The Digital Repository Integration with External Information Services[1]

Elena V. Kovyazina[1]

[1] Institute of Computational Modelling SB RAS, Krasnoyarsk, Russia, elena@icm.krasn.ru

Abstract. The evolution of society to the digital future has initiated major changes in the principles and approaches to science and education. One of the important aspects of moving along this way is open access to the results of scientific research, which can be implemented using technology and the open archives software platform. The implementation of open archives in the libraries of research and educational institutions requires solving many problems as the select and correcting of data schemes, the batch exchange of data between the library automation system and the digital archive, authentication and authorization, licensing for digital documents and collections, and etc. Some of these problems are relatively easy solved by web services that are actively developing in the process of creating a scientific communication infrastructure on the Internet and being its integral part. Their inclusion as part a digital archive, as well as exploring the possibilities and features of their functioning, is an actual task. Some services useful for libraries are considered, integration with which can be implemented using the DSpace 6.3, as well as the features of their configuration and operation.

Keywords: infrastructure of open science, digital repository, global network information services.

1 Introduction

The humanity evolution to a digital future has initiated major changes in the principles and approaches to science and higher education. The expansion of the Internet, the development of digital and information technology has created the conditions for a wide and open dissemination of knowledge. The international community is actively discussing the principles and technologies for the transition to open science and open education. New approaches for information work are created and actively developed, the concepts are transformed, and terminology is formed that is aimed at working with digital resources [1-4].

One of the important aspects of moving along this way is open access to the results of scientific research. Moreover, the results are interpreted in the broad sense of this concept - it is assumed that access will be open not only to texts - publications, reports, scientific notes, etc., but also to a variety of research data. This approach implies a gradual evolutionary unification of technologies for research digital libraries formed on a basis of institutional repositories and data centers [4].

As noted in [5], Russia is in the initial stage of the process of "digital transformation of science and education". However, certain steps in this direction have already been taken and the definition of directions and methods for further advancement is in demand. In the libraries of research and educational institutions, the technologies of open archives are becoming popular, allowing to solve many problems who not previously solved with the help of traditional library automation systems (SAB). In fact, a digital open archive is what the international library community is commonly called the research digital library (RDL), and the methods of working with documents in such an archive help library staff to perceive an electronic document not as a digitalized copy of a printed publication, but as a full-fledged digital object, as part of a conceptual model of a research information system [6]. In mind to the authors of the review [2], to ensure the digital transformation of science, each digital object must have FAIR (Findable, Accessible, Interoperable and Re-usable). The digital repository provides all the possibilities for matching the documents included in its composition with the specified requirements. It is implemented, as a rule, on one of the free software platforms, each of which has individual features that make it possible to solve specific problems relevant to a particular research institution [7]. When introducing open archive technologies into the work of a library of a research institution, it is necessary to solve many problems as the select and correcting of data schemes, the batch exchange of data between the library automation system and the digital archive, authentication and authorization, licensing for digital documents and collections, etc. Some of these problems are relatively easy solved by web

services that are actively developing in the process of creating a scientific communication infrastructure on the Internet and being its integral part. For this reason, their inclusion to a digital repository, as well as the study of the possibilities and features of their functioning, is an actual task. Unfortunately, until recently, a significant part of such services, for example, OCLC LC Name Authority Service, have not been Russified, and therefore can only be used in a limited way. However, the development of open archive technologies in non-English-speaking countries leads to evolutionary changes in this area. Connection to the external services of the digital repository of the Institute of Computational Modeling SB RAS was carried out on DSpace 6.3.

2 Integration of the repository with external web services

The digital archive based on DSpace is built on current network protocols, therefore it is well integrated into the infrastructure of the Internet. The issues of our consideration are external web services that require, as a rule, two-side actions - on the side of the service (registration with or without keys) and updating the internal settings of the digital repository. It should be noted that the list of external services under consideration is not exhaustive for the digital repository and includes only the most popular and relevant for libraries.

The weakness of many library automation system (SAB) is no indexing of content by Internet search engines. Common search engines, such as Google or Yandex, provide special services where you can specify the addresses of sites and their pages for indexing, for example, for Google https://search.google.com/search-console/. However, this way of indexing would open the site not only for crowler search engines, but also for bots, which increase the load on the repository. DSpace provides a set of rules to optimize search engine indexing of content. In particular, this is a daily generated sitemap, the links to the files of which are installed by default. This means that every time the sitemap index page registered at http://www.google.com/webmasters/sitemaps/ is updated, the corresponding script informs Google that the sitemap is updated, and the collection, metadata and bitstreams of a digital repository are accessible to the search engine. DSpace also blocks the known “bad” bots, a list of which is supported by Wikipedia (http://en.wikipedia.org/robots.txt), and which is updated with each update of the system.

Another weakness of the SAB is the licensing, that is, compliance with copyright and publishing limitations on access to documents submitted in a digital repository. Starting from v5 DSpace has been integrated with the database of publisher policies Sherpa/RoMEO, which allows observing publisher policies to access the document at the stage of submitting items in the repository. To use the service in full, you need to register on its website to receive a free API key, which then is included in the search requests of service. Service can be obtained at the request of ISSN, or the title of the journal at the time of filling out the metadata on the journal their paper is published, or use Sherpa/RoMEO autocomplete in authority control mode. The data obtained from the service allows you to determine the terms of the publishing embargo, enter it into the metadata value to ensure the copyright of the full text in automatic mode.

In addition to publishing requirements, DSpace provides an alternative for authors or other copyright owners to specify the desired restrictions on access to the full text and metadata of the submitted items. It can be specified using traditional copyright and attached to the metadata of a digital object as bitstream containing the license text. However, as an alternative, DSpace supports the Creative Commons licensing service (https://creativecommons.org/). Support for selected licenses is controlled by website configuration options, and each selected license includes an interactive link to the Creative Commons website. If this option is enabled, users can select a Creative Commons (CC) license during submitting the item in a digital repository or determine the absence of licensing. If you select licensing, metadata and, if desired, a copy of the license text is attached to the items in the repository, and there will be indication text and CC icon on the homepage of the item. To communicate with the service, The Creative Commons REST API is utilized. This allows DSpace to store metadata references to the selected CC license, while also storing the CC license as a bitstream. The following CC license information are captured [8]:

- The URL of the CC License is stored in the "dc.rights.uri" metadata field.
- The name of the CC License is stored in the "dc.rights" metadata field.
- The RDF version of the CC License is stored in a bitstream named "license_rdf".

Another important aspect of working with information is related to ensuring the stability of links to the source of the digital document. The development of the exchange of references in URLs teaches web users that sites can disappear or reconfigure without an announcement, then files containing research-critical results may become unavailable permanently or temporarily. To solve this problem, the basic property of DSpace was the creation of a persistent identifier for each digital object, collection, or community in the repository. To ensure persistence of DSpace identifiers, a mechanism for creating and managing identifiers independent of storage and localization is required. By default, DSpace uses the CNRI Handle System (http://www.handle.net/) to create them. For a moderate fee, the service provides a unique ‘prefix’, with the help of which identifiers of digital repository objects are generated. Typically, Handle identifiers are assigned to communities, collections, and items. Based on them, unique URIs of corresponding objects are formed. Although bundles and bitstreams are not marked with their own Handles, a unique URI is generated for each of them based on the corresponding identifier of the item, and thus, they can be referenced both as part of a larger object, and individually. The handle system also functions as a global permission infrastructure; thus, the end user can enter the Handle in any service (for example, a web page) that is capable of resolving them, and the end user will be redirected to the object (in the case of DSpace, community, collection or item object) identified by this handle.

Handle are just one of the options for global identification of digital objects. There are many different systems for persistent identification: Handle, DOI, urn: nbn, purl, and many more. DSpace allows the use of external identifier services different from Handle. The most popular in scientific organizations is the DOI identification service. DOIs are persistent identifier like Handle, but as many big publishing companies use DOIs, they are quite well-known to scientists. Some journals ask for DOIs to link supplemental material whenever an article is submitted. In DSpace, starting with version 4.0, you can use DOI in parallel to the Handle. By "using DOI" is meant the automatic generation, reservation and registration of DOI for every item that enters the repository. These newly registered DOIs will not be used as a means of building URIs for DSpace items [8]. Objects will still be identified by Handle.

To register a DOI, one has to enter into a contact with a DOI registration agency which is a member of the International DOI Foundation. There are several such agencies. Different DOI registration agencies have different policies. Some of them offer DOI registration especially or only for educational institutions, others only for publishing companies. Most registration agencies charge fees for registering DOI, and they all have different rules describing for what kind of item a DOI can be registered. DataCite is an international initiative to promote science and research, and a member of the International DOI Foundation. DataCite's members act as registration agencies for DOI. Some DataCite members provide their own APIs to reserve and register DOIs; others let their clients use the DataCite API directly. Starting with version 4.0, DSpace supports DOI administration by using the DataCite API directly or by using the EZID API (which is a service of the University of California Digital Library). This means you can administer DOIs with DSpace if your registration agency allows you to use the DataCite API directly or if the registration agency is EZID [8].

Another interesting service that can be used to authority control and determine the affiliation of authors is the ORCID service. Integration with it adds ORCID compatibility to all existing authority control solutions in DSpace. The string names of the authors are still stored in the DSpace metadata. The authority key field is used to store a uniquely created internal identifier that associates the author with his more advanced metadata, including the ORCID and alternative author names. ORCID authorities not only link a digital identifier to a name. It regroups a load of metadata going from alternate names and email addresses to keywords about their works and much more. The metadata is obtained by querying the ORCID web services. In order to avoid querying the ORCID web services every time, all these related metadata are gathered in a "reputable metadata cache" that DSpace can access directly [8]. The extensive capabilities provided by ORCID services are limited by the lack of their Russification; therefore, LDAP protocols are more often used in Russia to generate author metadata for CRIS systems [7].

To intensify the creation of digital archive content, a useful function is to import data from external sources. For importing DSpace offers several different methods, the most promising of which is importing by the REST API. Mainly, in this context we are talking about importing descriptive metadata of items, as well as their full texts, if they are placed in the public domain. For importing, metadata mapping between different data schemes are used, comparing the values of the metadata fields of the imported records with the corresponding meaning fields of the DSpace internal representation. As a format for downloading records from external sources, xml is used. Relatively simple way of importing due to the conformity of the schemes and data formats used by them is to import documents from foreign information resources. The practice of joint work of the international community with digital repositories has led to the accumulation of metadata mapping for the most common information resources - arXiv, PubMed, CrossRef, CiNii, etc. If necessary, the imported data is processed and can be visually presented in one of the bibliographic formats. Domestic SABs are configured for MARC formats, they rarely provide an API for importing and do not have external importing services, therefore transferring records from SAB to a digital archive requires the formation of their own metadata mapping and handlers.

And finally, the generated digital archive should be, at a minimum, registered by the services whose prerogative it is. In particular, the service who registers open access digital repositories is OpenDOAR, which records all information about the digital archive. Registration is free, but it takes some time associated with testing a digital resource, since it must meet certain requirements of the registration service:

1. It must be continuously available.
2. It cannot be an electronic journal.
3. It must not contain closed materials.
4. It cannot be only bibliographic, but must contain full texts.
5. It cannot be a library catalog or a collection of locally available e-books.
6. Access to the resource should not be limited by password or registration.
7. Should not be a proprietary database or subscription resource.

First level headings are all flush left, initial caps, bold and in point size 12. One-line space before the first level heading and 1/2-line space after the first level heading.

2 Conclusion

Distribution of digital repositories formed in open archive technologies is an important step towards open science. The repository provides global identification of content, the ability to combine text and data in a consolidated digital object, the safety of documents and their licensing. Integration of the repository with rapidly developing systems and Internet services is a guarantee of embedding the contents of the digital archive in the infrastructure of scientific communication.

References

[1] Castelli D.,Manghi P., Thanos C. A vision towards Science Communication Infrastructures // International Journal on Digital Libraries. 2013. V. 13. Issue 3-4. P. 155-169. https://doi.org/10.1007/s00799-013-0106-7 (accessed:21.10.2019).

[2] Ayris P., Ignat T. Defining the role of libraries in the Open Science landscape: a reflection on current European practice // Open Information Science. 2018. V. 2. Issue 1. P. 1-22. https://doi.org/10.1515/opis-2018-0001 (accessed: 20.10.2019).

[3] Charalampous A.,Knoth P. Classifying Document Types to Enhance Search and Recommendations in Digital Libraries // 21st International Conference on Theory and Practise of Digital Libraries (TPDL). Thessaloniki, Greece. 2017. P. 181-192. https://arxiv.org/pdf/1707.04134.pdf (accessed: 20.10.2019).

[4] Hidalgo Y.,Ortiz E., Febles J.P. A Method for Integrating Bibliographic Data from OAI-PMH Data Providers // IEEE Latin America Transactions. 2017. V. 15, № 9. P. 1695-1698. https://ieeexplore.ieee.org/document/8015075 (accessed: 20.10.2019).

[5] Качан Д.А., Богатко А.В., Богатко И.Н., Енин С.В., Кулаженко В.Г., Лазарев В.С., Лис П.А., Скалабан А.В., Юрик И.В. Интеграция информационных ресурсов открытого доступа для обеспечения научно-образовательного процесса в учреждениях высшего образования. // Открытое образование. 2018. 22(4). С. 53-63. https://doi.org/10.21686/1818-4243-2018-4-53-63 (accessed: 21.10.2019).

[6] Fedotov A.M., Fedotova O.A. Reference model of the scientific digital library // XVI Российская конференция «Распределенные информационно - вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции. Новосибирск: ИВТ СО РАН, 2017. С.393-409. http://elib.ict.nsc.ru/jspui/bitstream/ICT/1467/111/paper52.pdf (accessed: 21.10.2019).

[7] Fedotova O.A., Fedotov A.M., Zhizhimov O.L., Sambetbayeva M.A. Digital repository for research and education information systems // Труды ГПНТБ СО РАН. 2019. № 3. С.23-28. https://doi.org/10.20913/2618-7515-2019-3-23-28 (accessed: 20.10.2019).

[8] DSpace 6.x. Documentation. https://wiki.duraspace.org/display/DSDOC6x/ (accessed: 20.10.2019).