DOI: 10.26314/GARR-Conf19-proceedings-12

Managing research data at the University level: the experience of the University of Milan

Paola Galimberti[1], Susanna Mornati[2]

[1]University of Milan, [2]4Science s.r.l

Abstract. The University of Milan is the largest university in Lombardy and one of the largest in Italy, and the only Italian university to be part of the League of European Research Universities (LERU). As an observer on the international scene, it has witnessed a series of activities on research data at European level and the lukewarm transposition, at national level, of policies decidedly oriented towards the opening up of research processes. Among the recommendations concerning Open Science issued by the European Commission [1], the management of FAIR data [2] is one of the main pillars identified by LERU in the roadmap for universities [3]. Moving from observer to protagonist was a decisive step for the University of Milan therefore, confirming the responsibility of leading role among Italian universities assumed with the participation in the LERU. Little has been said about research data so far in Italy, always and only among technicians, little has been asked and little has been built. There is no central data infrastructure in Italy, although initiatives are being taken at group level. We are also of the opinion that the individual institutions should have a responsibility in the collection, management, care, conservation and transmission of the heritage of knowledge, a responsibility that has been instanced with the adoption of institutional repositories for scientific literature, and now requires a step forward, facilitated by the availability of ad hoc technologies, open and simple and inexpensive to adopt, to manage the research data produced at the university level. At the University of Milan, we tried to intercept the needs of users even before they were expressed. Researchers in the departments were interviewed to understand what the real needs were, but above all to try to understand the degree of awareness with regard to the issue of data storage and retention, data responsibility and the main critical issues in their management. The interviews revealed a varied picture with some points in common: need for technological support, need for training, need for legal support. With regard to technological support, we have tried to identify and make available to researchers a tool that could meet the requirements internationally, that was open source, that could also interface with other tools of the university. The choice fell on Dataverse, a free open source platform created within the University of Harvard and adopted by prestigious institutions around the world [4]. Following a one-year pilot project at the university, the production phase was completed at the beginning of 2019. The installation of Dataverse has been outsourced to 4Science, the only Italian company that collaborates with the international community of Dataverse and offers services in the cloud. The article explains how the choice of Dataverse came about and why it was decided to use an instrument other than AIR (the University's Institutional Repository), what the critical points were, how it is planned to proceed with the promotion of this instrument currently used by two departments, what integrations are desirable in the immediate future and what are the benefits expected from this new approach

Keywords. RDM, data preservation, data curation, data stewardship, Dataverse

1. Anticipate user needs

If at a global level there is a great deal of attention on research data and on the possibility of managing, archiving and preserving them, at a national and institutional level in Italy it is not yet clear what the needs of researchers might be, and therefore little has been achieved.

It is in an attitude of listening to the demands at international level and, above all, to the priorities of the LERU of which the University is a member, that an attempt has been made to understand and anticipate what the requests of researchers would have been once the requests of the research funding bodies had been accepted.

In order to understand what data was produced in the different disciplinary areas, how many, where it was stored and with what rules, how it was stored, who could access it and why, who was responsible for it, a series of interviews were carried out with professors and researchers from the different departments, the most evident outcome of which was the awareness that there was no management of the data, that there was no concern about its conservation, that the legal and ethical issues related to the data were very unclear and that its publication was in many cases completely excluded.

2. A Policy for research data management

The first step was therefore to develop a policy on the management of research data that provided for precise responsibilities on the data (by the university and by the Principal Investigators), which required that the management according to FAIR principles was the standard management, and defined a period of data retention. The definition of what the university means by research data and the subjects who are required to comply with the policy has been fundamental. In addition to requiring that data be managed in accordance with FAIR principles, the University has also undertaken to find and offer its researchers a suitable tool. The policy was submitted to and approved by the Academic Senate.

3. The search for a suitable instrument

Interviews showed that the researchers' needs for data management were clear, among them:

- have an easy-to-use tool
- that guarantees always and in any case the accessibility of the data
- which could also be accessed by colleagues from outside the university
- that complies with the FAIR principles
- that could be harvested by Open AIRE
- that could contain different versions of the datasets
- that attributed a DOI to the datasets and make them citable
- that would allow the publication of data with ad hoc licenses

Since the institutional repository for publications, AIR (based on DSpace technology) does not meet many of these requirements, a tool was sought that could meet all these needs.

Dataverse, a software platform for RDM developed by the Harvard University, was selected for a pilot project. It was chosen for several reasons that were assessed during the pilot and confirmed the choice:

- it is free and open source, providing a reliable and sustainable tool to support the institutional policy
- it is FAIR compliant, guaranteeing appropriate participation and high visibility in the research ecosystem

- it supports researchers' needs, namely the availability of a reliable, accessible, and durable tool to share data in a restricted group before making them public, to store data safely and make data discoverable for future reuse, to publish them in the appropriate version when funders require it, to cite and reference them in journal papers, and so on.
- Dataverse was born to meet a set of requirements to improve the data-sharing infrastructure of the scholarly community [5]:
- Recognition: as much as it is important for researchers to be recognised as the authors of a journal publication, likewise this need is felt for data sets;
- Public distribution: the dissemination of knowledge must be carried out through channels that are as public as possible and not through private agreements;
- Authorization: authors should be able to control access to their data sets, whether through a license, a guestbook, or a combination of methods;
- Validation: data's existence and validity should be verifiable even without access to the data itself;
- Verification: data's authenticity should be preserved over time, through the possibility of versioning and freezing the data sets:
- Persistence: data sets should be available for future access for an indefinite time;
- Ease of use: the tool to manage data sets should be used by researchers themselves, not necessarily mediated by professional archivists;
- Legal protection: licenses associated to data set should define responsibilities related to authorship, usage, etc.

4. Training and experimentation

Once the choice of the instrument had been made, it was proposed to the various departments, selecting in particular a department that would act as a tester for the different functions.

The choice fell on a department that had included the management of research data among the objectives of the three-year plan and that willingly carried out the experimentation.

At the end of the trial, which confirmed the choice of the tool[6], advertising began in the departments, both through a guide to the use of Dataverse that was distributed, and through ad hoc presentations aimed at framing the theme of data management, its purposes and the benefits that derive from it.

The theme has also been linked to that of research integrity, particularly with regard to the possibility of reproducing research.

5. The use of Dataverse

Currently there are two main ways of using Dataverse: on the one hand it is used to deposit the datasets that are then indicated in the journals as material accompanying the articles submitted, on the other hand it is used as a reference tool for data management in data management plans that accompany European projects and beyond.

6. An initial assessment

One year after the start of the trial and as the awareness of the importance of data management grows, the number of datasets stored has increased dramatically.

Most of the archive is not open at the moment, so the substantial part of the archive is not visible, even if it is accessible. There is still a lot of support work to be done, from producing guidelines and FAQs to help researchers find their way around, to setting up courses and tutorials.

At this point, however, we can say that a path has been undertaken, certainly challenging, but that will lead our researchers to a greater awareness about research processess, research reproducibility and its reuse even for purposes other than those for which it was designed.

References

[1] European Commission - Open Science: https://ec.europa.eu/research/openscience/index.cfm; last accessed 5 April 2019

[2] Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018, doi: 10.1038/sdata.2016.18

[3] League of European Research Universities (2018) Open Science and its role in universities: A roadmap for cultural change, Advice Paper no. 24 May, https://www.leru.org/files/LERU-AP24-Open-Science-full-paper.pdf

[4] The Dataverse Project, https://dataverse.org/, last accessed 5 April 2019

[5] Gary King (2007) An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research 36, pp. 173–199, https://gking.harvard.edu/files/abs/dvn-abs.shtml

[6] https://dataverse.unimi.it/

Authors

Paola Galimberti - paola.galimberti@unimi.it

Paola Galimberti is responsible for the supporting activities for the evaluation board, for the governing bodies and for the quality assurance. She manages the institutional archive IRIS (she is coordinator of the national working group Cineca), the platform of e-publishing OA (riviste.unimi.it) and the repository for research data (dataverse.unimi.it). She supports the activities of the OS committee of the University and is delegate in the LERU group on Open Science issues.

She is a member of the Board of AISA (Italian Association for the Promotion of Open Science); she is a member of the editorial staff of ROARS (Return on Academic Research), member of IOSSG, (Italian Open Science Support Group) and editor for Italy and Germany of the DOAJ.

Paola Galimberti is responsible for the supporting activities for the evaluation board, for the governing bodies and for the quality assurance. She manages the institutional archive IRIS (she is coordinator of the national working group Cineca), the platform of e-publishing OA (riviste.unimi.it) and the repository for research data (dataverse.unimi.it). She supports the activities of the OS committee of the University and is delegate in the LERU group on Open Science issues.

She is a member of the Board of AISA (Italian Association for the Promotion of Open Science); she is a member of the editorial staff of ROARS (Return on Academic Research), member of IOSSG,

(Italian Open Science Support Group) and editor for Italy and Germany of the DOAJ.

Susanna Mornati - susanna.mornati@4science.it

Susanna Mornati is COO at 4Science, Italy. She has extensive experience in the design and implementation of information systems for research, gained in thirty years spent at the University of Milan, CERN and university consortia for ICT. With her vast expertise in the research domain, in 2015 she directed the program of implementing DSpace-CRIS (IRIS) at 67 Italian HE and research institutions and the IRIDE project for ORCiD adoption at the national level in Italy. Both projects involved over 60,000 researchers and were successfully achieved in just a few months.

Susanna has gained an international reputation in the Open Science communities, participating in scientific boards and committees, and a speaker at numerous events. She is a member of the Research Data Alliance (RDA), the COAR Controlled Vocabularies Board, the DSpace Leadership and Steering Groups, the euroCRIS CRIS-IRs Task Group, the Italian Association for Open Science (AISA), the Italian Open Science Support Group (IOSSG).