EVALUATION OF FREE TECHNOLOGIES FOR THE MANAGEMENT OF METADATA IN DIGITAL REPOSITORIES FOR GOVERNMENTAL ENTITIES

Daniel Yucra[1], Julio Sotomayor[2], Jorge Mayhuasca[3], Antonio Arroyo[4], Carmen Chicasaca[5], Wilson Marin[6]

[1] Scientific University of the South, Peru.
[2,3]National University Federico Villarreal, Peru.
[4]National University of the Altiplano, Peru.
[5]Alas Peruanas University, Peru.
[6]César Vallejo University, Peru.

Email :[1]dyucra@cientifica.edu.pe,[2]jsotomayor@unfv.edu.pe,[3]jmayhuasca@unfv.edu.pe,
[4]arroyopaz@unap.edu.pe,[5]c_chicasaca@alu.uap.edu.pe,[6]wmarinv@ucv.edu.pe

ABSTRACT: This article presents the results obtained from the evaluation of free technologies for the management of metadata in digital repositories for governmental entities. The methods and procedures consisted of a literature search on digital repository software for data and metadata management, testing and evaluation of alternative solutions, as well as considering functional requirements as a basis: cataloging and data and metadata management systems. The result of this work highlights the characteristics, capabilities, and functionalities of each software that presents resources for building a data repository, sharing data, and representing it to users. In addition, technical evaluation criteria for testing, evaluation, and selection are presented. This proposal allows government sector entities to achieve technological sovereignty and independence, in addition to promoting the publication of open data for open government management.

KEYWORDS: Digital Repository, Metadata, Open Data.

1. INTRODUCTION

This work is motivated so that government sector entities can publish more open data, this involves implementing a repository solution to manage metadata, and make this knowledge known to citizens, researchers and the general public, so as to encourage and promote open and free technologies, in its implementation for the exchange of data in a decentralized manner.

An example is the National Open Data Platform (datosabiertos.gob.pe), which allows gathering data and information from different sectors and generating knowledge, this platform is free to access, organized by different themes, whose objective is the preservation, standardization, and versioning of the deposited data, always thinking of strategies to facilitate the reuse of data by the community.

In order to have more data and information available, it is necessary that the government entities that provide the data publish their metadata, using different tools for open data management, using standards of the Open Archives Initiative (OAI), which allows interoperability between digital repositories, both at the level of provider entities, as consumers and the national open data platform, to facilitate the reading of metadata performed by search engine algorithms, facilitating searches [1].

In Peru, there is the framework of Data Governance of the Peruvian State is constituted by technical and regulatory instruments that establish the minimum requirements that the entities of the Public Administration must implement according to their legal, technological and strategic context to ensure a basic and acceptable level for the collection, processing, publication, storage and opening of the data it manages [2].

An important factor to implement digital repositories by public entities is the software that will allow the management of metadata, the same must present characteristics, functionalities, strengths and often have to make a selection with technical criteria, support open standards to ensure the publication of more data and information for its citizens, researchers and the general public.

Open Data supports public oversight of governments and helps reduce corruption because it allows for greater transparency. For example, Open Data makes it easier to monitor government activities, such as tracking public budget expenditures and impacts [3].

In this opportunity, we present an evaluation of existing tools for metadata management that work with free licenses and collaborative schemes, which allows us to support the management of open data from its installation, deployment, and production of digital repositories for government entities.

DOI: 10.9756/INT-JECSE/V14I3.670

2.REVIEW OF THE LITERATURE

2.1. Open Data

According to the United Nations [4], we are facing a "data revolution", characterized by an explosion in the volume, in the speed at which data is generated and disseminated, and in the various sources and devices in which it is available.

This represents a challenge for governments, which must become stewards of information and data, for which they must foster collaborative spaces that enable the generation of value through the use of open data.

The International Open Data Charter defines open data as digital data that are made available to the public with the necessary technical and legal characteristics so that they can be freely used, reused and redistributed by anyone, anytime, anywhere [5].

2.2. Digital Repository

The Digital Repository is a repository of digital documents, files, and data, whose objective is to organize, store, preserve and disseminate in Open Access model.

2.3. Metadata

Metadata is data that describes other data. In general, a metadata group refers to a group of data describing the information content of an object referred to as a resource.

They are the descriptions of the data and services available; it is the documentation that allows the user to know the characteristics of quality, timeliness, availability, ownership, among others of the data, and the technical capabilities of the services such as type of service, version, availability information, rates, payment instructions and even restrictions.

2.4 DataSets

Refers to the categorization of public data in data catalogs, organized in datasets to be easily indexed and located. Therefore, fields are used that define the dataset such as description, update frequency, format, license of use, among others [6].

2.5. Fair Principle

Guiding principles to make research data easy to find, accessible, interoperable, and reusable [7].

2.6. Metadata used in government open data catalogs

Data Catalog Vocabulary (DCAT) is a definition for open data sets defined by the World Wide Web Consortium (W3C), adopted by Europe and used internationally in different government open data catalogs, such as data.gov and data.gov.uk. DCAT is a Resource Description Framework (RDF) vocabulary that facilitates interoperability between government open data catalogs, in which metadata for dataset identification is defined by classes [8].

2.7 Formats for data used in government open data.

The W3C published a working document called Publishing Open Government Data, in which it urges, in the implementation of open government data projects, to use data formats that allow their widespread use, guaranteeing access, use and reuse of the data. Therefore, it recommends XML, and RDF structured formats because they allow manipulation and reading through tools, databases, and query languages, such as SPARQL, XQuery and JavaScrip. This makes them open standards for the production and publication of this type of data and for leveraging linked data [9].

2.8 Government open data value chain.

Implicitly, open government data includes two values: 1) that of investment, used to create, store and maintain the data; 2) business and profits, which can be generated by using the data to create consumer applications. This growth suggests the need to create an infrastructure that manages the data value chain (Figure 01), from data creation, data storage, data processing and packaging, to finally delivering the data to stakeholders [10].

2.9 Implementation process for data openness

Under the process approach, five stages with their respective instruments of the implementation process for data openness are considered, according to Figure 01:

DOI: 10.9756/INT-JECSE/V14I3.670

Figure 01. Implementation Process for Data opening
Source: National Office of Electronic Government - Peru (2016) [11]

3. PROBLEM DEFINITION/FORMULATION

The high cost of software licenses for this solution, and the growing demand for free solutions with access to source code to manage digital repositories, make a government entity select a tool that allows, implement, manage, operate, publish and support a wide variety of data from the entity to the citizenship, This way, they can have repositories that interoperate with other platforms and return to the citizen the interest to be informed, to participate in what their authorities are doing, to share the information to which they have access and to propose innovative solutions to the great challenges facing the country, and allows to change the climate of distrust of the authorities and rulers.

4. METHOD OF ANALYSIS

The management of metadata management in digital repositories for government entities, induces Open Data analysts of public sector entities, to review the literature, perform comparative analysis, perform proof of concepts, and review free or licensed technologies, with the purpose of publishing repositories with open data.

Therefore, the evaluated software was defined by consulting the Peruvian National Open Data Platform (datosabiertos.gob.pe). Figure 2 presents the three most used free and open-source software for metadata management in repositories until August 2021 [12].

For this project, 3 software based on free licensing were selected: Dspace, Dataverse and CKAN, which are supported by free software licenses, therefore, government entities have access to download, implement and put into production respectively.

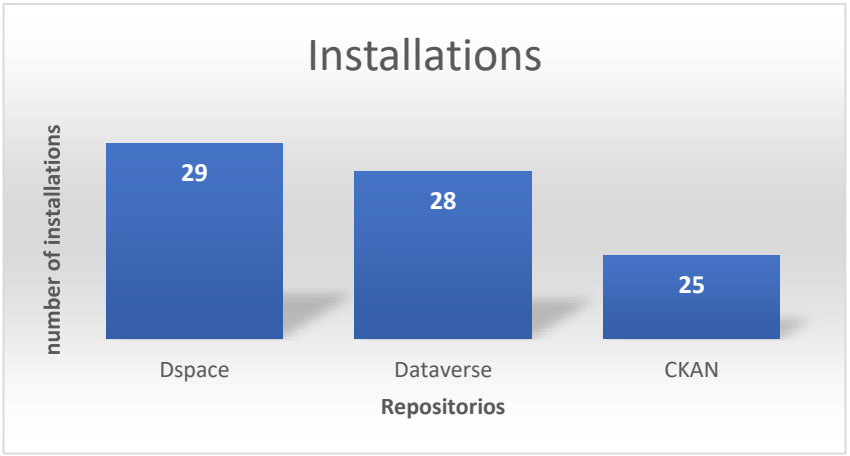

Figure 02. Number of Open Data Repository Software Installations in Public Entities
Source: National Open Data Platform [12]

For the evaluation of free technologies for metadata management, 12 criteria formulated by Edward C. Luczak and Jennifer L. Marill of the National Library of Medicine® (NLM) of the National Institutes of Health (NIH) of the United States [6] were established to evaluate, test and recommend software and digital repository systems to support NLM collection and preservation of a wide variety of digital objects.

5. RESULTS AND DISCUSSION

5.1 Results obtained

For the evaluation and selection of the free tool for the management of metadata in digital repositories for governmental entities, it is necessary to first establish a Technological Architecture on which the repository software will be implemented.

DOI: 10.9756/INT-JECSE/V14I3.670

Table 1. Digital Repository Software Products based on open-source software evaluated

Product evaluated	**Evaluated version**	**Operating system used**	**Download URL**
Dspace https://duraspace.org/	7.0 Backend y 7.0 User Interface	CentOS Linux 8	https://duraspace.org/dspace/download/
Dataverse https://dataverse.org	5.6	CentOS Linux 8	https://github.com/IQSS/dataverse
CKAN https://ckan.org/	2.9	CentOS Linux 8	https://github.com/ckan/ckan

The following are the technological and licensing characteristics of the solutions evaluated:

Table 2 - Technological characteristics of the free solutions

Product evaluated	**License**	**Programming Language**	**Orientation**
Dspace	BSD license	Java / Angular	Institutional bibliographic repository.
Dataverse	Apache License Version 2.0	Java	Research data repository.
CKAN	Affero General Public License	Python	Repository, Data storage and distribution.

The criteria were assigned a weighting, according to the range of evidence evaluations according to Edward C. Luczak and Jennifer L. Marill of the National Library of Medicine® (NLM) of the National Institutes of Health (NIH) of the United States [15]:

Table 3. Scoring criteria

Score	**Level**
0	None
1	Low
2	Moderate
3	High

Finally, the Agile UP methodology was applied for the use and development of an application for metadata management, where the selected solutions were used for evaluation and testing, and the necessary software and hardware configurations were used. The DataVerse stands out for the higher demand for RAM memory. This occurs because the software architecture model allows the creation and management of several Data Verses from the installed software. In the case of the DSpace software, it requires more storage space on the hard disk because we need to configure, in addition to DSpace, another dependency to perform the data management, which justifies the size of the hard disk.

Table 4. Minimum software and hardware requirements

CKAN	**Hard Disk:**25GB **Memory ram:**4 GB **Processor:**4 cores of 2 GHz

DOI: 10.9756/INT-JECSE/V14I3.670

DSpace	**Hard Disk:**48GB **Memory ram:**4 GB **Processor:**2 cores of 2 GHz
DataVerse	**Hard Disk:**25GB **Memory ram:**8 GB **Processor:**2 cores of 2 GHz

Taking into account the minimum installation requirements of the analyzed software, Table 5 presents the evaluation of free software technologies for the management of metadata in digital repositories for governmental entities, obtaining the following results:

Table 5. Evaluation of Free Software Technologies for the management of metadata in digital repositories for government agencies

Criteria to be considered	**Dspace**	**Dataverse**	**CKAN**
Functionality	3	3	3
Scalability	3	3	3
Interoperability	3	3	2
Ease of implementation	2	3	3
System security	2	3	2
System performance	3	3	3
Physical environment	3	3	3
Proven successful implementations	3	3	3
System support	3	3	3
Development community strength	2	2	2
Stability of the development organization	2	3	2
Strength of the technology roadmap for the future	3	2	2
Score Obtained	**32**	**34**	**31**

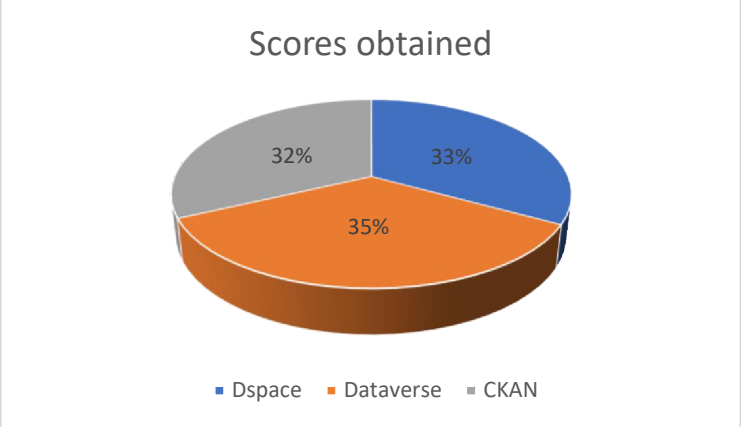

Figure 03. Scores obtained according to the evaluation criteria
Source: Own elaboration

DOI: 10.9756/INT-JECSE/V14I3.670

5.2. Discussion

Taking into account the results of Table 5 and Figure 3, it was decided to choose and select the free information technology offered by Dataverse in its version 5.6, under Apache License Version 2.0, implemented in the CentOS Linux 8 Operating System.

Support for FAIR data principles, version control, data citation for datasets and files, OAI-PMH (harvesting), API for interoperability and custom integration, search customization, file hierarchy, DataCite integration, usage statistics and metrics, use of Schema. org JSON-LD, notifications, object storage, security services checking (MD5, SHA-1, SHA-256, SHA-512, UNF), big data handling, downloading in R and TSV (data tabulation) language formats, among other functional features.

On the other hand, it has the ability to integrate with APIS and different platforms (TwoRavens, WorldMap, OJS, OpenSci, DataCite, Piwik, among others) and future integrations such as ORCID, Almetric, DMPTool among others.

6. CONCLUSIONS

This selection of the Dataverse solution is a complementary part of an open data publishing and digital repository implementation problem, most importantly the sustainability, support, maintenance and administration of metadata management and thus ensure the production in the publication of open data.

Metadata management helps users who consume or use the data to make efficient decisions and achieve support public oversight of governments and help reduce corruption because they allow greater transparency in a country.

DataVerse, presented as a robust software not only able to work with the data but also to perform its versioning. Both Dataverse and DSpace were developed to be a solution-focused on scientific research, metadata management, the assignment of persistent unique identifiers, and the possibility of interoperability. Dataverse is presented as a solution that presents easy access to reuse datasets, in addition to providing an application programming interface (API) for depositing and querying deposited DataSets.

This Dataverse solution constitutes a viable, sustainable and highly eligible technological alternative for organizations of government sector entities that want to publish their data and information and thus improve the delivery of data and information in a way that is relevant to the community and that has no legal restrictions for its dissemination and can be freely used, reused and redistributed by anyone who considers it useful.

As future work, an analysis will be carried out that includes operating system performance, process time measurement, database performance and evaluation of hardware versus software performance, the level of interoperability and the impact it has on the entities that implement it.

7. REFERENCES

[1]. "Open Archives Initiative", https://www.openarchives.org/, (2019). [Online]. Available: [Accessed: 16-oct-2021]

[2]. "DecretoLegislativo N° 1412, aprueba la Ley de Gobierno Digital, el Marco de Gobernanza de Datos del Estado Peruano", https://cdn.www.gob.pe/uploads/document/file/353216/decreto-legislativo-que-aprueba-la-ley-de-gobierno-digital-decreto-legislativo-n-1412-1691026-1.pdf, ((2018) [Online]. Available: [Accessed: 16-oct-2021]

[3]. "Políticas de DatosAbiertos- Banco Mundial", http://opendatatoolkit.worldbank.org/es/starting.html#policies, (2009), [Online]. Available: [Accessed: 17-oct-2021]

[4]. "A World that Counts Mobilising the Data Revolution for Sustainable Development", https://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf, (2014), [Online]. Available: [Accessed: 17-oct-2021]

[5]. "Open Definition", http://opendefinition.org, (2015), [Online]. Available: [Accessed: 15-oct-2021]

[6]. Jennifer L. Marill& Edward C. Luczak, "Evaluation of Digital Repository Software at the National Library of Medicine", Volume 15 Number 5/6, ISSN 1082-9873.(2009)

[7]. Mark. Wilkinson, M. Dumontier. "The FAIR Guiding Principles for scientific data management and stewardship", [Online]. Available: https://www.nature.com/articles/sdata201618, [Accessed: 14- Sep-2021].

[8]. "Data Catalog Vocabulary (DCAT)", [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/, (2014), [Accessed: 14- Sep-2021].

[9]. "Publishing Open Government Data", [Online]. Available: https://www.w3.org/TR/gov-data/, (2014), [Accessed: 14- Sep-2021].

DOI: 10.9756/INT-JECSE/V14I3.670

[10]. V. Grieken, "Open Data: A design for the provisioning of Dutch government public and geospatial transport data", Bachelor Thesis. University of Groningen. (2011)
[11]. "Guíarápidaparaapertura de DatosGubernamentales -OficinaNacional de GobiernoElectrónico", [Online]. Available: https://sgp.pcm.gob.pe/wp-content/uploads/2016/10/Guia-Rapida-ADG.pdf, (2016), [Accessed: 29- Sep-2021].
[12]. "PlataformaNacional de DatosAbiertos", https://www.datosabiertos.gob.pe/, (2019). [Online]. Available: [Accessed: 16-oct-2021]

AUTHORS

Daniel Alejandro Yucra Sotomayor
Professor at the School of Systems Engineering of the Universidad Nacional Federico Villarreal, Master in Advanced Project Management, Master in Computer Science, was Deputy Secretary of Information Technology of the Secretariat of Digital Government of the Peruvian Government (2019), specialist in interoperability, data governance, enterprise architecture and digital transformation.

Julio Elmer Sotomayor Abarca

Professor at the School of Systems Engineering of the Universidad Nacional Federico Villarreal, Systems Engineer, Master in Computer Science and Master in Administration and Accounting. Graduated from the Graduate School - Doctorate in Systems Engineering, Universidad Nacional Federico Villareal Lima. Currently studying for a Master in Administration and Advanced Project Management at the Catholic University of Murcia, Spain.More than 20 years of teaching experience in different universities in Lima and the country.

Jorge Victor Mayhuasca Guerra
PhD in Engineering, Master in Systems Engineering, Industrial Engineer, Research Professor, Senior Lecturer appointed to the Faculty of Industrial and Systems Engineering.Director of the Academic Department of Systems Engineering.Director of the Professional School of Industrial Engineering.Member of the Scientific Committee of the School of Industrial and Systems Engineering.President of the quality committee of the Professional School of Systems Engineering.

Antonio Arroyo Paz
Professor of the Master's Degree in Systems Engineering, Statistics and Computer Science at the Universidad Nacional del Altiplano Master's Degree in Systems Engineering, mention in Software Engineering. PhD (c) Science, Technology and Environment from the Universidad NacionaldelAltiplano.

Carmén Rosa Chicasaca Arpita
Professor at the professional school of Systems Engineering and Computer Science at Alas Peruanas University, Systems Engineer, Master in Administration. Graduated as Doctor in Systems Engineering at the Universidad Nacional Federico Villarreal. Master in Public Management at EUCIM Business School in Spain. I have 10 years of experience as a university teacher in Lima-Peru.

Wilson Ricardo Marin Verastegui
Professor of the professional school of Systems Engineering at the Universidad César Vallejo, Systems Engineer, Master in Information Technology Management. Master in University Teaching. Graduated as Doctor in Systems Engineering at the Universidad Nacional Federico Villarreal. I have 9 years of experience as a university teacher in Lima-Peru.

DOI: 10.9756/INT-JECSE/V14I3.670