DALIA: An Open Data Repository for the Open Digital Products of the Bologna Cnr Research Area Library

Gabriela Carrara [1,2], Silvana Mangiaracina [1], Debora Mazza [1], and Alberto Candiani [1]

[1] *CNR Biblioteca Area territoriale di Ricerca di Bologna, Via Gobetti 101, 40129 Bologna, Italy*
[2] *CNR IMM-BO, Via Gobetti 101, 40129 Bologna, Italy*

Abstract

The Bologna Research Area Library of the National Research Council is very active in many fields such as educational projects, services and support scientific research.
NILDE (Network for Inter-Library Document Exchange), a software developed by the Bologna Library and adopted by a vast community of Italian and foreign libraries is one of the main projects carried out by the Library, started in 2001 and still active. More than 20 years of activity have brought the NILDE project to produce a large amount of digital data (multimedia materials from conferences, video interviews, multimedia teaching resources...) that need to be preserved and managed. Faced with the demand of consultation and reuse of these materials by the NILDE community, consisting of librarians, researchers and students for their daily work, the Bologna Research Area Library devised a simple way to organize the digital data produced by the projects in which it has been involved following the Open Science principles. The open data repository DALIA, accessible via web, was built to the purpose making use of the well proven open source tool CKAN.
The archive has been populated starting from the data belonging to the NILDE project. The tests carried out with DALIA suggest its potentiality of a new service to be proposed to the Bologna Research Area researchers in order to introduce them to the procedures for archiving, describing and preserving data as suggested from EU.

Keywords

NILDE, resource sharing, inter-library loan, document delivery, Open Science, open data

1. Introduction

The Bologna Research Area Library (known as BdA) of the National Research Council (CNR) is a multidisciplinary science library which was created in 1995 to collect, integrate, enhance and disseminate the scientific bibliographic heritage of the institutes that merged into the CNR Research Area in Bologna[1]. The BdA plays a role of support and technological development for accessing scientific documentation and its free circulation and has promoted the emergence of innovative services for users of academic and research libraries in Italy. The BdA was and currently is involved in several national and international projects, the main among them being NILDE (Network for Inter-Library Document Exchange) [1]. NILDE is a web-based software for libraries and end-users, developed from 2001 by the BdA [2,3], which is currently used by more than 900 university, public research and health libraries in Italy, Spain and Greece, totaling more than 85,000 end-users (mainly researchers and students) registered through their libraries. A collaborative network of libraries using the NILDE

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy
gcarrara@area.bo.cnr.it (G. Carrara); mangiaracina@area.bo.cnr.it (S. Mangiaracina)
0000-0002-5172-0820 (G. Carrara); 0000-0003-0717-1227 (S. Mangiaracina)

[1] The CNR Bologna Research Area is a research campus that hosts seven CNR research institutes, two INAF (Italian Astrophysics National Institute) research institutes, one Tecnopolo Laboratory to support innovation of Emilia Romagna companies, as well as common facilities to support the research activities, such as the central library and the congress center.

software has formed, to mutually and freely share their bibliographic resources: the NILDE community, consisting of about 2.000 professional librarians who use the system daily, and their end-users. Since 2003, under the aegis of the BdA, the NILDE community periodically meets in national conferences focused on Document Delivery and interlibrary cooperation and organizes training courses dedicated to the librarians' community and focusing on the needs of end-users. After more than 20 years of activity, the NILDE project and its community have produced a rich variety of publications [4] and a large amount of digital data of various types: multimedia materials from conferences, video interviews, multimedia teaching resources, photographs, graphs, etc.

In 2019, the BdA decided to collect this large amount of material, dispersed throughout Italy, and to use an open data repository, accessible via the Web, to allow the NILDE community learning about the historical evolution of the NILDE project and finding useful material for their work.

We created the DALIA (Dati Aperti della bibLioteca di Area) repository with the twofold aim of preserving the digital data produced by the several initiatives and projects in which the BdA has been involved, and distributing them in an open way to the public. To this scope the whole corpus of digital data produced by the NILDE project represents a perfect starting point.

This short paper describes the main steps followed in constructing our repository and our future plans.

2. Repository building phases

DALIA is the result of the following building phases, briefly described hereafter:

1. Identifying the repository software/hardware platform
2. Repository setup
3. Mapping of available/existing data
4. Data retrieval
5. Data harmonization
6. Data input and metadata definition
7. Graphic interface

2.1. Identifying the repository software/hardware platform

The digital data held by the BdA are of various types and formats (multimedia materials from conferences, video interviews, multimedia learning resources, photographs, graphs, etc.) and document the intense activity carried out since 1995, and the NILDE project is no exception. In particular, the needs of the NILDE community, especially of the various professional working groups dedicated to training, communication and internationalization, are those of having a common repository of reference where to find, in a quick, efficient, updated and secure way, the digital materials mentioned above for the creation of new digital objects or simply to reuse them in new initiatives. Hence the need to store, catalog and publish these data by grouping them into homogeneous bins using an open-source archiving system that is simple to manage internally, that could store data in many different formats and that would allow BdA to distribute such data on the web in an open way under CC-BY-SA license.

Talking about open data also implies a reflection on how much these data can be made FAIR (Findable, Accessible, Interoperable and Reusable) as strongly suggested by the European Community.

Following the EU indications, in order to transform data into FAIR Digital Objects "*it is necessary to assign them a Persistent Identifier (PID) and create metadata rich enough to allow them to be reliably found, used and cited. The data should also be represented in commonly accepted formats, and be richly documented using the metadata standards and vocabularies adopted by the relevant research community. Finally, the data should contain provenance information to enable interoperability and reuse. The latter includes reporting on how the data were created (e.g., survey protocols, experimental*

processes, calibration information, and sensor locations) and information on data reduction or transformation processes to make the data more usable, understandable, or "science-ready"" [5].

The EU also suggests that "*FAIR is a scale, and various degrees of FAIRness can be applied to different data sets. It may not make sense, or even be feasible, to apply all FAIR principles to all outcomes. But a minimum level of FAIRness should be applied to the data being retained (e.g. discovery metadata, persistent identifiers, and access to the data or metadata)*" [5].

This latter concept applies to the data in our hands and after a careful analysis of the available open data catalogs (Invenio, Dataverse, CKAN) we avoid solutions that were oversized for our needs.

Our choice was therefore focused on the CKAN (Comprehensive Knowledge Archive Network) platform [6] because most of the data in our possession have a PID, the link of a resource (maybe already present in a recognized repository) can be inserted in CKAN, and CKAN allows to manually customize the metadata fields and preserve them even when the original dataset has been removed or is absent.

In addition, the platform is simple for both users and operators who need to input data and has a plugin-system that allows adding functionality; altogether it seemed the right choice to build a system adaptable to the future needs of the BdA. CKAN allows us to i) manage, publish, search datasets and documents in open format; ii) visualize data in tables, graphs and maps; iii) have the history of operations performed on datasets by one or more operators; iv) use APIs to manage and query datasets; and v) integrate with WordPress portals (which the BdA uses). In addition, CKAN's resources are held in trust by the Open Knowledge Foundation, a non-profit organization with best practice policies on open government and brand usage. Currently, our CKAN repository is mounted on a docker platform on our server.

2.2. Repository setup

Data are published in units called "*datasets*" in a CKAN system. Each dataset contains two items: 1) information or "*metadata*" about the data (i.e. title and publisher, description, date, formats available, released license, etc.); 2) several "*resources*", which hold the data itself without restriction on the format. CKAN can store the resource internally, or store it simply as a link, the resource itself being elsewhere on the web. A dataset can contain any number of resources and must belong to an *organization*. In addition, it is possible to define some *groups* to collect datasets even belonging to different organizations, in themes or projects. On these bases, we organize our data.

For our purposes, we rename "organization" into "ARCHIVES (ARCHIVI)" and "groups" into "TOPICS (TEMATICHE)" with the aim to facilitate the user's navigation. NILDE represents one of our ARCHIVES.

2.3. Mapping of available and existing data

The first bulk of datasets we decide to store inside the repository belongs to the NILDE project. As shown in **Figure 1**, the NILDE project is very articulated and has produced materials of various kinds such as training courses, conference proceedings, manuals and technical documentation, scientific publications or newspaper articles, graphics, and administrative documents.

In this first phase, we focused on materials related to the conferences. To date, 12 conferences have been held all around Italy. To organize the work of data mapping and data retrieval, specific guidelines were drawn up to determine the conference accompany material: website, program, conference proceedings, book of abstracts, collection of presentation slides, posters, photos, and videos.

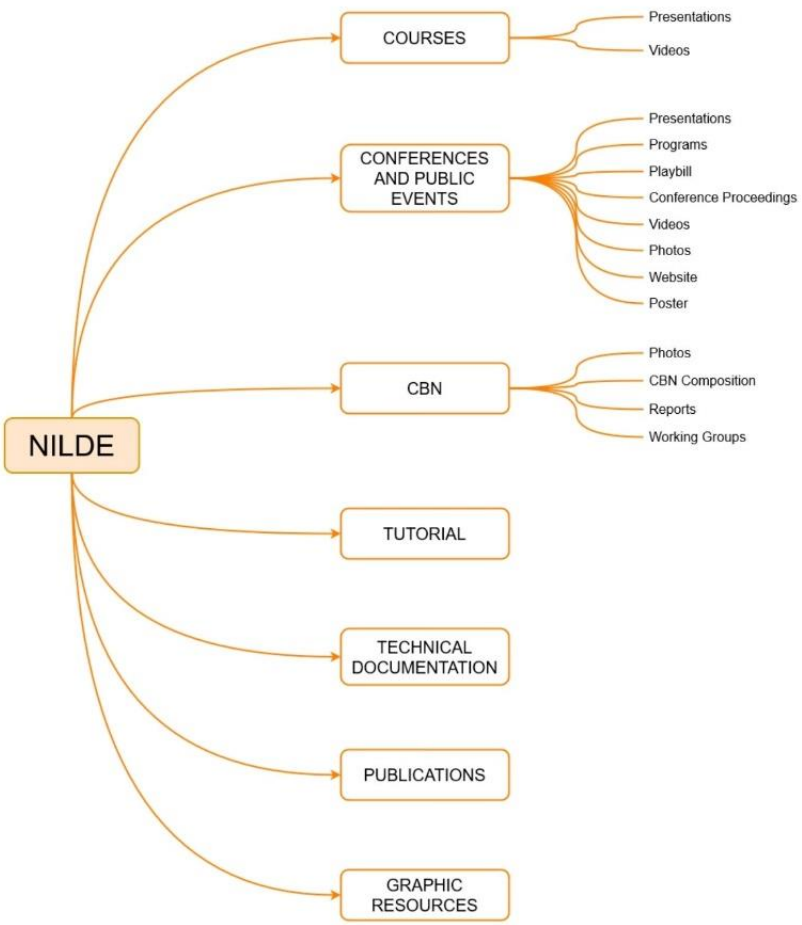

Figure 1: NILDE data mapping schema.

2.4. Data retrieval

A great deal of effort was required to recover the original data. As mentioned before, the 12 conferences were organized under the aegis of the BdA, but were hosted by various Italian institutions. So, the original data were not always present in their entirety within the BdA.

To retrieve every possible data, we explore each conference chronologically, listing all the materials mentioned before in a working sheet and checking all possible combinations and websites of hosting. Most of the difficulties we encountered were related to the early conferences, where there were very often broken or outdated links. In these particular cases, for example, we contacted directly the referent(s) of the individual hosting institutions, which are collaborators of the BdA. The aim of this inquiry was both to have the widest view of all the materials present and to go to the primary source of the data.

Then for unstable or "very-likely-to-break" links and files we decided to collect and archive a copy of these sets of data and store it locally. In one case, we even decide to make a copy of the conference's website using the tool Wayback Machine [7] hosted by Internet Archive [8].

Therefore, this work has allowed us to reconstruct and document the entire evolutionary path of the NILDE project in order to guarantee its preservation.

2.5. Data harmonization

As explained before, data has been retrieved from different sources so we've been forced to standardize them to provide users with a comparable view of data. Following our guideline, we defined some internal rules, and an example of the relationships between data type and formats we would like to use is shown in Table 1.

Table 1
Document data type and used data format

Document Type	Data Format
Video	Mp4
Presentation Abstract	pdf
Slide	pdf
Photographs	jpg

2.6. Data input and metadata definition

The input work is still in progress but up to now, more than 200 resources relative to 12 datasets (see congresses) are present inside DALIA. Basic metadata have been already described for each dataset that is distributed under a Creative Commons License Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). We are still refining the vocabularies and the ontologies we want to apply to our data to guide our potential users in the discovery.

2.7. Graphic interface

Our goal is to facilitate data discovery by our users (Researchers or librarians involved in the NILDE project), so we dedicate particular effort to building a graphic interface that could guide our users in consulting and finding data (see **Figure 2**: DALIA web-interface (prototype).).

The based color used for the graphic interface is a websafe red with a nuance pink-red (#993333), a brighter shade that reclaims the red color on the BdA Library logo. It conveys positivity, energy, and involvement. As compared to the tranquility emanating from the blue palette (the other color used for the BdA logo), red exalts the contents of the website and it is easy to remember/memorize by the users, unlike a blue-colored website. The icons are monochrome white, as colored icons could have been confused, not standing out through the red background.

DALIA has a portable interface, that is adaptable to all devices, even small screens such as mobile phones. Today these kinds of devices are becoming an increasingly agile tool and allow the users to retrieve data at any time and in any place.

Figure 2: DALIA web-interface (prototype).

3. Future work

DALIA is currently working but has not yet taken its final form and there is still a lot of data that needs to be included within it. The next steps will be to achieve the right combination in defining metadata, keywords and themes that can group the various datasets. The use of specific vocabularies and appropriate ontologies will also allow for increased interoperability between search systems, thus promoting the visibility of the NILDE project's open data on the web.

In the meantime, some Area Library researchers, about to retire, have expressed the need to organize their data and make them available to the community.

The tests carried out with DALIA suggest the potentiality of a new service to be proposed to these researchers in order to "educate" them on the correct procedures for archiving, describing and preserving data. We are evaluating the scalability of the system and the definition of some basic rules to allow and facilitate the management of data, for example, within a single research project or broader themes, or allowing individual researchers to reorganize their data scattered around the world with a view to future preservation. In this case, the repository that is created can be a useful tool for the organization of research work: both to allow an almost immediate retrieval of data, and to maintain in the same place a local copy of data already deposited in other international and certified repositories such as Zenodo[7] or other listed into ROAR[8], etc.

The advantages for a single researcher are the extreme simplicity of the storage system and the possibility to make the data openly accessible at any time. The ability to group different datasets belonging to different organizations (read researchers) under the umbrella of a project, for example, will allow data to be shared between researchers of different disciplines relatively easily.

Or, within the same institution, it might be possible to bring together all the work done by individual researchers working on the same topic. Last but not least, this system can be used as the basis for writing a Data Management Plan (DMP) required by the new European research program Horizon Europe. One of the future activities to support the proposal of this new service will be the organization of training meetings about Open Data, FAIR principles and the creation of Data Management Plans. An opportunity not only to provide a service to those who need to deposit or organize their material but

also to introduce these key issues and raise awareness among researchers of the Bologna CNR Research Area.

4. Acknowledgements

We thank all the librarians that are working in NILDE projects who in some ways helped us in retrieving the data. We thank R. Magno for retrieving the original HTML files of the conference website in which NILDE originated. A special thanks to M. Greco and M. Rossi who gave us a huge photographic archive documenting NILDE since its beginning. We thank A. Tugnoli and G. Resci that helped us physically building the infrastructure with their IT competence. We thank A. Frezzini for the graphic solutions.

5. References

[1] NILDE (Network for Inter-Library Document Exchange). URL: https://nildeworld.bo.cnr.it/
[2] Mangiaracina S., Giannuzzi M., Pistoia B., Guazzerotti M. (2005), Il sistema NILDE per il Document Delivery: dalla sperimentazione alla cooperazione, dal progetto al servizio, in "Biblioteche oggi", 23 (1), pp. 29-39.
[3] Mangiaracina S. et al (2008), NILDE: Developing a New Generation Tool for Document Delivery in Italy, in Interlending & document supply, 36 (3), pp. 167-177.
[4] NILDE project publications link. URL: https://nildeworld.bo.cnr.it/it/pub
[5] Open Science. URL: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science
[6] Turning FAIR into Reality: Report and Action Plan. https://doi.org/10.2777/1524
[7] CKAN, The world's leading open-source data management system. URL: https://ckan.org/
[8] Wayback Machine. URL: https://web.archive.org/web/
[9] Internet Archive. URL: https://archive.org/
[10] Zenodo. URL: https://zenodo.org/
[11] OpenAIRE. URL: https://www.openaire.eu/
[12] Registry of Open Access Repositories. URL: http://roar.eprints.org/