Analysis of the Open Data Landscape in Mexico

Joanna Alvarado Uribe

Institute for the Future of Education

Tecnologico de Monterrey

joanna.alvarado@tec.mx

Paola Mejía Almada

Institute for the Future of Education

Tecnologico de Monterrey

gabriela.almada@tec.mx

Alma Beatriz Rivera Aguilera

Biblioteca Francisco Xavier Clavigero

Universidad Iberoamericana

alma.rivera@ibero.mx

Benjamín Alejandro Guerrero Olvera

Biblioteca Francisco Xavier Clavigero

Universidad Iberoamericana

benjamin.olvera@ibero.mx

María Teresa Villalón Guzmán

Instituto Tecnológico de Celaya

Tecnológico Nacional de México

teresa.villalon@itcelaya.edu.mx

María Guadalupe Vega Díaz

Biblioteca Daniel Cosío Villegas

El Colegio de México

guvega@colmex.mx

JOAQUÍN GIMÉNEZ HÉAU

DIRECCIÓN GENERAL DE REPOSITORIOS UNIVERSITARIOS

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

joaquin@dgru.unam.mx

EJE TEMÁTICO

Datos abiertos

ABSTRACT

The concept of open data originated from the idea of making government data available to anyone. In 2013, the Mexican government launched the National Digital Strategy to encourage the adoption and development of information by implementing five key enablers, including an open data portal to make government data more accessible to citizens. In 2014, the beta version of the datos.gob.mx portal was launched. Due to the significant potential of data to generate benefits, it has been referred to as the new gold. However, open data also presents some challenges, such as lack of interoperability and reproducibility and the resources needed to provide and maintain open data available to all. Therefore, based on the transversal agenda REMERI-ANUIES-EduTraDi focused on Open Science Ecosystems, the objectives of this research are to identify and analyze institutional repositories promoting Open Data in Mexico in order to document their characteristics and determine possible areas for improvement in such initiatives. This investigation allowed the identification of ten repositories promoting open data in Mexico and the integration of 16 characteristics that helped to describe and analyze the open data landscape, showing that the different open data management efforts carried out by Mexican institutions are mostly consolidated.

KEYWORDS

Open data; repository; Open Data Landscape; Mexico.

Introduction

The concept of open data originated from the idea of making government data available to anyone, as government institutions collect a vast amount of information from a variety of sources. Such a concept gained momentum in the late 2000s due to the rise of the Internet and the proliferation of digital information. During this period, there was a concerted effort by governments and various organizations to provide access to these resources to a large number of users (The World Bank Group, Open Data in 60 Seconds, 2021). The first government policies appeared in 2009, while in 2010, the World Bank organization launched the first “Open Data Initiative”. This initiative had the goal of finding solutions to complicated development problems by giving researchers access to a wide range of data related to the global economy (The World Bank Group, Learning About the Open Data Initiative, 2012).

In addition to the Open Data Initiatives launched every year, more than 250 countries along with almost 50 developed and developing countries, as well as other organizations (such as the United Nations), have introduced Open Data initiatives (The World Bank Group, Open Data in 60 Seconds, 2021). In 2013, the Mexican government launched the National Digital Strategy to encourage the adoption and development of information by implementing five key enablers, including an open data portal to make government data more accessible to citizens (Digital Government Unit, 2015). This became a reality in 2014 when the beta version of the datos.gob.mx portal was launched. Since then, this portal has been available and has gone through several updates (Coordinación de Estrategia Digital Nacional, 2015).

The definition of open concerning open data and content states that: Open means that anyone can freely access, use, modify, and share it for any purpose (Open Knowledge Foundation, 2015). Moreover, it is also important to emphasize that open data must be available in a common, machine-readable format and licensed so that anyone can manipulate data. For example, transforming, combining, and sharing it with others, even commercially

(European Data, n.d.). Due to the significant potential of data to generate benefits, it has been referred to as the new gold. Some of the benefits of providing and using open data are transparency, public service improvement, innovation and economic value, and efficiency (The World Bank Group, Open Data in 60 Seconds, 2021). However, open data also presents some challenges, such as loss of data privacy and security, lack of interoperability and reproducibility, improvement of data quality, intellectual property rights, and the resources needed to provide and maintain open data available to all (Goben & Sandusky, 2020).

One of the most recent studies related to open data repositories highlights the importance of the role of infrastructure, as it becomes crucial in the dissemination, visibility, and openness of data for research processes (Science, et al., 2022). Therefore, based on the transversal agenda REMERI-ANUIES-EduTraDi focused on Open Science Ecosystems, as well as on the aforementioned background and related work, the objectives of this research are to identify and analyze institutional repositories promoting Open Data belonging to the public and private sectors in Mexico in order to document their characteristics and determine possible areas for improvement in such initiatives.

The rest of the article is organized as follows. Section 2 presents the methodology followed in this research project. Subsequently, Section 3 provides the results, analysis, and discussions about them. Finally, Section 4 gives the conclusions and proposals for improvement.

Methodology

The identification and analysis of institutional repositories that incorporate Open Data belonging to the public and private sectors at the national level are intended to document their characteristics and determine areas for

improvement in such initiatives. To achieve this purpose, the methodology presented in Figure 1 and described below was followed:

1) Initial identification of repositories incorporating open data in Mexico. This first phase was based on two aspects: 1) on the knowledge and experience of the members of the research project, whose profiles are mainly related to the management of data repositories in Mexican institutions, and 2) on the search for repositories whose description or keywords included references to open data through different digital resources, such as websites, web platforms, research and dissemination publications, reports, among others.
2) Definition of the characteristics (variables) reviewed and registered for each repository identified in this investigation. To carry out this definition, the characteristics were established considering the purpose of the research, as well as different standards and guides related to types of data and areas of knowledge.
3) Collection of values for each characteristic of the repositories integrating open data in Mexico. To document the defined characteristics, the information provided on the platforms or websites of each repository was consulted and the values obtained were registered in a spreadsheet structured for this purpose. In addition, in some cases where the information was not available, the person responsible for the repository was contacted to request their support in providing the missing information.
4) Discussion of the information collected for each repository by the person responsible for its registration and all the members of the research project. This activity allowed to resolve doubts about the registered information and thus improve the quality and standardization of the information collected.
5) Elaboration of a quantitative analysis of the characteristics of the repositories. In order to discover the most used aspects and the areas for

improvement in the repositories promoting open data, statistics of the values and different visualizations by characteristic were carried out. Subsequently, the analysis of these representations and the information itself was performed.

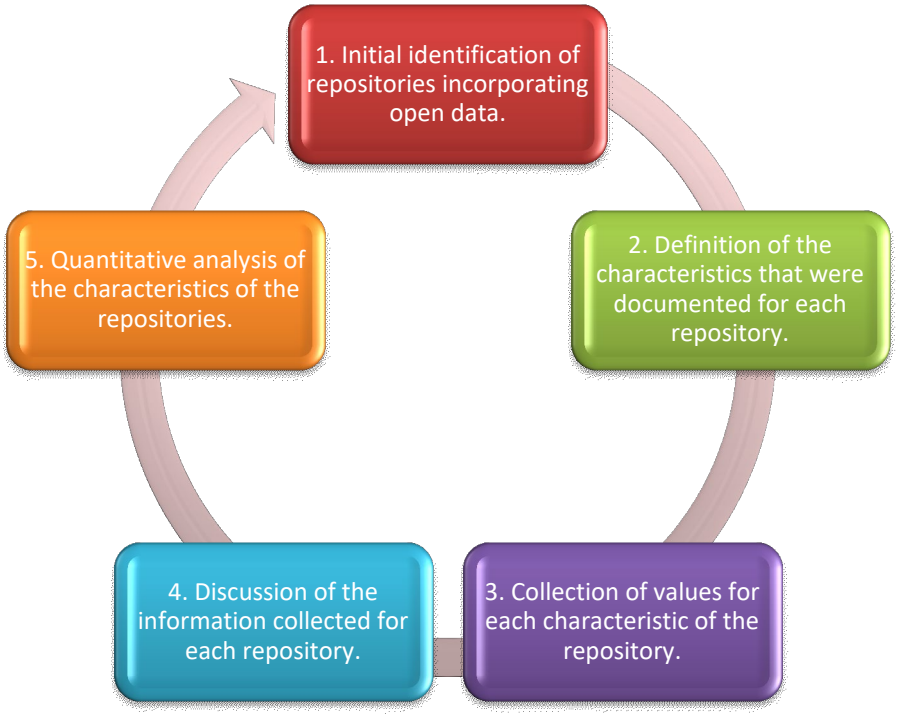

FIGURE 1. Methodology used to analyze the open data landscape in Mexico. Own elaboration.

Results

Firstly, the characteristics considered in the documentation of the repositories promoting open data, partially or completely, in Mexico are described. Later, the information about the identified repositories and the graphs visualizing the collected characteristics are shown. Finally, the analysis of the information presented is provided.

Definition of the Collected Characteristics

The 16 characteristics defined for the analysis of the repositories, promoting open data in Mexico, are related to their identification (such as name and Uniform Resource Locator (URL)), the area of knowledge of their content, their typology, numerals, metadata standards, use licenses, among other aspects oriented to data and software. The detailed list of features and their description is given below:

1) Repository name.
2) Repository URL.
3) Description of the repository. Overview of the repository content.
4) Institution. Institution name responsible for the repository.
5) Sector. The sector to which the institution responsible for the repository belongs is indicated.
6) Name of the person in charge. Full name of the person responsible for the repository or of the person related to the activities of the repository.
7) Email of the person in charge.
8) Data typology (LEARN project, 2020).
    a) By its type. It indicates whether the data is in the type of electronic text documents, data sheets, metadata, among others.
    b) By its file formats: It is specified if the data is textual, numerical, structured, among others, as well as the file extensions, for example, .csv, .txt, .pdf.
    c) By its data processing level: the data can be indicated as 1) raw or unprocessed, 2) processed, or 3) analyzed.
    d) By the data generation source: the data can be considered as 1) canonical or reference, 2) experimental, 3) models or simulations, 4) derived or compiled, or 5) Observational.
9) Classification based on the data structure. It indicates whether the repository is structured into Collections, Datasets, Geospatials, or Other.

10) Areas of Knowledge. The area of knowledge of the repository content is indicated according to the areas defined by the Consejo Nacional de Ciencia y Tecnología (CONACYT) (CONACYT, 2022).
11) Metadata standard.
12) Technology/Platform. The platform hosting the repository is specified.
13) Data use license.
14) Data use restriction.
15) Numerals. The figure (obtained or calculated) of collections and resources in the repository is provided.
16) Open Data. It indicates whether the repository contains only open data.

Repositories Promoting Open Data in Mexico

The identification of repositories promoting open data at the national level was carried out from September 9, 2022, to November 2, 2022. To consider the integration of the repository into this research, the following aspects were taken into account:

1) Repositories must present at least one open data record.
2) Open data must come from Mexican institutions.
3) In case of that an institutional repository indicates that its open data is in an exclusive open data repository, the repository that should be considered for the analysis is the one that only integrates open data.

Based on these considerations, a total of ten repositories presenting open data were selected to carry out their complete documentation. The overview of these repositories can be seen in Table 1, which is updated to the review period of the characteristics from May 8 to 12, 2023.

No.	Repository Name	Repository URL	Institution	Sector	Open Data Exclusive
1	Portal de Datos Abiertos UNAM	https://datosabiertos.unam.mx/	Universidad Nacional Autónoma de México (UNAM) / Dirección General de Repositorios Universitarios	Public	Yes
2	Data Hub del Tecnológico de Monterrey	https://datahub.tec.mx/	Tecnológico de Monterrey	Private	Yes
3	Datos Abiertos de México	https://datos.gob.mx/	Gobierno de México	Public	Yes
4	Laboratorio Nacional de Políticas Públicas	http://datos.cide.edu/community-list	Laboratorio Nacional de Políticas Públicas	Public	Yes
5	CIMMYT Research Data & Software Repository Network	https://data.cimmyt.org/dataverse/root?q=&types=dataverses&sort=dateSort&order=asc&page=1	International Maize and Wheat Improvement Center	Public and private	Yes
6	Repositorio Digital del Servicio Sismológico Nacional	http://www2.ssn.unam.mx:8080/catalogo/	UNAM / Servicio Sismológico Nacional / Instituto de Geología	Public	Yes
7	Datos abiertos - INEGI	https://www.inegi.org.mx/datosabiertos/	Instituto Nacional de Estadística y Geografía (INEGI)	Public	Yes
8	Repositorio de Documentación sobre Desapariciones en México (RDDM)	https://rddm.mx/	Center for Research Libraries; El Colegio de México; Universidad Iberoamericana; UNAM / Instituto de Investigaciones	Public	Partially

No.	Repository Name	Repository URL	Institution	Sector	Open Data Exclusive
			Jurídicas		
9	Datos Abiertos - SSA/DGIS	http://www.dgis.salud.gob.mx/contenidos/basesdedatos/Datos_Abiertos_gobmx.html	Secretaría de Salud (SSA) / Dirección General de Información en Salud (DGIS)	Public	Yes
10	Repositorio IBERO	https://ri.ibero.mx/	Universidad Iberoamericana	Private	Partially

Table 1. Description of the ten repositories promoting open data in Mexico.

Open Data Landscape in Mexico

As can be seen in Table 1, ten repositories promoting open data in Mexico were identified. Of these, more than half correspond to public institutions (such as universities and government institutions) and two to private educational institutions. In addition, 80 % of the identified repositories exclusively publish open data mainly in the areas of knowledge of "Social Sciences", "Physical-Mathematical and Earth Sciences", and "Engineering and Technological Development". Considering the number of collections and data records (resources), some repositories can be expected to have up to 40 collections and 2,106,655 data records, while some have at least one collection and six resources.

Regarding the data typology, Figure 2 shows that the data is mainly provided in Datasheets (22.73 %), followed by Electronic text documents (20.45 %), and also with their Metadata (13.64 %). This has a relationship to Figure 3, where the .csv, .pdf, .xlsx, and .txt file formats are more represented (larger) in the word cloud and therefore, more published in open data repositories.

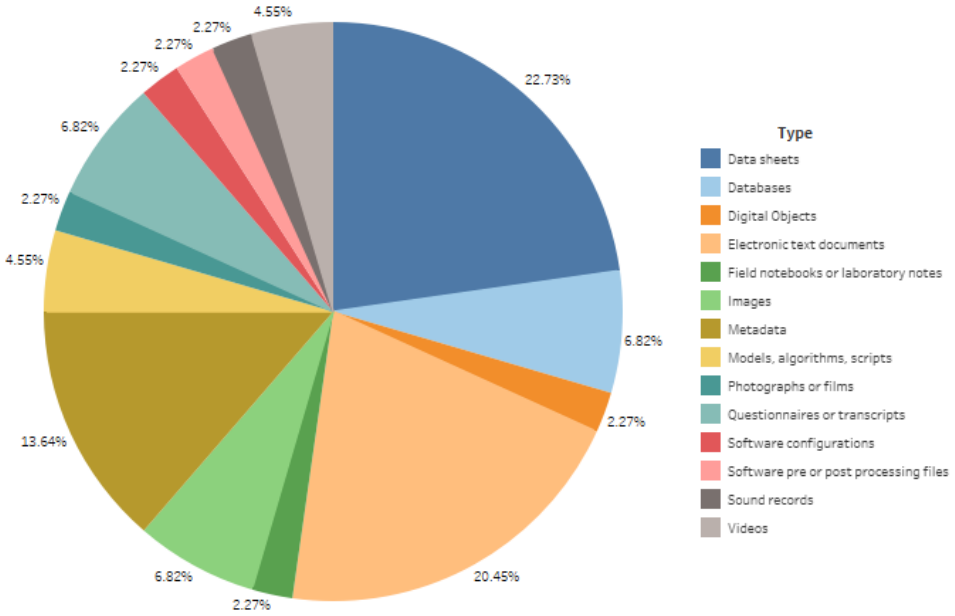

FIGURE 1. Data typology according to its type. Own elaboration

FIGURE 2. Data typology according to its file formats. Own elaboration

From Figures 4 and 5, it can be seen that all repositories contain processed data, followed by analyzed data (90 %), generated mainly from other data (derived or compiled) and considered as curated datasets (canonical or reference).

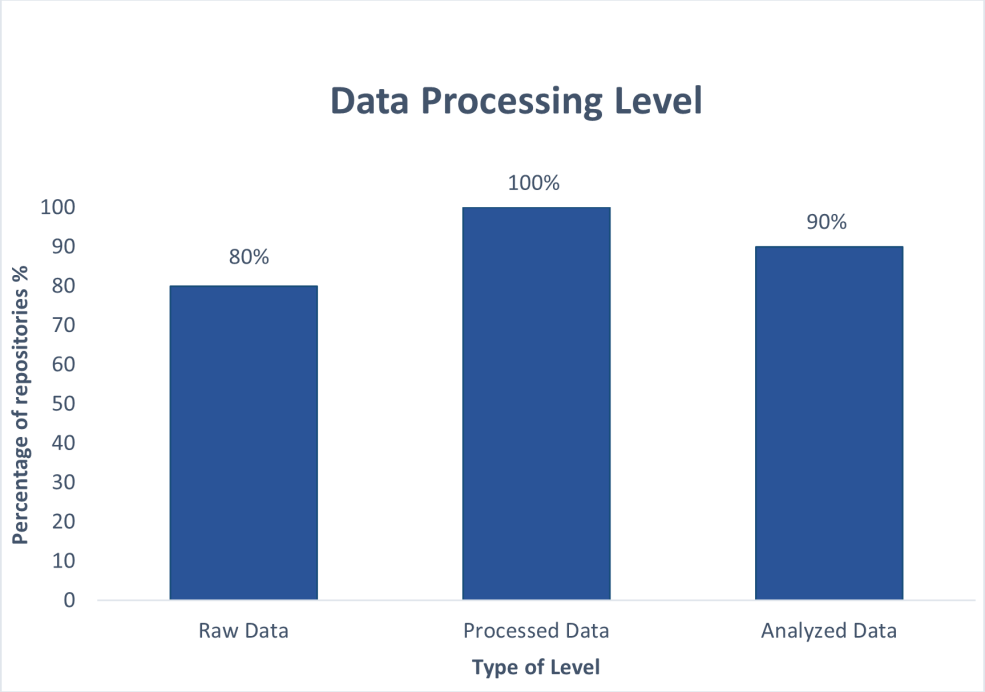

FIGURE 3. Data typology according to its data processing level. Own elaboration

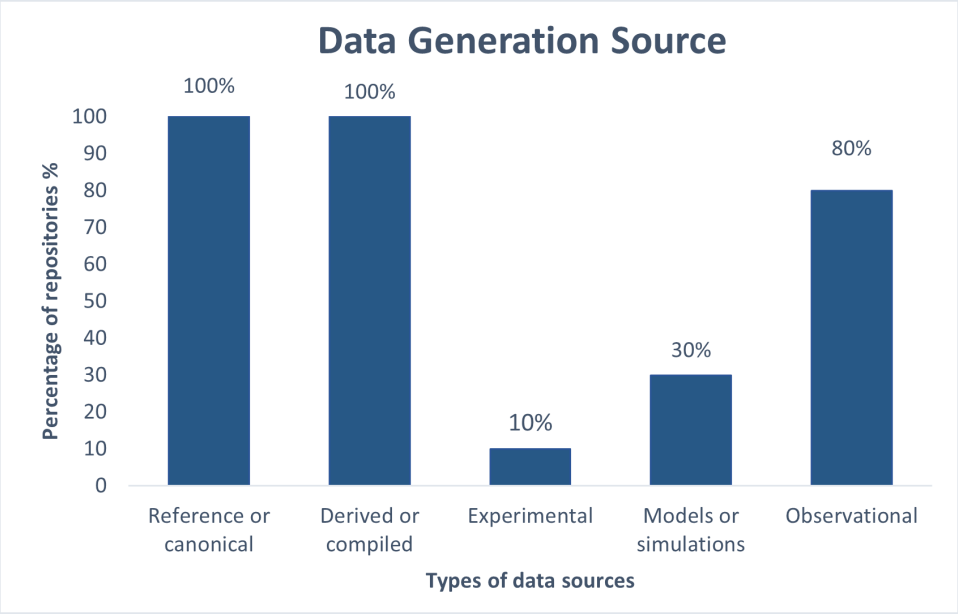

FIGURE 4. Data typology according to its data generation source. Own elaboration

On the other hand, the Dublin Core Metadata Initiative (DCMI) is the most followed by repositories by incorporating its Dublin Core (DC) vocabulary. However, in terms of the technology or platform hosting the repositories, there

is no adoption trend in the ten repositories reviewed since Dataverse and DSpace present the same percentage of use (20 %). In addition to the fact that there are repositories where the information on their platform or technology is not provided (20 %), as can be seen in Figure 6.

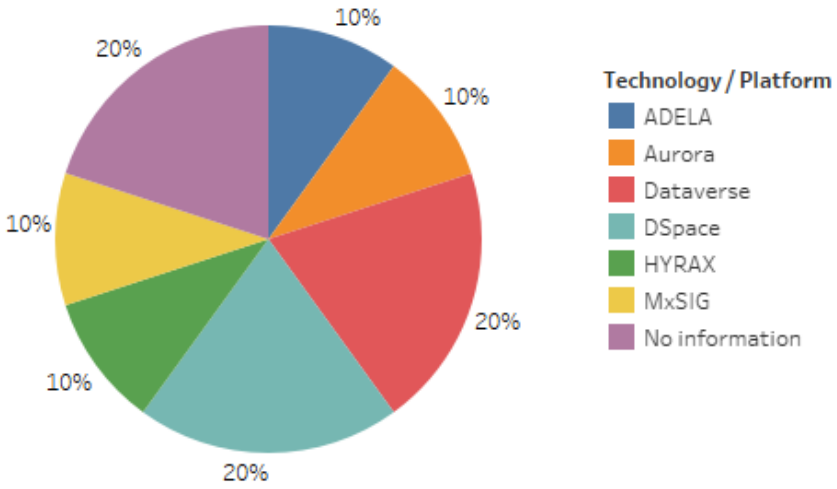

FIGURE 5. The platform or technology hosting the repository. Own elaboration

Finally, the widely popular Creative Commons (CC) licenses were the most widely used in repositories. However, some restrictions of use and access were found, such as 1) to freely use the data, the source of origin must be cited, 2) limitation of access to resources (which can be accessed upon request), and 3) data embargo.

Conclusions

This investigation allowed the identification of ten repositories promoting open data in Mexico and the integration of 16 characteristics that helped to describe and analyze the landscape of open data published in said repositories until May 12, 2023. The results of this research show that the different open data management efforts carried out by Mexican institutions are mostly consolidated since they incorporate metadata standards, platforms, and use

licenses that enable free data access, use, and sharing. However, some areas for improvement were identified, which are organized into four proposals:

1) Establish an agreement on the information to be provided and its structuring for a better description, analysis, and evaluation of the data and the repositories. Some recommendations are:
   a) Enrich the documentation to establish record guidelines and examples of the different data types. For example, in geospatial data, it is suggested to add the coordinates.
   b) Provide a base of descriptive characteristics of the repository. For example, the metadata standards followed and the platform or technology hosting the repository.
   c) Define the same classification of resources by areas of knowledge, type, file formats, among other aspects, such as those mentioned in this investigation, for their correct review and analysis. These categories could be presented as content filters.
2) Promote the integration of open data in a unique repository or in different repositories where it is sought to unify metadata standards and the hosting technology or platform to allow:
   a) Interoperability among data.
   b) Informed decision-making regarding an area or topic of interest based on the available data.
3) Strengthen the adoption of laws and agreements on the protection of personal data and the treatment of sensitive data. In this way, although it seeks to promote and facilitate access to data and its documentation, possible inappropriate use of them is being prevented, especially when the integrity of people may be affected.
4) Encourage the integration of more private institutions in the Open Data initiative of Mexico in order to promote the publication of their data and, in this way, support research boosted in favor of improving the quality of life in the country and the world.

As future work, it is proposed to continue completing the information that could not be obtained from the repository to ensure that all the characteristics defined in this investigation are documented. In addition, carry out an evaluation of these repositories through an evaluation standard widely used and reported in the literature.

ACKNOWLEDGMENTS. The authors thank Gerardo Castañeda Garza for his collaboration in the early phases of the project, as well as Rosalina Vázquez Tapia, Antonio Razo, and Martín Adalberto Tena Espinoza de los Monteros for the coordination and support within the framework of the transversal agenda REMERI-ANUIES-EduTraDi to which this research belongs.

References

CONACYT, C. (2022). Términos de referencia. Convocatoria de Ciencia Básica y/o Ciencia de Frontera: <https://conacyt.mx/wp-content/uploads/convocatorias/ciencia_de_frontera/paradigmas_y_controversias/2022/TdR_Paradigmas_y_Controversias_de_la_Ciencia_2022_VF.pdf>

COORDINACIÓN DE ESTRATEGIA DIGITAL NACIONAL. (2015). Datos.gob.mx: Impulsa los Datos Abiertos en Mexico. <https://datos.gob.mx/blog/datosgobmx-impulsa-los-datos-abiertos-en-mexico?category=noticias&tag=desarrollo>

DIGITAL GOVERNMENT UNIT, M. O. (2015). Digital government toolkit. National Digital Strategy. <https://www.oecd.org/gov/mexico-digital-strategy.pdf>

EUROPEAN DATA. (n. d.). What is open data? <https://data.europa.eu/elearning/en/module1/#/id/co-01>

GOBEN, A., & SANDUSKY, R. (2020). Open data repositories. Current risks and opportunities. <https://crln.acrl.org/index.php/crlnews/article/view/24273/32092and#:~:text=Another%20challenge%20is%20the%20potential,continue%20to%20pursue%20grant%20funding>

LEARN PROJECT, L. A. (2020). Gestión de datos de investigación. <https://biblioguias.cepal.org/gestion-de-datos-de-investigacion>

OPEN KNOWLEDGE FOUNDATION. (2015). Open Definition. <https://opendefinition.org/>

SCIENCE, D., GOODEY, G., HAHNEL, M., ZHOU, Y., JIANG, L., CHANDRAMOULISWARAN, I., DAY, L. (2022. The State of Open Data 2022. doi: <https://doi.org/10.6084/m9.figshare.21276984.v5>

THE WORLD BANK GROUP. (2012). Learning About the Open Data Initiative. <https://www.worldbank.org/en/news/feature/2012/03/22/learning-about-the-open-data-initiative>

THE WORLD BANK GROUP. (2021). Open Data in 60 Seconds. <http://opendatatoolkit.worldbank.org/en/open-data-in-60-seconds.html>