Curation of bibliographic metadata of the institutional repository on the Invenio-JOIN^2 platform

I.A. Filozova, T.N. Zaikina, G.V. Shestakova, R.N. Semenov

Joint Institute for Nuclear Research, Dubna, Moscow Region, Russia

Abstract. Content filling of the institutional repository and keeping the entered data "up to date" is a very resource-intensive task that requires organizing the coordinated actions of operators to enter data into an information system (IS). To resolve one helps the curation of bibliographic metadata — a set of actions and measures aimed for updating, managing and preserving digital objects throughout their life cycle in educational and the scientific interests of the community. This work considers the issues of bibliographic descriptions curation of publications by JINR (Joint Institute for Nuclear Research) employees, their enrichment of metadata entered into the JINR institutional repository from external sources: the Scopus bibliographic and abstract database, the Web of Science search Internet platform, the information platform in High Energy Physics INSPIREHEP. The development of information services for solving the problem of current accounting of the publication activity of JINR staff is described.

Keywords: *Open Access (OA), Institutional Repository (IR), CRIS & OAR, software platform Invenio-JOIN^2, Utility, Application Programming Interface (API), Authority Record, Publication Activity.*

DOI: 10.14357/20790279230102

Introduction

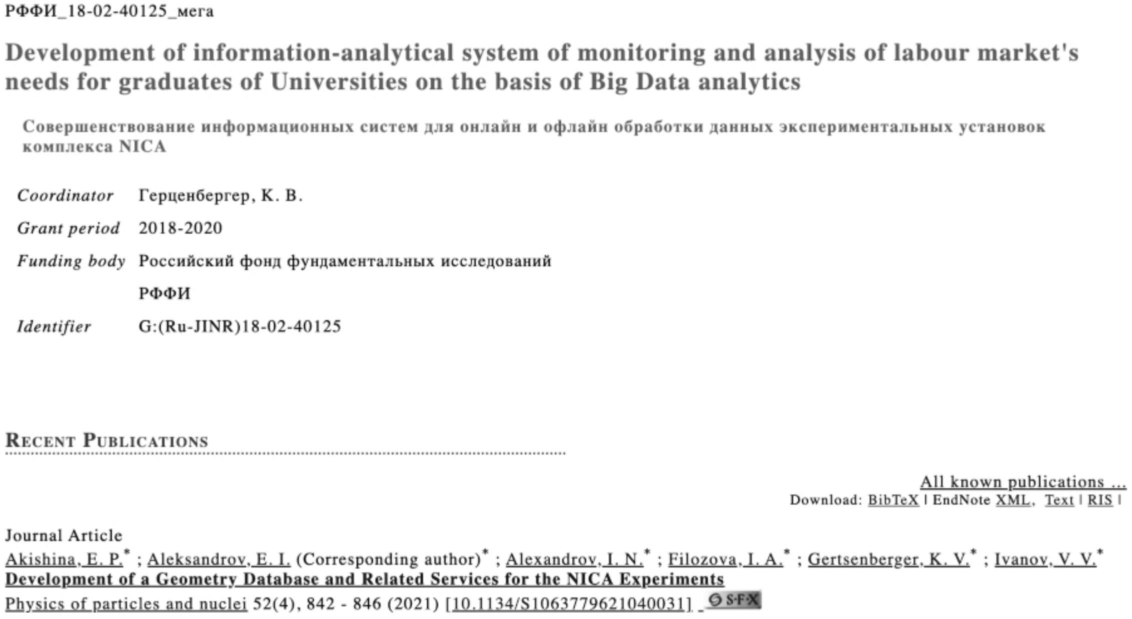
Keeping the content of the institutional repository up to date is an important task that requires significant time and human resources. The quality of the content of any information system is one of the key factors that makes it attractive to end users and able to meet their changing information needs. The way to organize such laborious work is to ensure the process of continuous curation of bibliographic metadata (the main digital objects of such kind information systems) and include a set of actions and measures aimed at updating, managing and preserving this metadata throughout their entire life cycle. The tasks of curation include: search and analysis of missing bibliographic metadata (not included in the repository at the current time for some reason); search and analysis of missing metadata of bibliographic descriptions uploaded to the repository; tracking changes of the publication's status (made earlier preprint can be transformed into an article published in a peer-reviewed journal); detection of input errors; updating user account data; updating data on the structural divisions of the organization, etc.

Currently, JINR is implementing an institutional repository based on the Invenio-JOIN^2 software platform [1; 2]. Closed beta testing is underway now.

This paper considers the issues of identifying missing bibliographic descriptions of JINR staff publications, adding bibliographic descriptions of reference books (grants, experiments, persons); enrichment of metadata (their clarifications and additions) entered into the JINR institutional repository with data from external sources: the bibliographic and abstract database Scopus, the information platform in the field of high energy physics INSPIREHEP [3]. These tasks can be partially automated. As a solution, a set of utilities and services has been developed to facilitate the implementation of some specialized standard curation processes for the JINR Publications Server.

1. Approaches and Practices for the Curation of Institutional Repositories

Most literature on the research topic presents the implementation of IRs from the viewpoint of end users. Thus the description is restricted by the front-end component of such IS, that is only "iceberg tip" and does not give a complete picture of the complexity of such information systems. Detailed structured descriptions of activ-



РФФИ_18-02-40125_мега

Development of information-analytical system of monitoring and analysis of labour market's needs for graduates of Universities on the basis of Big Data analytics

Совершенствование информационных систем для онлайн и офлайн обработки данных экспериментальных установок комплекса NICA

Coordinator	Герценбергер, К. В.
Grant period	2018-2020
Funding body	Российский фонд фундаментальных исследований РФФИ
Identifier	G:(Ru-JINR)18-02-40125

RECENT PUBLICATIONS

All known publications ...
Download: BibTeX | EndNote XML, Text | RIS |

Journal Article
Akishina, E. P.* ; Aleksandrov, E. I. (Corresponding author)* ; Alexandrov, I. N.* ; Filozova, I. A.* ; Gertsenberger, K. V.* ; Ivanov, V. V.*
Development of a Geometry Database and Related Services for the NICA Experiments
Physics of particles and nuclei 52(4), 842 - 846 (2021) [10.1134/S1063779621040031]

Fig. 1. Authority record of grant "РФФИ 18-02-40125_мега"

ities related to the management of research data, the needs and actions of various categories of users, used tools are a good basis for developing best practice guidelines and infrastructure templates for IRs [4]. These knowledge tools can then be used by institutions that are currently implementing institutional repositories. In many respects, the approaches and practices of IRs content curation depend on their specifics and the specifics of the institution that holds the content [5]. The resources management of an institutional repository includes the task of curation [6; 7]. The JOIN^2 project pays a lot of attention to the issue of curation. All instances provide updating authority records Persons and Institutes in automatic or semiautomatic mode [2]. One more example: curation process to enrich data in JOIN^2 repository with Metadata from PubMed. The script produces several output-files that are uploaded in batch and make updating records [2]. The implementation of curation processes depends on the set of protocols and authentication tools used by the institution. The project ISTINA [8] uses back-end algorithms to find sustainable teams of authors, research performance evaluation, authorship disambiguation and so on. The paper [4] describes data curation and use activities in IRs, their structures, roles played, skills needed, contradictions and problems present, solutions sought, and applied approaches.

2. Utilities and Services for JDS-JOIN^2 Prototype

To solve the above tasks, the following set of useful utilities has been developed: New Grants, People&Institutes Corrections, Inspire Corrections, HAC List. Services are implemented as a backend solution for the JINR institutional repository and are executed as server processes with a given frequency — tasklets. Publications Server is carried out as an Open Access archive of JINR scientific output [9].

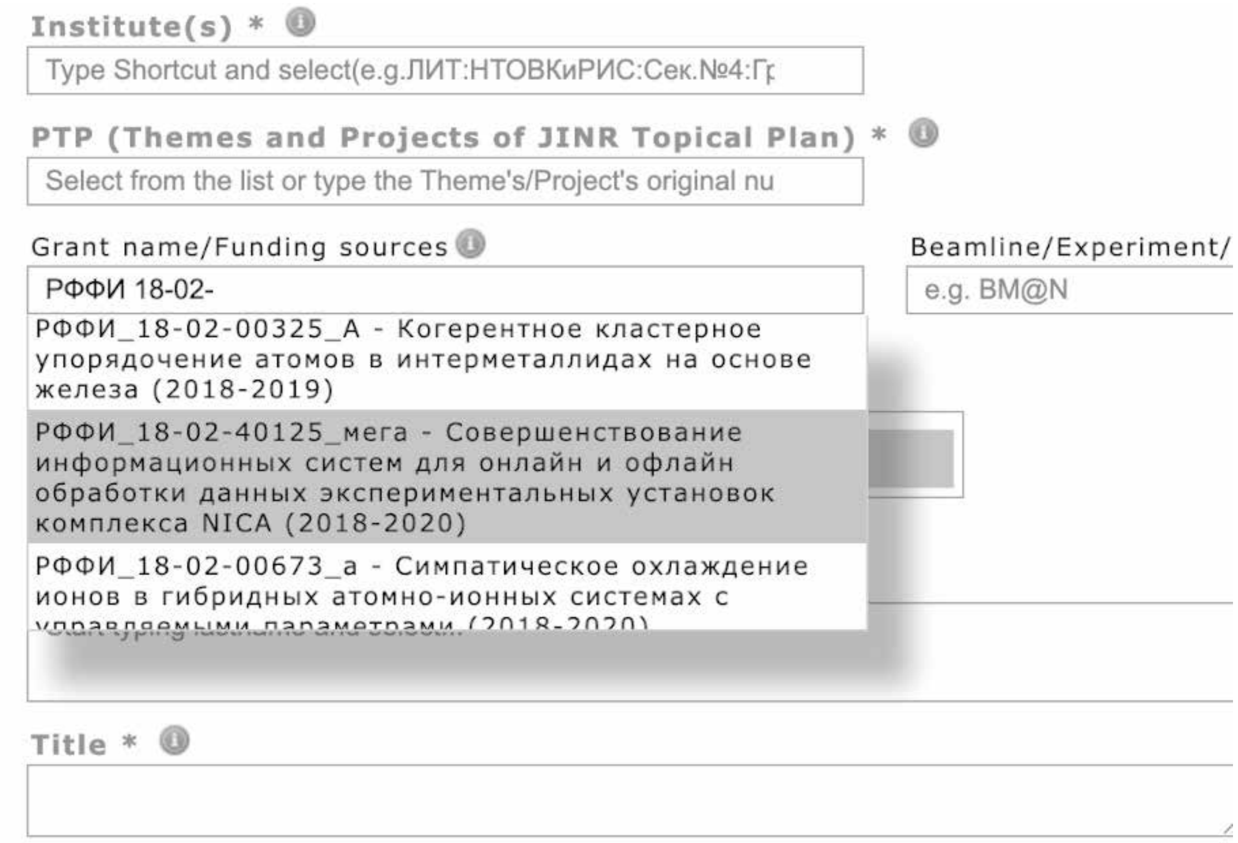
2.1. Utility New Grants

To reflect the information about grant support of the research work, resulting or describing in a publication, it is necessary to add the appropriate metadata to bibliographic description of this publication. Grant metadata is stored in the Authority Collection Grants (fig. 1) – catalog of research funding sources (JINR Themes/Projects, RFBR and RSF grants).

The Authority record of the Grant "РФФИ 18-02-40125_мега" contains the title, coordinator name, grant period, as well as a list of publications loaded into the system and linked (logical connection of the publication belonging to this funding source) to it. To specify the relationship between publication and Grant, the submitter should enter the necessary metadata in publication through the web-Submit tool (fig. 2).

The input subsystem, using the import module by the values of digital identifiers (in this case, DOI), recognizes public bibliographic metadata and distributes them among the form fields. In the Grant name/Funding sources section, the user selects the desired grant from the drop-down list (it must first be loaded into the system).

Information about JINR projects, funded by the Russian Foundation for Basic Research and the



Institute(s) *

Type Shortcut and select(e.g.ЛИТ:НТОВКиРИС:Сек.№4:Гр

PTP (Themes and Projects of JINR Topical Plan) *

Select from the list or type the Theme's/Project's original nu

Grant name/Funding sources

РФФИ 18-02-

РФФИ_18-02-00325_А - Когерентное кластерное упорядочение атомов в интерметаллидах на основе железа (2018-2019)

РФФИ_18-02-40125_мега - Совершенствование информационных систем для онлайн и офлайн обработки данных экспериментальных установок комплекса NICA (2018-2020)

РФФИ_18-02-00673_а - Симпатическое охлаждение ионов в гибридных атомно-ионных системах с управляемыми параметрами (2018-2020)

Beamline/Experiment/

e.g. BM@N

Title *

Fig. 2. Web-form for entering a publication

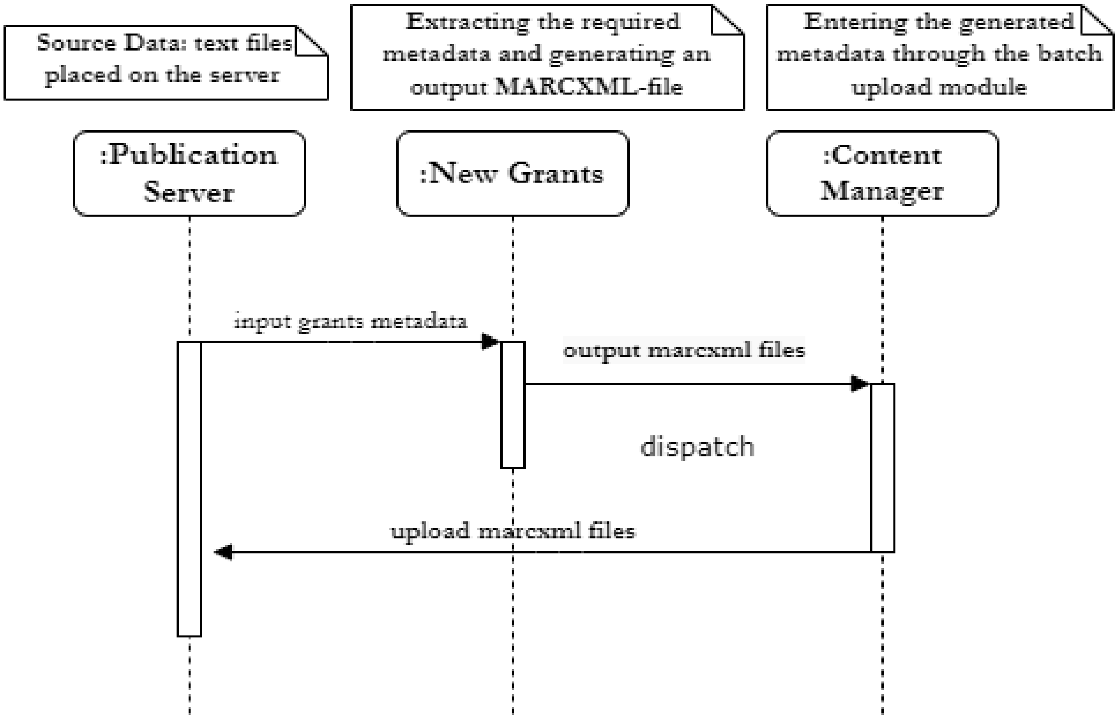

Source Data: text files placed on the server
Extracting the required metadata and generating an output MARCXML-file
Entering the generated metadata through the batch upload module
:Publication Server
:New Grants
:Content Manager
input grants metadata
output marcxml files
dispatch
upload marcxml files

Fig. 3. Interaction of the server, module New Grants and content manager of institutional repository in the scenario of generating and loading metadata about new grants

Listing 1. Example Authority Record Grant РНФ 19-75-20121

```
<record>
<datafield ind1=”7” ind2=” “ tag=”024”>
<subfield code=”a”>G:(Ru-JINR)19-75-20121</subfield>
<subfield code=”0”>I:(Ru-JINR)_400000</subfield>
<subfield code=”d”>19-75-20121</subfield>
</datafield>
<datafield ind1=” “ ind2=” “ tag=”035”>
<subfield code=”a”>G:(Ru-JINR)19-75-20121</subfield>
</datafield>
<datafield ind1=” “ ind2=” “ tag=”150”>
<subfield code=”a”>Новые гибридные и углеродные аэрогели - синтез и анализ структуры методами малоуглового рассеяния (на ИБР-2)</subfield>
<subfield code=”y”>2019-2022</subfield>
</datafield>
<datafield ind1=” “ ind2=” “ tag=”371”>
<subfield code=”a”>Горшкова Ю.Е.</subfield>
<subfield code=”0”>P:(Ru-JINR)P002369</subfield>
</datafield>
<datafield ind1=” “ ind2=” “ tag=”450”>
<subfield code=”a”>РНФ 19-75-20121 Проведение исследований на базе существующей научной инфраструктуры мирового уровня</subfield>
<subfield code=”y”>2019-2022</subfield>
<subfield code=”w”>d</subfield>
</datafield>
<datafield ind1=”1” ind2=” “ tag=”510”>
<subfield code=”a”>Российский научный фонд</subfield>
<subfield code=”0”>I:(DE-588b)1228775-1</subfield>
<subfield code=”b”>РНФ</subfield>
</datafield>
<datafield ind1=” “ ind2=” “ tag=”980”>
<subfield code=”a”>G</subfield>
</datafield>
<datafield ind1=” “ ind2=” “ tag=”980”>
<subfield code=”a”>AUTHORITY</subfield>
</datafield>
</record>
```

Russian Science Foundation is the source of data for the New Grants module. This information is presented on the official website of JINR (http://www.jinr.ru/grants/) [11]. The New Grants module receives a text file with data about grants as input, and generates an output file in MARCXML format, which is loaded into the system in batch loading mode by the content manager of the JINR publications server (fig. 3).

This utility implements the following functions:

- Extract grant metadata from source file.
- Checking for duplicates (whether there is an entry with an identical number in the Grants authority records collection).
- Generation of a marcxml file with a description of the grant for subsequent loading into the system (see Listing 1. — an example of the description of the Grant authority record for the RSF project 19-75-20121 in marcxml format).

Using the New Grants utility, 270 records of RFBR grants for the period 2013-2022 and 60 entries about RSF projects for the period 2014-2022 were uploaded to the JINR Publications Server.

2.2. Service People&Institutes Corrections

This application updates the authority records of People and Institutes – collections – directories of JINR employees (potential authors of scientific publications) and JINR structural divisions of all levels of the hierarchy, respectively (fig. 4).

The main functions of *People&Institutes Corrections*:

- Updating data about working employees: modifying email, adding SSO login, adding identifiers of various external author identification systems like Inspire, ORCID, Scopus (if available).
- Deleting authority records of non-working (dismissed) employees.
- Making authority records of new (hired) employees.

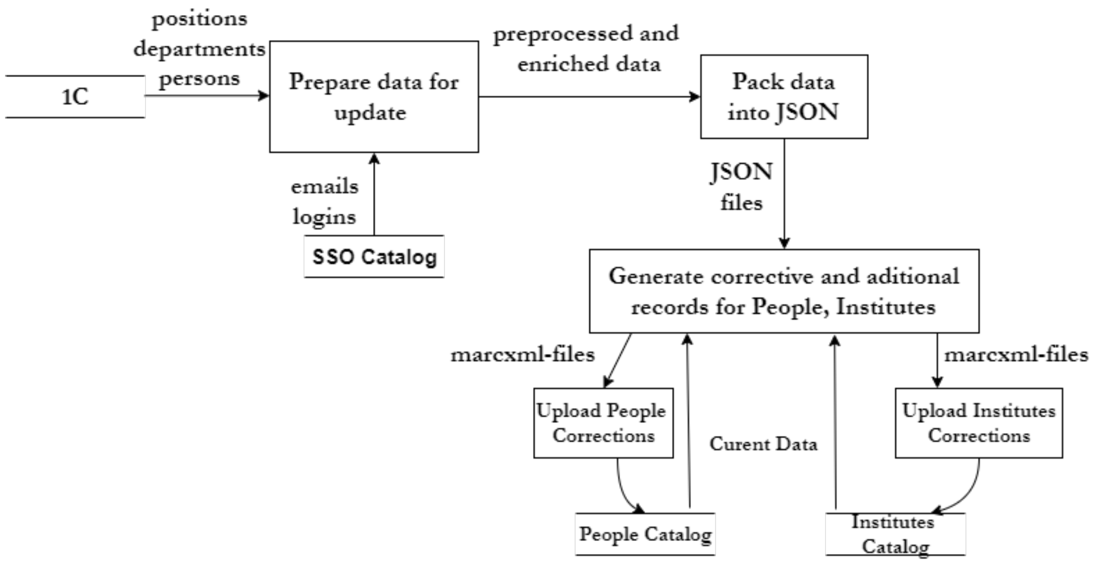


Fig. 4. Update People and Institutes Catalogs

- Updating data when an employee moves from one department to another.
- Updating data on current structural divisions (renaming, transferring to another level of hierarchy, etc.).
- Deleting authority records of liquidated structural divisions.
- Entering authority records of new (newly created) structural units.

Deletion of the record happens after attaching a specific tag, the record becomes inactive (invisible to the user) but is not destroyed from the database.

This service is currently running weekly. On average, the volume of corrective entries is: 25-35 updates (including dismissed and hired people) per week, 3-4 updates for departments per month. When the institutional repository is set to production mode, the service will be launched daily.

2.3. Service Inspire Corrections

INSPIRE (https://inspirehep.net/) is the trusted information hub for the high energy physics community [3]. It serves as a one-stop information platform for the HEP community, including 8 interconnected databases of literature, conferences, institutions, journals, researchers, experiments, jobs, and data. INSPIREHEP works in collaboration with CERN, DESY, Fermilab, IHEP, IN2P3 and SLAC and has been serving the scientific community for nearly 50 years.

Experience has shown that reports on publication activity often lack publications for a given reporting period from subject repositories such as INSPIREHEP. This information platform has collected more than 35 thousand publications of JINR employees for the period from 1956 to the present (fig. 5).

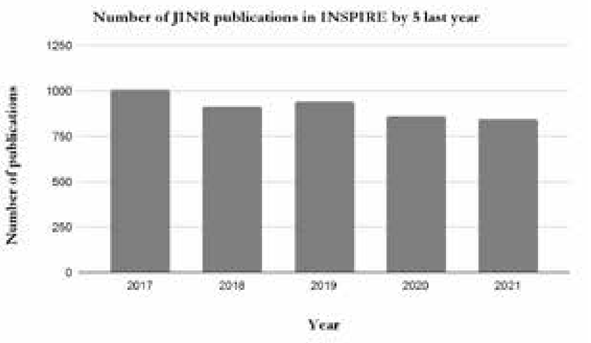
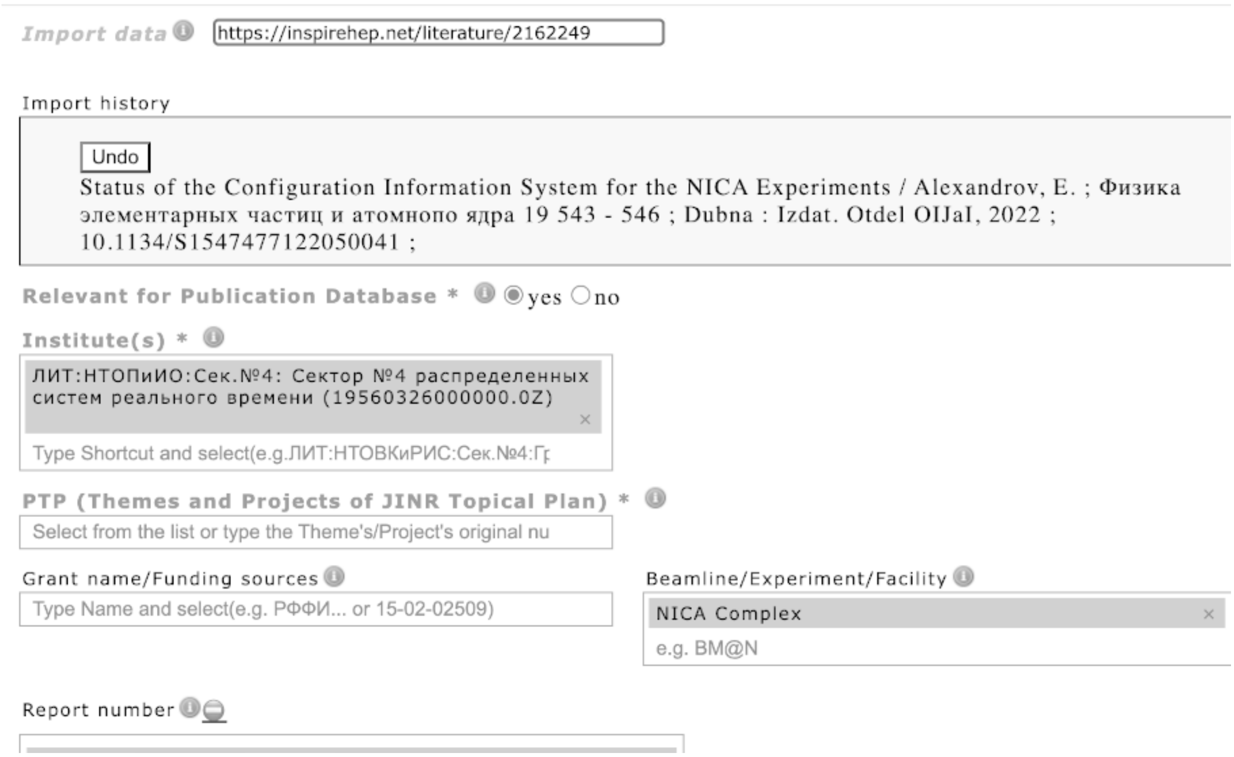


Fig. 5. JINR Publications in INSPIREHEP (2017-2021)

Over the past five years, the number of publications of JINR staff members stored in INSPIREHEP has been at the level of 850 ÷ 1000 per year.

To identify missing (missing descriptions in the institutional repository) INSPIREHEP publications, the *Inspire Corrections* service was developed, which performs the following functions:

- Retrieve bibliographic metadata from the INSPIREHEP hub using the public API (https://github.com/inspirehep/rest-api-doc) for a specified period, as an example: for the last month (configurable options).
- Duplicate Check: Checks for the publication metadata retrieved from INSPIREHEP and stored at the JINR Publication Server.
- Sending an email notification to the co-author (or responsible person performing archiving by proxy) about unloaded INSPIREHEP publications with references and INSPIRE's IDs. Using this information, an employee can easily submit these publications to the JINR Publication Server by using the *Import Data* function of the web-submit module (fig. 6).



Import data https://inspirehep.net/literature/2162249

Import history

Undo

Status of the Configuration Information System for the NICA Experiments / Alexandrov, E. ; Физика элементарных частиц и атомнопо ядра 19 543 - 546 ; Dubna : Izdat. Otdel OIJaI, 2022 ; 10.1134/S1547477122050041 ;

Relevant for Publication Database * ● yes ○ no

Institute(s) *

ЛИТ:НТОПиИО:Сек.№4: Сектор №4 распределенных систем реального времени (19560326000000.0Z) ×

Type Shortcut and select(e.g.ЛИТ:НТОВКиРИС:Сек.№4:Гр

PTP (Themes and Projects of JINR Topical Plan) *

Select from the list or type the Theme's/Project's original nu

Grant name/Funding sources

Type Name and select(e.g. РФФИ... or 15-02-02509)

Beamline/Experiment/Facility

NICA Complex ×

e.g. BM@N

Report number

Fig. 6. Web-form for new submissions

2.4. Utility HAC List

The experience of pre-production prototype of institutional repository “JINR Publications Server” showed that one of the frequent types of user queries is the search for publications in journals registered in various international and national catalogs and databases (Database Coverage). For example, the search of publications published in the Directory of Open Access Journals (DOAJ) indexed by Scopus, or journals included in the List of Higher Attestation Commission a list of leading peer-reviewed scientific journals included by the Higher Attestation Commission of the Russian Federation, recommended for publishing the main scientific results of a dissertation for the degree of candidate and doctor of science) [12].

The functionality of the Invenio-JOIN^2 software platform, on which the JINR Publications Server institutional repository prototype is deployed, allows the filtering of publications described above. To do this, the Statistics key – special Authority Record, should be created and associated with the necessary journal. Then all publications of this journal entered into the system will be indexed as belonging to this Statistics key, and can be found by the search attribute StatID = value. Authority Record with internal value **2002** was added to the Statistics keys catalog. The aim of it is to indicate that the article is/was published in a journal from the LIST of the peer-reviewed scientific publications, approved by the Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation.

The HAC List published at the official website in a pdf file (the latest up-to-date version – 04/27/2022 includes 2679 journal titles). The file contains a table with columns: *No.; Name of the journal; ISSN; Scientific specialties and branches of science corresponding to them, in which academic degrees are awarded; The date of inclusion of this journal in the List.* The content of the pdf-file in its original form is not suitable for machine processing. An application was implemented

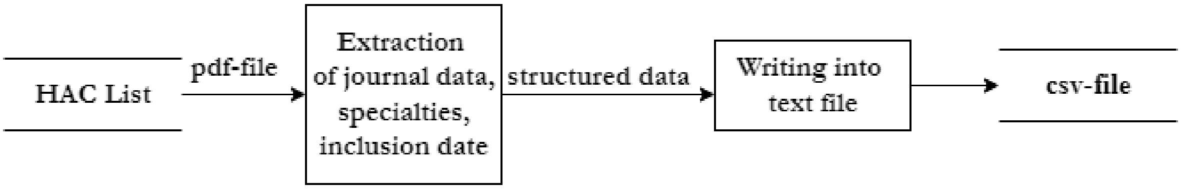


Fig. 7. Extraction of structured data from source HAC List

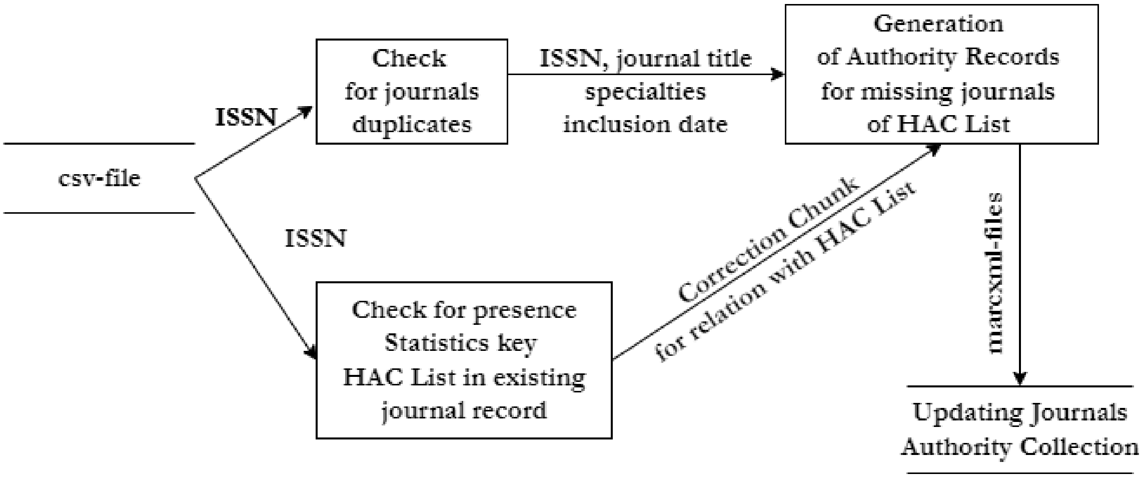

csv-file
ISSN
Check for journals duplicates
ISSN, journal title specialties inclusion date
Generation of Authority Records for missing journals of HAC List
ISSN
Check for presence Statistics key HAC List in existing journal record
Correction Chunk for relation with HAC List
marcxml-files
Updating Journals Authority Collection

Fig. 8. Indication journals records by statistics key HAC List

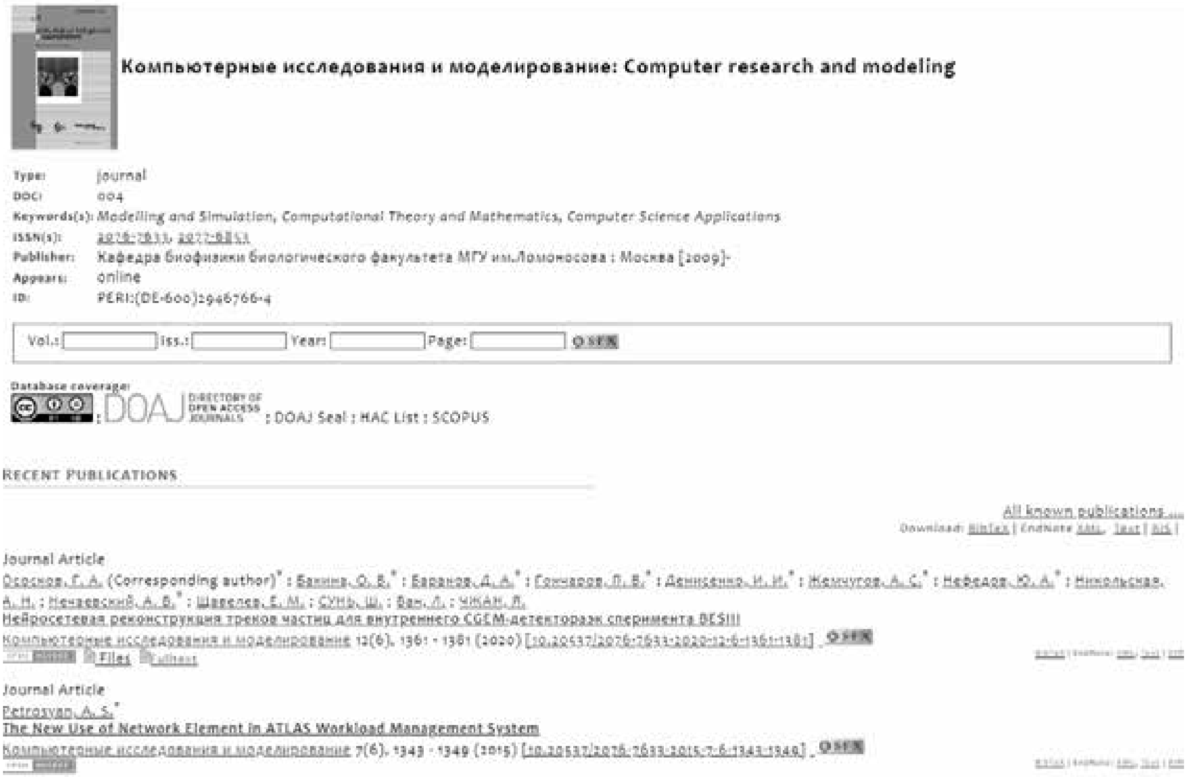
that extracts data from this pdf-file and structures the data according to the above columns, and generates a text csv file (fig. 7).

The resulting csv file is fed to the input of another application execute the following functions:

- Checking the journal for duplicates by ISSN: whether there is a record with an identical ISSN value in the Journals authority collection.
- Checking if an authority record has a *HAC List* Statistics key label, which means that the journal is identified in the system as belonging to the HAC List.
- Creating the corrective chunk with *HAC List* Statistics key in its absence**.**
- Generation of journals authority records metadata absened in a system, in marcxml format for uploading with Statistics key HAC List**.**

Fig. 8. illustrates it.

An example of this approach is presented in fig. 9. The authority record of the journal *Computer research and modeling* displays a list of publications published in it. The Database Coverage block displays the *HAC List* label.



Компьютерные исследования и моделирование: Computer research and modeling

Type: journal
DOC: 004
Keywords(s): *Modelling and Simulation, Computational Theory and Mathematics, Computer Science Applications*
ISSN(s): 2076-7633, 2077-6853
Publisher: Кафедра биофизики биологического факультета МГУ им.Ломоносова : Москва [2009]-
Appears: online
ID: PERI:(DE-600)2946766-4

Vol.: ____ Iss.: ____ Year: ____ Page: ____

Database coverage:
; DOAJ Directory of Open Access Journals ; DOAJ Seal ; HAC List ; SCOPUS

RECENT PUBLICATIONS

All known publications ...
Download: BibTex | EndNote XML, Text | RIS |

Journal Article
Ососков, Г. А. (Corresponding author)* ; Бакина, О. В.* ; Баранов, Д. А.* ; Гончаров, П. В.* ; Денисенко, И. И.* ; Жемчугов, А. С.* ; Нефедов, Ю. А.* ; Никольская, А. Н. ; Нечаевский, А. В.* ; Щавелев, Е. М. ; СУНЬ, Ш. ; Ван, Л. ; ЧЖАН, Я.
Нейросетевая реконструкция треков частиц для внутреннего CGEM-детекторазк сперимента BESIII
Компьютерные исследования и моделирование 12(6), 1361 - 1381 (2020) [10.20537/2076-7633-2020-12-6-1361-1381]
Files Fulltext

Journal Article
Petrosyan, A. S.*
The New Use of Network Element in ATLAS Workload Management System
Компьютерные исследования и моделирование 7(6), 1343 - 1349 (2015) [10.20537/2076-7633-2015-7-6-1343-1349]

Fig. 9. Authority record of journal *Computer research and modeling*

Conclusion

Updating the content of an institutional repository is a complex and multifaceted task that requires continuous curation by the accompanying staff. Curation is organized on the basis of typical business processes, some of which can be automated. Automation consists in the implementation of a set of specialized utilities (auxiliary scripts that complete the functionality of the software platform in order to perform typical curation tasks) and information services. The work shows the usage of this functionality on the JINR Publication Server as an example. In particular, the *New Grants* Utility*,* Service *People&Institutes Corrections,* Service *Inspire Corrections* have been implemented and included in the pre-production version of the system. The Utility *HAC List* is implemented with limited functionality and is still under development. The relevance to continue the development of *Scopus Corrections* Service with functionality similar to *Inspire Corrections* Service is discussed.

References

1. The official web-site Invenio project. Available at: https://invenio-software.org/ (accessed November 23, 2022).
2. The official web-site JOIN² Project. Available at: https://join2.de (accessed November 23, 2022).
3. Information platform for HEP community. Available at: https://inspirehep.net/ (accessed November 23, 2022).
4. *Lee DJ, B. Stvilia.* 2017. Practices of research data curation in institutional repositories: A qualitative view from repository staff. PLoS ONE 12(3): e0173987. Available at: https://doi.org/10.1371/journal.pone.0173987 (accessed November 23, 2022).
5. *Kidney A., C. R. Borges, D. Molodenskiy, et al.* 2020. SASBDB: Towards an automatically curated and validated repository for biological scattering data. Protein science 29(1): 66 – 75. doi:10.1002/pro.3731.
6. *Redkina N.S.* 2022. The libraries and Open Science: vectors of interaction. Scientific and technical libraries 3:105–126. doi:10.33186/1027-3689-2022-3-105-126. (In Russian).
7. *Redkina N.S.* 2019. Modern trends in research data management. Scientific and technical information. Series 1: Organization and methodology of information work 4:1–7. (In Russian).
8. *Afonin S.A., and others.* Ed. Academician V.A. Sadovnichiy. 2014. Intellectual system of thematic research of scientific and technical information. M.: Moscow University Press. 262p. (In Russian).
9. JINR Publications Server, Available at: https://publications.jinr.ru (accessed November 23, 2022).
10. *Filozova Irina,Tatiana Zaikina, Galina Shestakova, Roman Semenov, Martin Köhler, Alexander Wagner, Laura Baracci on behalf of the JOIN² project.* 2020. JINR Open Access Repository based on the JOIN² Platform. Proceedings of the Data Analytics and Management in Data Intensive Domains 2020. CEUR Workshop Proceedings, 2790:142-155. Available at: http://ceur-ws.org/Vol-2790/paper14.pdf (accessed November 23, 2022).
11. Materials of the JINR official website. Available at: http://jinr.ru (accessed November 23, 2022).
12. Official web-site of the Higher Attestation Commission under the Ministry of Science and Higher Education of the Russian Federation. Available at: https://vak.minobrnauki.gov.ru/ (accessed November 23, 2022).

I.A. Filozova. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: fia@jinr.ru

T.N. Zaikina. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: ztanya@jinr.ru (correspondent author)

G.V. Shestakova. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: shestakova@jinr.ru

R.N. Semenov. Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia, e-mail: roman@jinr.ru