences. In the multi-disciplinary environment of

the Institute 530 academic staff (375 researchers and 155 Ph.D. students) [8] work on problems in experimental and theoretical physics, chemistry and physics of materials, organic and physical chemistry, biochemistry, molecular biology and medicine, environmental and marine research and computer science and electronics. Institute has 11 divisions, 3 centers, a library and sections for maintenance, technical service and administration. In the year 2008, the RBI have had 136 projects in basic research, which are funded by the Ministry of Science, Education and Sports (MSES), as well as 41 international projects, 67 applied and technological projects and 4 HITRA projects [8]. RBI staff is also active in teaching at universities and in 2008 they contributed 78 undergraduate courses and 245 graduate courses to the program of higher education in Croatia. The total number of research articles published by RBI scientists in 2008 was 446, whereof the majority was published in high ranking international journals [8].
Institutional repositories are "digital collections that capture and preserve the intellectual output of a single or multi-university community" [3]. Benefits from IR are many. Individual researchers gain better visibility of their papers which can, therefore, be cited earlier and more often than papers which are not in OA. Research community can find and access information more easily and institutions gain on their visibility and prestige by collecting all its scientific output in one place, rather than to be spread amongst hundreds of journals [10].
On July 17th, there were 4 active institutional repositories in Croatia (School of Medicine, Faculty of Philosophy, Faculty of Mechanical Engineering and Naval Architecture, all from University of Zagreb and Digital repository of the Information Sciences Department at the Faculty of Philosophy, University of Osijek) and Portal of scientific journals of Croatia – HRČAK. At the same time, there were 1429 repositories registered in Directory of Open Access Repositories - OpenDOAR (4 from Croatia) [13] and 1411 in Registry of Open Access Repositories - ROAR (3 from Croatia) [7].

The idea about RBI digital repository

The idea of opening the science to the public was the main "spiritus movens" that led towards the RBI Digital Repository project proposal. As stated in the document entitled "Science and Technology Policy of the Republic of Croatia 2006-2010", scientific and research output produced within publicly funded projects should be freely available to the public [15]. Considering the size, status and importance of the RBI for the Croatian and International academic society, and new trends in scientific communication, the RBI Library came up with idea about implementation of institutional repository at the RBI which should be such platform for depositing and disseminating the results of publicly funded projects. The idea was initially born in 2006 and Library wrote a draft of the proposal project for an institutional repository, which was introduced to colleagues from Public Relations office (PR Office) and from Center for Informatics and Computing. They supported the idea and suggested that the Project should be expanded and incorporate

other digital content produced on RBI. These suggestions were adopted so the project proposal was rearranged with their help and its final version was created in July 2007 and presented to the RBI administration [4]. The Project was approved, as well as its financial construction for a period of first two years needed for acquisition of necessary hardware. Also, one person was designated to the library for working part-time on digitization and organization of RBI old documentary materials (photographs).

The main objectives of this Project were:

- archiving and preservation of digital content of the RBI
- gathering all scientific output of the Institute on the single site and offering it in OA to the community
- creating a digital archive for archiving of documental and press clipping materials about the RBI
- increasing of the RBI's visibility and it's scientific contribution to the science
- promoting of the OA initiative at the RBI and in Croatia
- helping RBI staff to publish their work on their personal web pages without fears of breaking copyright law

RBI digital repository will consist of three virtually separated parts:

- self-archiving online platform of RBI's **scientific output** based on OA principles. Such platform is commonly referred as "institutional repository";
- **documentary archive** - online digital storage of important historical and current multimedia content produced by RBI or with RBI as main theme of such contents;
- **press-clipping archive** - online digital storage of press-clipping content.

It was planed that the implementation of the RBI digital repository will take place in 7 phases:

1. Setting up hardware and software
2. Resolving copyright and licensing issues
3. Training the personnel for digitalization and administration of the repository
4. Digitalization of the documentary materials and initial data archiving
5. Presenting repository to RBI staff and OA advocacy
6. Depositing materials and regular maintenance of the system
7. Establishing institutional self archiving mandating policy and depositing license

RBI digital repository

Choosing appropriate software system

After deciding what kind of repository does RBI need, it was necessary to choose appropriate open source software for it. Software that suit RBI's needs would have to fulfill following requirements:

- open source software
- functional and extendible integrated search engine

- OAI-PMH compliance
- integrated standard internationalization an localization functions
- variety of standards and data formats for metadata representation
- group or role based access right privileges system
- fully customizable collection tree structure.

Therefore testing of several software options (CDS Invenio, EPrints and DSpace) was conducted (table 1). As a result of this testing, CDS Invenio (http://cdsware.cern.ch/) was recognized as the most promising solution for RBI’s multi-purpose repository and it was decided to do more detailed tests on it. Despite that, it is still possible to change it if better software appears on the market. Until now a significant amount of work was done on localization and analyzing administration interface, especially the submission process. CDS Invenio is very flexible system and despite the fact that it requires lots of work on its customization, as well as other tested products, it was concluded that with Invenio it is possible to do more in less time.

Table 1: Feature comparison of tested software products

Tested functionality	**CDS Invenio**	**DSpace**	**Eprints**
Programming language	Python	Java	Perl
Database engine	MySQL	PostgreSQL	MySQL
Localization	Yes – standard .po files	Yes – Language packs based on Java Standard Tag Library	Yes – XML files
Customization of UI and depositing system	Web interface, extensive HTML and Python programming required.	Web interface, larger changes require Java coding and recompiling application.	Majority of configuration is handled by XML files and some HTML templates. Basic knowledge of respective technologies required.
Default metadata standard	MARC with possible export to other standards	DC	DC
User authentication	Local database, LDAP, Shibboleth	Local database, LDAP	Local database, LDAP
Access control	Role based	Groups as roles	3 default groups
Search engine	Python custom	Java Lucene	Perl custom

Complexity of installation and maintenance was not rated nor compared because it depends on the extent in which one would have to customize default features of an application and, of course, it depends on competence and availability of IT staff team members involved in the process of implementation.

Type of archived materials and supported formats of files

As already mentioned, Digital repository of the RBI will archive three different types of materials: RBI's scientific output, documentary materials and press clipping materials about RBI. Those materials will be archived in textual, video, audio and pictorial form and repository will support uploading of all available file formats, although a certain formats for specific data types of archived materials will be preferred. Table 2 brings the list of preferred file formats for different types of digital content. Mentioned formats are preferred because of the possibilities to represent them on web UI, for e.g., showing images, streaming audio and video. Furthermore, OGG (theora for video and Vorbis, Flac for audio) are open-source and patent free.

Table 2: Preferred file formats of archived materials for different types of digital content

Type of digital content	Preferred file formats of archived materials
Textual materials	PDF, DOC, ODF, RTF, PPT
Video materials	OGG (Theora)
Audio materials	OGG (Vorbis, Flac, Speex)
Pictorial materials	TIFF, JPG, PNG, PPT, ODF

RBI scientific output

This is the type of materials which is usually archived in IR. It is scientific material in a form of articles, book chapters, books, reports, posters, data sets etc. Many of mentioned items are copyright protected and it will be necessary to investigate the terms under which those materials can be archived and in which version. This will be elaborated later in this paper when talking about copyright issues.

Documentary materials

Under this type of materials photographic materials of the Institute, its building, staff, equipment and all kind of promo materials (brochures, posters etc.) will be included. For that reason, a project of digitizing of those materials was conducted. All photographic materials were gathered, organized and digitized. During this process, it was realized that there's a great number of material, whereof lots of pictures are duplicates or very similar to each other and that the process of selection of those materials will be needed. It was also realized that it would be very useful to identify people on the photographs and to include those information to its description. Therefore it was decided to have a inter phase in which photographs will be uploaded in lower resolution into the online gallery (which currently holds over 6000 photographs from 1950s until today) and they will be visible to RBI staff who will be able to tag people on them and describe wider context of taken photographs. Number of visits to uploaded photographs will be one of the criteria for choosing which photographs will be archived to

the RBI digital repository, but the most important criteria will be historical significance of a certain photograph. It will be possible to archive single photograph or store more thematically related photographs into an album along with their short descriptions. Photographs will be archived in their original resolution and quality with smaller images for quicker preview.

Press clipping material about RBI

The third type of materials included to the RBI digital repository is press cut materials of all published articles about RBI, recorded TV or radio shows, as well as archive of press release material from RBI's Public Relations (PR). The idea is that RBI's PR write press release, deposit it to RBI digital repository and all interested media can download it from the repository and publish it. Permission to deposit press cut and press release materials will be given only to PR Office.

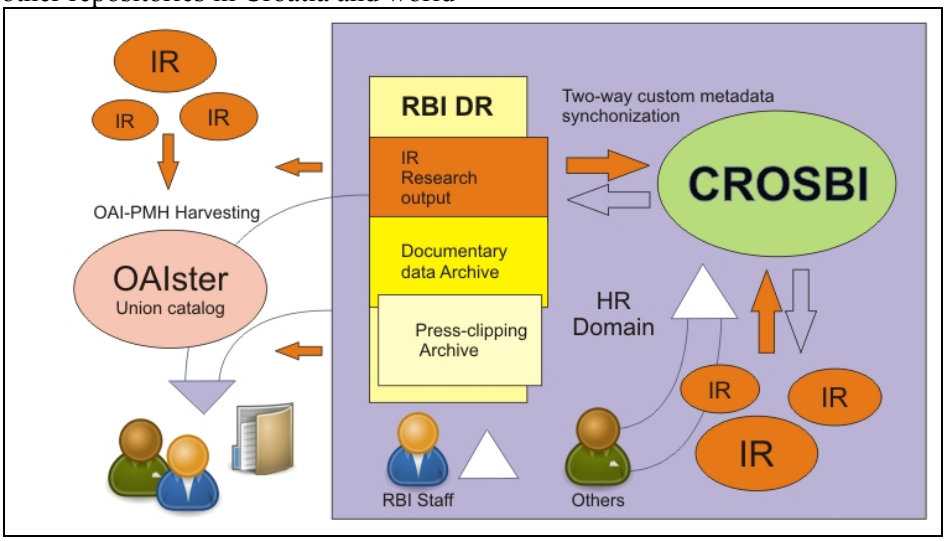
Metadata structure and self archiving

In 2002 the RBI Library started the project called Croatian Scientific Bibliographic Database (CROSBI) (http://bib.irb.hr/), a bibliographic database in which all Croatian scientists are "forced" by MSES to deposit metadata because they are obliged to attach the listing of their published works from it for their projects reviews, scientific promotion etc. Although it is primary bibliographic database, CROSBI also has an option for uploading full text documents. Until July 27th 2009, there were more than 243.000 bibliographic records archived into the database, whereof 13.000 full text documents [5].

Considering the fact that RBI staff has to deposit metadata about published papers in CROSBI database, and that the development of the new RBI digital repository is in progress, the plan is to make those two services compatible. Experience of institutions with similar situation were studied [1, 2] and it was decided that data synchronization between two systems should be enabled. That's why it was necessary to have metadata compatibility issue on mind when creating set of metadata needed for description of different types of materials expected to be hosted in the repository. While CROSBI is not based on any metadata standard, the metadata scheme used for describing items in the repository is based on MARC standard, but it could be mapped to other standard metadata formats as well (such as DC for OAI-PMH compliance). Therefore a mapping of CROSBI and RBI repository metadata fields will be done in order to enable this metadata exchange. Main goal regarding implementation of interoperability between existing services is to wipe out the need for multiple depositing of documents and/or multiple metadata submitting procedures. The idea is to insert metadata (and deposit document) only once, into researchers "home" repository and afterwards the system will do the replication process of certain types of content (scientific articles, posters etc.). Scientist would only have to supplement replicated record where applicable (e.g. add custom metadata regarding

publication details and category of an article in CROSBI). There will also be a possibility to harvest metadata from CROSBI to the RBI digital repository for initial import in Institutes repository. Successful implementation of this module could be later replicated by other institutional repositories in Croatia (Figure 1).

Figure 1: Interoperability of the Digital repository of the RBI with CROSBI and other repositories in Croatia and world



On the basis of metadata description, virtual collections will be formed. These virtual collections will be based on the RBI organizational structure and its specific needs, as well as on the fact that repository will archive three different types of materials which should be clearly visible at the repository homepage as separate collections.

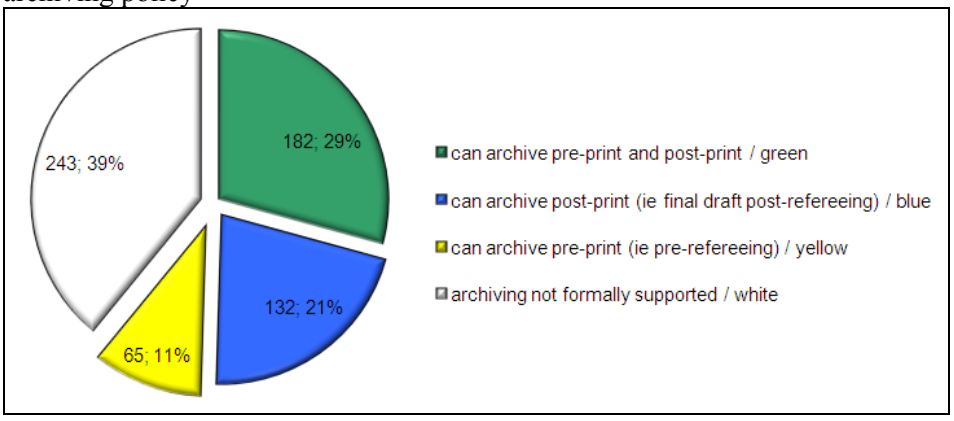
One of the often mentioned problems in literature about institutional repositories is problem about gathering content for repository [2, 6]. Some institutions decided that library staff will deposit items into the repository in behalf of their faculty staff [2], while others decided for self-archiving [11]. RBI scientists already have experiences with depositing metadata to CROSBI. To take advantage of this fact it was decided that our repository will be based on self-archiving, rather than on library staff depositing. Furthermore, it will be necessary to create institutional self-archiving mandating policy. This obligation shouldn't be a big problem for the RBI staff because they are already used to enter bibliographic data into CROSBI, and considering planned interoperability between two systems, it shouldn't take to much extra time for depositing.

Copyright, licensing and access rights

Copyright issues are crucial for every institution which is implementing an institutional repository and has to be considered very carefully. Copyright issues with documentary materials are in most cases very clear – RBI is a copyright holder and it can publish them in its repository. Press releases and press cut materials are, according to Croatian copyright law, free of copyright law and therefore, can be archived in our digital repository [14] with the exception of audio/video materials (TV and radio shows) which cannot be freely published.

However, the situation with scientific content is much more complicated because in majority of cases publications rights are transferred to the publisher. The SHERPA/RoMEO database of publishers is here out of great help. This database is used to determine the rights of authors to include papers published in scientific journals in the IR. At July 24th, SHERPA/RoMEO database included details of the policies of 662 publishers [9]. The publishers are divided into four groups according to archiving policy, and every group is represented with different color. Figure 2 brings data about publishers in RoMEO database according to their archiving policy. Most of publishers (61%) allow depositing of pre- or post-print version of paper in an IR.

Figure 2: Number and percentage of publishers in RoMEO according to their archiving policy



As mentioned before, archiving to the repository will be based on author's self-archiving and the final goal is to teach them how to check by themselves in SHERPA/RoMEO database which version of published paper can they archive to the repository. At first, the library staff will be there to help the authors to check the journal's policy about self archiving, but the library intervention is to be reduced to the minimum with time.

A suitable deposit license needs to be created and embedded into the submission form. This license will be based on the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported model (http://creativecom-

mons.org/licenses/by-nc-nd/3.0/) and the SHERPA model license and it will consist of clauses covering the institution's rights to store, disseminate and preserve deposited items, as well as providing assurance that copyright is owned by the author or permission has been given by the owner. Submitter of material will have to agree to the licensing terms and conditions or an item will be rejected.

Besides freely available contents, RBI digital repository will contain a great amount of digital data (documents, multimedia) that require access control mechanism, e.g., articles during embargo period, datasets that can be accessed only by few people, experimental data available to lab staff etc. CDS Invenio uses "role based access control" (RBAC) so it is possible to create fine grained rule sets for authorization purposes. User authentication and authorization mechanism relies on RBI's existing LDAP user database and further development is on the way to implement "Single Sign On" integration with Croatian Academic Authentication and Authorization infrastructure – AAI@EduHr. The same authentication and authorization mechanism users use for logging in to CROSBI and many other services developed by the RBI Library and other institutions in Croatia. Single sign-on principle is based on existence of central identity provider which guarantees that certain person is really who she pretends to be and it guarantees that she has certain rights when accessing service on service provider side. In Croatia every academic and research institution has its LDAP user database and there is central service maintained by SRCE (http://www.aaiedu.hr/) that acts as proxy between end user and service provider who wants to authenticate and/or authorize users through this mechanism. User can use one username and password but there is no interoperability so that remote independent services have the information that user is once logged in into system and can access any available resource. There comes Shibboleth (http://shibboleth.internet2.edu/) in place. Implementing of Sibboleth into CROSBI database is currently in the process and it was decided to use the same service for authenticating and authorizing users of RBI digital repository.

Benefits of such approach is that users have to log in only once and then they can access any available resource identified by Single sign-on system without the need for any further entering of login details. In that manner it is possible to restrict some content to much wider audience. For example, it is possible to configure group of Croatian students that have rights to download certain datasets, a group of users that belongs to some specific institution and, of course, RBI internal groups based on LDAP attributes such as Division, Labs, etc. that can access their "private" area within the repository.

Future perspectives

To fulfill its goal and meaning IR must continuously grow and researchers must accept it as a way of scientific communication. A literature cites numbers of reasons why faculty participation rates are often so low and why are IR recruiting new content to slow: from the lack of awareness of the existence of institutional repositories [12], concerns about copyright and intellectual property is-

sues, concerns that depositing into the IR will be considered prior publication [6] to the fact that self-archiving is additional obligation for them and they are just not motivated to do it. There's no clear strategy that will lead for sure toward continuous and strong growth of the IR, but Mark and Shearer came to the conclusion that it is very important to promote repository on campus, because it raises awareness of the existence of the repository, and to establish institutional self archiving mandating policy [6]. That's why it is very important for RBI to complete institutional self archiving mandating policy as soon as possible to ensure better chances for successful growth of the repository. Furthermore, it is important for the RBI to stay on track with current achievements in exploring long-term data preservation strategies (backup solutions, possibility of access to data after longer period of time) and with ongoing activities concerning interoperability and disseminating of archived content e.g. OAI-ORE (http://www.openarchives.org/ore).

References:

[1.] Afshari, Fereshteh; Jones, Richard. Developing an integrated institutional repository at Imperial College London. // Program: electronic library and information science. 41 (2007), 4; 338-352

[2.] Barwick, Joanna. Bulding an institutional repository at Loughborough University: some experiences. // Program: electronic library and information science. 41 (2007), 2;113-123

[3.] Crow, R. The Case for Institutional Repositories: A SPARC Position Paper, Scholarly Publishing and Academic Resources Coalition, Washington, DC. 2002. http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf (Access date: July 17th, 2009)

[4.] Digitalni repozitorij Instituta "Ruđer Bošković": prijedlog projekta. Knjižnica Instituta "Ruđer Bošković", Zagreb, 2007. (interni dokument)

[5.] Hrvatska znanstvena bibliografija. Skupni statistički podaci. http://bib.irb.hr/skupni_podaci (Access date: July 27th, 2009)

[6.] Mark, Timothy; Shearer, Kathleen. Institutional repositories: a review of content recruitment strategies. (2006). http://archive.ifla.org/IV/ifla72/papers/155-Mark_Shearer-en.pdf

[7.] Registry of Open Access Repositories (ROAR). http://roar.eprints.org/ (Access date: July 17th, 2009)

[8.] Ruđer Bošković Institute Annual Report 2008. Zagreb : Ruđer Bošković Institute, 2009

[9.] SHERPA/RoMEO. Publisher copyright policies & self-archiving. http://www.sherpa.ac.uk/romeo.php?stats=yes (Access date: July 24th, 2009)

[10.] Sparc Europe. Institutional Repositories: A Guide to Open Electronic Archive. http://www.sparceurope.org/resources/hot-topics/institutional-repositories (Access date: July 10th, 2009)

[11.] Sutradhar, B. Design and developement of an institutional repository at the Indian Institute of Technology Kharagpur. // Program: electronic library and information science. 43 (2006), 3; 244-255

[12.] Swan, Alma; Brown, Sheridan. Authors and open access publishing. // Learned Publishing, 17 (2004), 3; 219-226

[13.] The Directory of Open Access Repositories - OpenDOAR. http://www.opendoar.org/ (Access date: July 17th, 2009)

[14.] Zakon o autorskom pravu i srodnim pravima. Narodne novine. 79 (2003). http://narodne-novine.nn.hr/clanci/sluzbeni/2003_10_167_2399.html

[15.] Znanstvena i tehnologijska politika Republike Hrvatske: 2006.-2010. Zagreb : Ministarstvo znanosti, obrazovanja i športa, 2006.