[EN] 09. Creating Institutional Repositories Based on the dLibra System

On this page, selected aspects of the use of the dLibra system for creating institutional repositories are discussed. The information is complementary with respect to the rest of the documentation of the system and, on its own, does not constitute an exhaustive presentation. In the appropriate places on this page, links to selected fragments of the documentation are provided. The page has the structure of questions and answers. We encourage our readers to submit new questions in the comments to this page or in the question and answer service of the Digital Library Federation.

Question List

What is the difference between digital libraries and institutional repositories?

In their basic scopes, both digital libraries and institutional repositories have the same aim: providing access to digital objects – usually broken down into collections – and offering additional tools, such as searching, browsing, or indexes. Therefore, those phrases are often used interchangeably, especially in the technical context. The difference usually lies in the type of the offered resources. The term "digital library" is customarily applied to services which provide access to library collections in the digital form (both digitized collections and collections which have been 'born digital'). The term "institutional repository" describes services which provide access to the effects of the work (usually scientific and contemporary) of the people employed in the given institution. The materials do not have to be officially published – they can also be presentations or reports (for example, technical reports) created in the institution. There are various approaches to the interrelationships of the two systems. Depending on the context, the terms are sometimes used interchangeably. Sometimes, the term 'digital library' is said to have a wider meaning while the term 'repository' is viewed as a subset (for example, a collection) of what can be stored in a digital library. On the other hand, some people believe that a 'repository' is the more general term because a repository may contain works which have not been reviewed or officially published but are only made available online. There are other synonymous terms, such as 'digital museum' or 'digital archive', which refer to variants of the systems described above, different with respect to the offered resources or the way in which those resources are described (the metadata schema) or presented (the user interface, added services).

Apart from the potential differences with respect to the nature of the collected and offered resources, the way in which the resources are collected is often quoted as a distinguishing characteristic of digital repositories. In the case of typical digital libraries, a library user collects the materials and enters them to the system. In the case of institutional repositories, it is often assumed that materials are collected largely by means of the self-archiving method, that is, a process in which authors send their works, for example, through a special online form, to the repository system. Next, the administrators/moderators of the repository verify the material and decide whether it should be made availble or returned to the author for the necessary corrections to be made. Whatever the principles of operation, some changes (for example, completing an object descriptions or assigning objects to appropriate collections) can be introduced by repository administrators, and some can be made automatically (for example, conversion of files saved in editable formats, such as DOC, to archiving /distribution formats, such as PDF/A). In such a case, the institutional user authentication system is often integrated with the institutional repository authorization and authentication system.

Also see

- the "Digital libraries and institutional repositories – what is the difference?" question in the "Questions and answers" service of the Digital Library Federation.

Should a digital library and a repository within the framework of one institution be separate initiatives or is it better to combine them?

That is always an individual decision. Below, we discuss a few basic advantages of both approaches.

- Combining a Digital Library with a Repository
  - **For end users: easier access to a greater number of objects** – one access point for the resources of the (digital) library and the works of the employees of the institution.
  - **For administrators/editors of a digital library / repository: simpler and more comfortable management of collecting and providing access** – there is one system of collecting resources and making them available for maintenance and management.
  - **For IT specialists: simpler and cheaper system maintenance** – one information system for maintaining, administering, updating, monitoring, and protecting; one equipment infrastructure.
- A Digital Library and a Repository as Separate Initiatives
  - **For end users: easier access to a specialized scope of materials** – the access point only to the contemporary works of the employees of the institution.
  - **For institution employees: greater opportunities for presenting one's professional profile and the profile of the institution** – a dedicated server which only presents the publications and other materials which are being created in the institution.
  - **For administrators / repository editors: new options for promoting the repository** – the option to register the data and scientific content in the browsers and data aggregators which only accept services matching the definition of an (open) institutional repository.

When the two theoretically alternative approaches are analyzed, one may come to the conclusion that all the advantages could be enjoyed in a system allowing the collection of various types of data in one database and only breaking them down to separate interfaces/portals at the presentation level. Such an approach is possible in the dLibra system, based on an appropriately designed collection structure.

For example, let us assume that we want to construct one system which would encompass three types of materials:

- historical library collections,
- contemporary publications of the employees of the institution, and
- current copywrighted publications, only available in Poland.

For that purpose, the following collection structure can be designed in the dLibra system:

- (M) The main collection of the dLibra system
  - (A) The digital library of institution X
    - Collections
      - Subcollections
  - (B) The institutional repository of institution X
    - Collections
      - Subcollections
  - (C) The internal repository of institution X
    - Collections
      - Subcollections

It is a complex example, but it is presented with the view to illustrating the broad scope of possibilities offered by the mechanisms of the dLibra system. Usually, the main collection (M) is the basis for the reader application in the dLibra system, in accordance with the simple rule of "one implementation = one website interface with a complete set of resources". The idea behind such a structure is that the main collection (M) should not be accompanied by a dedicated website but by three portals based on, respectively, collections A–C. Each of those three portals can have its own structure of collections and subcollections, and a consistent policy of making resources available should be created. The policy should encompass all the portals in such an implementation of the dLibra system. Since an object in the dLibra system can belong to more than one collection, one publication can be assigned to several collections visible in various portals.

For example, the following rules could be adopted in the example above:

- digitized library resources are assigned to portals A and C,
- all employee publications are assigned to portals A, B, and C, and
- contemporary copyrighted publications which are only to be published on the premises of the institution are assigned to portal C

Given those assumptions, we obtain:

- portal A which contains the information about all (library and repository) resources which institution X possesses and can make available to the public or which are authored by the employees of institution X;
- portal B which contains the information about all the works authored by the employees of institution X; and
- portal C which contains the resources from portals A and B and the information about the publications which are only published on the premises of institution X.

Access to publications in portals A and B can be authorized if need be. Access to the whole portal C is restricted to the range of the IP addresses of the local network of institution X. On the premises of the X institution, it is best to use portal C. Outside of the premises, a user can use portal A or portal B. Moreover, portal B can be promoted and registered as an institutional repository.

Also see

- "Can digital libraries and repositories be combined?" in the "Questions and Answers" service of the Digital Library Federation. Pytanie "Czy można łączyć biblioteki cyfrowe i repozytoria?" w serwisie FBC "Pytania i odpowiedzi"

dLibra is a system for building digital libraries, so can institutional repositories be created on its basis?

Yes, the dLibra system can be successfully used for building institutional repositories. It has all the functions necessary for collecting digital objects and metadata and making them available; it also supports self-archiving.

One important characteristic of institutional repositories is that authors can submit their works on their own (so-called self-archiving). Does the dLibra system support that?

Yes, it is described in detail in chapter 04. Alternative Presentation Versions (Multi-Format). The dLibra system can also be integrated with external user authentication systems; see 03. Integrating with Single Sign-On systems.

I want to make materials available with open licenses and make some materials available only to the employees of my institution – all the materials are the employees' publications. Should I create two separate repositories or is it possible and worthwhile to have those materials combined within the framework of one system?

Two such repositories can be run as one information technology system with two website portals. It can be done in a similar way to combining a digital library with a repository, which has been described in answer to the question: Should a digital library and a repository within the framework of one institution be separate initiatives or is it better to combine them?

The advantages of separating such repositories to two separate website interfaces are, first of all, the possibility of promoting at least a part of publications – the open ones – as a 100% open repository (a repository which only contains open publications). However, it is worth considering whether 100% opening is, indeed, necessary. For example, the OpenDOAR service, one of the most popular global aggregator of metadata from open repositories, quotes the following reasons as the most popular ones for rejecting registration applications:

- the service is not regularly available;
- the service is a journal and not a repository;
- the service does not contain any open-access materials;
- the service only contains bibliographic information and, possibly, links to external services, and it does not contain full content;
- the service is a library catalog or an e-book service, with the content only available in the local network;
- the service requires logging in (even if it is partially free) for gaining access to the theoretically open-access materials in it; and
- the service provides access to content commercially (the access is paid).

Apparently, the OpenDOAR service does not eliminate repositories which contain both open-access and restricted-access (for example, only available to employees on the premises) materials.

How can I combine a digital library and an institutional repository in the dLibra system?

The answer has been given in response to the question Should a digital library and a repository within the framework of one institution be separate initiatives or is it better to combine them?

I would like the materials from my repository to be visible in the Google Scholar web search engine. Does the dLibra system support it somehow?

Yes, the dLibra system fulfills the requirements of the Google Scholar web search engine concerning indexing, for example, on the page on which the metadata of particular objects in the HTML code are presented, in the <HEAD> section, there are appropriate tags with publication metadata. The dLibra system also makes it possible to download metadata in the RIS and BibTeX format, which makes it easier to later use the metadata in a scientific work. On the pages of the Google Scholar web search engine, the dLibra system is not listed as recommended (only two popular open-source repository systems and one hosted repository service provider are mentioned), but that does not mean that the dLibra system is not compatible with the requirements of Google Scholar.

Apart from the support on the part of the dLibra system, it is very important for files with scientific works published in the dLibra system to be prepared with respect to the requirements of Google Scholar. The absolute minimum is publishing files in the PDF format with a text layer (not only scans). For more information, see the Google Scholar information service.

How should PDF files be prepared for the Google Scholar web search engine to index them easily?

For detailed information about creating PDF files, see the page with the rules for creating PDF files for Google Scholar (in English). Generally speaking, a PDF file should be prepared in accordance with the following convention:

document title – it should be the greatest fragment of text, at the top of the page, in at least the 24-point font size; the same font should be used for the whole title, and all other texts on the page should be in a smaller font size than the title font size – otherwise, the other, greater text can be erroneously interpreted as the title;

document authors – they should be entered just after the title, in a slightly smaller font size which is, however, greater than standard text (so the size could be, for example, from the 16–26 point range); the same font should be used for all author names, and the font of the names should be greater than the font of section headings – otherwise, the other, greater text can be erroneously interpreted as author names;

particular authors should be separated with commas or semicolons, and their affiliations, degrees, and certificates should be omitted; where applicable, the following format can be used: “Author: John Smith”;

bibliography – it should be placed at the end of the document and have an appropriate title, for example, “Reference List” or “Bibliography”;

references – the subsequent references in the text should be numbered in the following way: “1. - 2. - 3.” or “[1] - [2] - [3]”; the text of every reference should contain a citation in a commonly used format, for example, “J. Biol. Chem., Vol. 234, No. 8, p. 1971/75, August 1959”; if the bibliography has not been published yet, the date of its current version should be entered, for example, “August 12, 2009”;

font type – type 3 fonts should be avoided because they are often generated with a missing or erroneous size and/or encoding, which makes it difficult for the Google tool to process the document; the font type can be checked in the “Properties” item of the “File” menu in the Adobe Acrobat Reader program.

How fast will materials published in a repository in the dLibra system be indexed by the Google Scholar web search engine?

There is no unequivocal, constant time for it – it all depends on the Google rules at the given moment and on how often Google updates the data from the repository. Having said that, properly prepared publications should become visible in a week or two. At the end of April and beginning of May 2013, we have made a small experiment to check that. For more information about that subject, see:

http://dlab.psnc.pl/2013/05/06/repozytorium-instytucjonalne-na-systemie-dlibra-i-google-scholar-maly-eksperyment/

I would like the materials from my repository to be visible in the Google Books index.. Does the dLibra system support it somehow?

The Google Books service is primarily for publishers interested in selling books. Moreover, it contains copies of scans the creation of which was financed by the Google company. Institutions which want to make their resources available in the Google Books index, should contact the Google company. Once the technical requirements and the principles of cooperation have been established (they are not available publicly), the Poznań Supercomputing and Networking Center can implement the necessary mechanisms in the dLibra system.

In my institution, there is a publishing house which publishes and sells (online) books, scripts, and journals created by the institution. Can I somehow take those resources in to account in the institutional repository?

The best solution would be to sign an agreement with the publishing house, to enable the transfer of publications to the digital library (and, possibly, after a time, an embargo since the moment of the publication of the object)..

If that is not possible, the simplest technical solution would be to use the mechanism of linking publications, available since version 5.5 of the dLibra system. It will allow the introduction of the following elements into the dLibra system:

- a description of the publication available on a web page of the publishing house,
- a link to the publishing house web page to which the user is to be redirected after having attempted to open the content of the publication in the dLibra system; and
- links to the publication files which the dLibra system should index for the purpose of full-text search.

If the publishing house website is configured in such a way that it will allow the dLibra system (for example, on the basis of a permanent IP address) to access, with the use of the HTTP protocol (that is, in a way similar to the way in which readers use a web browser), PDF, HTML, or other files with article texts, then it will be possible to search the digital library / repository for a publication on the basis of its description and content, and a reader trying to open the content will be redirected to the appropriate web page of the publishing house, where he or she will able to, for example, purchase the article.

The authors from my institution publish their works in international publishing houses, which make the articles available in return for payment. Can I somehow take those resources in to account in the institutional repository?

Publications from external publishing houses (including foreign ones) can be included by means of a process similar to the one described in response to the In my institution, there is a publishing house which publishes and sells (online) books, scripts, and journals created by the institution. Can I somehow take those resources in to account in the institutional repository?

Can information about publications the content of which is not available online, in any way, also be collected in the repository?

At this moment, the only form of a publications without content (and without a link to content) in the dLibra system are planned publications. Information about publications the content of which is not available online in any way should be collected in a separate bibliographic base. If there is a need to take such data into account in the digital library, the proposed way to do that is to import the metadata from the bibliographic base to the digital library and to add return links to the appropriate records in the base to those data. However, that practice may be disappointing for users who expect full access to a text and not only bibliographic data in a repository / digital library. That is why the recommended solution is linking from the bibliographic base to the repositor (off course, for those publications which are available in the repository), and not the other way round.

My institution employs a few thousand researchers. We are planning to collect all publications of the employees in the repository and, perhaps, adopt the open-access mandate policy. Will a few thousand users entering their works simultaneously to the system not cause a system overload?

In the case of large implementations of the dLibra system, the number of new publications reaches from a few thousand (the Digital Library of Wielkopolska) to tens of thousands (the Jagiellonian Digital Library). They are mainly scans of library resources, that is, publications which are a greater burden for the servers of the digital library than typical scientific articles (PDF files with a text layer and possibly a few figures). The following aspects may be taken into account when preparing an institutional repository for a great number of new publications entered by its users:

- The necessary server infrastructure should be ensured, which will use the scaling possibilities of the dLibra system (see 13. Scaling the dLibra System). [*] It may be especially important to put the service which will index the content and metadata of the documents on a separate server and to ensure fast functioning of the database used by the dLibra system.
- A copy of the website of the digital library should be prepared where the authors will be able to enter publications with the use of the self-archiving method. Such a copy of the website can be run on a separate server, and its functionality may be restricted to: accessing user accounts, adding new publications, and viewing the already added publications. In this way, even if the authors generate great traffic, readers will be using an independent portal.
- There should be an appropriate number of publication moderators who will supervise author groups and the publications entered by those authors (for example, supervisors of particular institutes). That may be achieved by assigning home directories and supervisors of those home directories to particular authors. Many authors can be assigned to one moderator, and many moderators can be assigned to one author group.
- An emergency solution for a possible system overload should be prepared. For example, if all researchers will be trying to enter their publications at the last minute (because of an instruction of the institution), the system may be overloaded regardless of any technical precautions taken in advance. It is worth having an emergency plan for such a situation – for example, having an appropriately large email box which may fulfill the function of a buffer when the self-archiving mechanism becomes unavailable.

A PDF file should be prepared in accordance with the following convention:

document title – it should be the greatest fragment of text, at the top of the page, in at least the 24-point font size; the same font should be used for the whole title, and all other texts on the page should be in a smaller font size than the title font size – otherwise, the other, greater text can be erroneously interpreted as the title;

document authors – they should be entered just after the title, in a slightly smaller font size which is, however, greater than standard text (so the size could be, for example, from the 16–26 point range); the same font should be used for all author names, and the font of the names should be greater than the font of section headings – otherwise, the other, greater text can be erroneously interpreted as author names; particular authors should be separated with commas or semicolons, and their affiliations, degrees, and certificates should be omitted; where applicable, the following format can be used: “Author: John Smith”;

bibliography – it should be placed at the end of the document and have an appropriate title, for example, “Reference List” or “Bibliography”;

references – the subsequent references in the text should be numbered in the following way: “1. - 2. - 3.” or “[1] - [2] - [3]”; the text of every reference should contain a citation in a commonly used format, for example, “J. Biol. Chem., Vol. 234, No. 8, p. 1971/75, August 1959”; if the bibliography has not been published yet, the date of its current version should be entered, for example, “August 12, 2009”;

font type – type 3 fonts should be avoided because they are often generated with a missing or erroneous size and/or encoding, which makes it difficult for the Google tool to process the document; the font type can be checked in the “Properties” item of the “File” menu in the Adobe Acrobat Reader program.

For detailed information about creating PDF files, see the page with the rules for creating PDF files for Google Scholar (in English).