Mathematical Problems of Computer Science 46, 81--86, 2016.

# Increasing the Visibility of Scientific Data in Armenia Using Persistent Identifiers

Hayk A. Grigoryan

Institute for Informatics and Automation Problems of NAS RA
e-mail: hayk-grigoryan@ipia.sci.am

## Abstract

During the development of computer technologies, the scientific research becomes more data intensive and collective than in the past. Data practices of researchers such as data sharing, discovery, reuse and preservation can be useful for other researchers in the same domain. Data sharing allows the verification of results and extends scientific research from previous results. Many scientific fields such as Biology, Astronomy, Weather forecast, etc., produce a vast amount of data, these data need to be shared and increase its accessibility, because sharing data has an important role for today’s science communities. In this paper, we introduce a deployed infrastructure to enable data-sharing using metadata which increases the accessibility of this data.

**Keywords:** Persistent identifier, Scientific data, Data sharing, Metadata.

## 1. Introduction

Data form the basis for good scientific decisions, management, and use of resources and informed decision-making. In addition, “science is becoming data-intensive and collaborative” [1]. Due to developments in computational simulation and modeling and communication technologies, the amount of data collected, analyzed and stored has increased enormously [2] and the science confronts with big data and complex data structures that traditional data processing applications are insufficient to operate with them. Digital data are not only the outputs of research but provide inputs to new hypotheses, enabling new scientific conceptions and driving innovation [3].

Data sharing becomes more important as science becomes more data intensive and amount of data daily increasing. Because of the huge size of scientific data, it’s good practice to share not only the data but also the metadata (data about data).

Metadata recapitulates basic information about data, which can make easier to find and work with specific instances of data. Metadata is structured information that declares, locates, explains, or otherwise makes it easier to retrieve, use, or manage an information resource. It can describe a different kind of resources such as single, collection or even a part of larger resources. As stated in [4] “metadata is a key to ensuring that resources will survive and continue to be accessible into the future”. Author, date created and file size are one example of basic document metadata. The ability to filter through that metadata helps someone to find the specific document.

In a phase where data-intensive science is moving towards automated processes, the usage of persistent identifiers (PID) for any type of Digital Object is crucial [5].

Currently, existing solutions provide good services for dealing with data sharing but usually they had some complex structures and the scientist who does not have knowledge of working with that kind of services may confront with difficulties of using that. Also, that services provide a metadata which may not match with our scientific rules. Our interface is simple to use and as it targets to the Armenian scientific communities, it is more flexible and can be changed depending on the community needs.

## 2. Persistent Identifier

A persistent identifier is a long-lived reference to a digital resource. It has two components: a unique identifier which ensures the provenance of a digital resource; and a service. When the resource location changes, the server locates the resource in the course of time and guarantee that the identifier resolves to the current location. The aim of Persistent identifiers is to solve the problem of the persistence of accessing cited resource, particularly in the scientific data. It can be used also for scientific data which is stored in the web network. Frequently, web addresses fail to take users to the expected referenced resource because of the technical problems with server or event more often by human-created failures. Organizations shift journals to new publishers, reconstruct their websites, or moving forward without using the older content, leading to broken links. If the referenced resource is essential for medical, legal or scientific reasons this can be frustrating for users [6].

## 2.1 Data Sharing

The data lifecycle cannot be considered independently from research lifecycle. Starting from Ideas and finalizing with publications the researchers confront with the search process which is a crucial part of this lifecycle (Figure 1) [7].

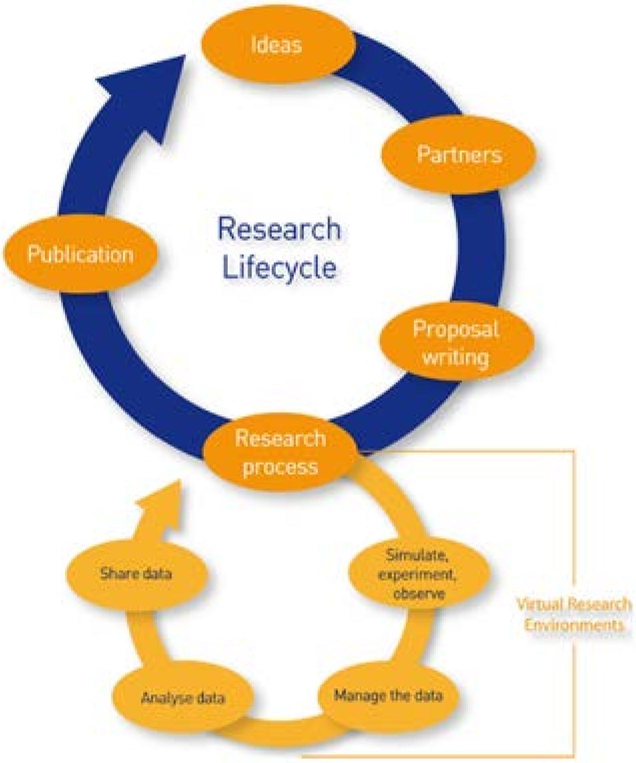

Fig. 1. Joint Information Systems Committee (JISC), Stages of the research and data lifecycle.

## 3. Related Works

In the area of data sharing already exist working services providing the scientists to share and save their data on the server which can be accessible for other communities.

ePIC can be considered as one of the famous providers of such services which was founded in 2009 by a consortium of European partners. It's providing the PID services based on the handling system [8] for the European Research. Handling Systems are used for the allocation and resolution of persistent identifiers. For the scientific research community ePIC provides 4 services: PID Service, PID Resolution, PID Replication and Global Handle Mirror Server. The ePIC API provides a software stack for a PID service [9].

There is another service provided by EUDAT which offers common data services, supporting multiple research communities as well as individuals, through a geographically distributed, resilient network of 35 European organizations. These shared services and storage resources are distributed throughout 15 European nations, and data is stored beside some of the most powerful supercomputers in Europe. Covering both access and deposit, from informal data sharing to long-term archiving, and addressing identification, discoverability and computability of both long-tail and big data, EUDAT's services address the full life cycle of research data [10].

The existing services resolve the problem with data sharing and allowing scientific communities to share their data within the network. However, they suggest some constraint rules for metadata and all data stored on their servers.

Our deployed infrastructure described in this paper will allow to be more flexible for storing and sharing data and will construct local data sharing rules. As our platform is based on the PID service provided by ePIC, it will allow our local scientific communities easily register their data and share them within the European Research.

## 4. Deployed Infrastructure

The schematic representation of the deployed infrastructure is presented below (Figure 2).

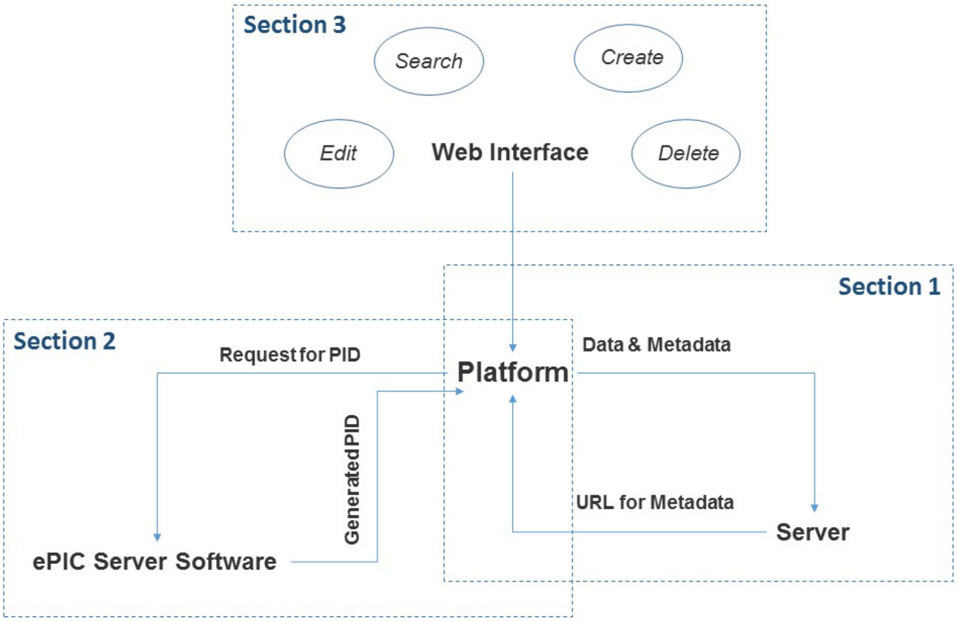

Fig. 2. The model of infrastructure.

The Model of Infrastructure can be divided into 3 sections.

**Section 1**: Registered communities can create metadata for their data and provide a reference for their resources. Metadata will be saved on our server and will generate the URL for that.

**Section 2**: Using the API provided by ePIC, the platform will request for PID generation passing the URL which was generated in Section 1. ePIC API allows to construct a generated PID by providing a custom suffix and prefix to the GUID . As it's working with one account that the ePIC provides to us the suffix of the generated PID will have a structure:

<Registered INST in ePIC system>-<GUID>-<Registered community INST in our system>

**Section 3**: Infrastructure will assign that generated PID with the metadata in our server and communities which are using our system can edit their metadata without touching the already existed persistent identifier. They can also delete their metadata which automatically will delete the PID from the ePIC system. We were also providing the search functionality within ePIC persistent identifiers and within our server data.

The infrastructure consists of two sides: back-end and front-end. For programming backend side the Node JS programming language with Express library was used. For storing the created metadata used the document based on MongoDB database engine. The front-end part is constructed with AngularJS technique which is based on the JavaScript script language.

## 5. Conclusion

In this paper, we have presented our infrastructure for data sharing which has a simple structure and provides an easy way to store and access data by scientists. The primary benefit of the work is to increase the visibility of the scientific community in Armenia by making their scientific output data more visible, trusted and accessible, and this in its turn will increase the productivity and collaboration between Armenian research communities and international communities. For future work the infrastructure can be improved by adding local replica storages which can be used for faster disaster recovery.

## References

[1] (2010) National Science Foundation. Press release 10-077 Scientists seeking NSF funding will soon be required to submit data management plans. Online. [Available]: http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928. Accessed 2010 Oct 2.

[2] (2009) National Academies of Science, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age Ensuring the integrity, accessibility, and stewardship of research data in the digital age. Online. [Available]: http://www.nap.edu/catalog.php?record_id=12615. Accessed 2010 Oct 5.

[3] (2010) National Science Foundation, Office of Cyber infrastructure Directorate for Computer & Information Science & Engineering (2008) Sustainable digital data preservation and access network partners (DataNet) program solicitation - NSF 07-601. Online.[Available]:http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141. Accessed 2010 Sep 22.

[4] (2013) NISO Press "Understanding Metadata", Online. [Available]: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

[5] (2015) Gary Berg-Cross, Raphael Ritz, Peter Wittenburg "RDA DFT Core Terms and Model", December 02, 2015, Online. [Available]: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

[6] (2010) JuhaHakala "Persistent identifiers – an overview", TWR Technology Watch Review, Online.[Available]:http://www.metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/

[7] (2011) Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, Mike Frame , "Data Sharing by Scientists: Practices and Perceptions", Online. [Available]: http://dx.doi.org/10.1371/journal.pone.0021101.g001

[8] (2016) The IEEE website. [Online]. Available: http://www.handle.net/

[9] (2016) The IEEE website. [Online]. Available: http://www.pidconsortium.eu/

[10] (2016) The IEEE website. [Online]. Available: https://eudat.eu/

**Submitted 04.08.2016, accepted 12.11.2016.**

# Հայաստանում գիտական տվյալների տեսանելիության բարձրացումը մշտական նույնացուցիչների միջոցով

Հ. Գրիգոյան

### Ամփոփում

Համակարգչային տեխնոլոգիաների զարգացման ընթացքում գիտական հետազոտությունների տվյալները դարձել են ավելի ինտենսիվ և հավաքական: Հետազոտողների տվյալների օգտագործման մեթոդները, ինչպիսիք են՝ տվյալների փոխանակում, բացահայտում, վերակազմակերպման օգտագործում և պահպանություն, կարող են օգտակար լինել  միննույն ոլորտի այլ հետազոտողների համար: Տվյալների տարածումը արդյունքների ստուգման հնարավորություն է տալիս և ընդլայնում է առկա գիտական հետազոտությունների արդյունքները: Շատ գիտական ոլորտներում, ինչպիսիք են՝ կենսաբանություն, աստղագիտություն, եղանակի կանխատեսում և այլն, արտադրվում են հսկայան քանակությամբ տվյալներ, որոնք պետք է լինեն ընդհանուր և բարձր հասանելիության, քանի որ տվյալների տարածումն ունի շատ մեծ դեր ժամանակակից գիտական հասարակությունում: Այս հոդվածում մենք ներկայացրել ենք տեղակայված ենթակառուցվածք, որը մետատվյալների օգտագործմամբ տվյալների փոխանակման հնարավորություն է տալիս, որը  մեծացնում է այդ տվյալների հասանելիությունը:

# Повышение видимости научных данных в Армении с использованием постоянных идентификаторов

А. Григорян

### Аннотация

При разработке компьютерных технологий, данные научных исследований становятся более коллективными и интенсивными, чем в прошлом. Методы использования данных исследователей, таких как обмен данными, открытие, повторное использование и сохранение, могут быть полезными для других исследователей в той же сфере. Обмен данными позволяет проверку результатов и расширяет научные исследования от предыдущих результатов. Многие научные направления, такие как биология, астрономия, прогноз погоды и т.д. производят огромное количество данных, эти данные должны быть общими и повысить его доступность, поскольку обмен данными играет важную роль для современных научных сообществ. В этой статье мы представляем инфраструктуру, позволяющую обмен данными с использованием метаданных, что увеличивает доступность этих данных.