REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487 • ISSN 2708-3411
Vol. 2 • Nro. 1 • enero-marzo 2021 • págs. 144-158

COMUNICACIONES BREVES

Marco de trabajo para la publicación de Datos Abiertos en Cuba

Framework for Publication of Open Data in Cuba

Reynaldo Alvarez Luna
rluna@uci.cu • https://orcid.org/0000-0002-5279-7598

Héctor Raúl González Diez
hglez@uci.cu • https://orcid.org/0000-0002-7601-4201

UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS, CUBA

Alberto Torres Reyes
albertocorreoficial@gmail.com • https://orcid.org/0000-0003-0121-0213

UNIVERSIDAD DE ARTEMISA, CUBA

Alain Rodriguez Torres
alainrodrigez0904@gmail.com • https://orcid.org/0000-0003-2848-6711

MINISTERIO DE LAS COMUNICACIONES, CUBA

Recibido: 2020-11-16 • Aceptado: 2021-01-19

RESUMEN

Los datos abiertos promueven el desarrollo de la sociedad a partir de un marco tecnológico y legal que contribuye a la transparencia en la gestión del gobierno y la sociedad. La publicación de datos permite que los científicos y empresas innoven a partir de problemáticas locales y generen mejoras en los servicios y gestión de diversos sectores mediante el uso de métodos de inteligencia artificial y *Big Data*. La presente investigación aborda los principales conceptos asociados a los datos abiertos, la relevancia de su adopción como práctica gubernamental y su influencia en la relación academia-sociedad en función de la transformación social. Se exponen las principales características de las plataformas para la publicación de datos abiertos y los requisitos de los datos para ser publicados. El análisis de los criterios fundamentales identificados para la selección de herramientas para la preparación y publicación de los datos permitió proponer una solución adecuada para el entorno cubano. El desarrollo de un marco de trabajo mediante la personalización de sistemas para la publicación de datos abiertos basados en estándares, que permitan a las organizaciones cubanas la publicación de datos y

la consiguiente contribución al desarrollo de soluciones de análisis de datos. Se identifican posibilidades de desarrollo futuro alrededor de la adopción de estándares para la calidad de los datos abiertos y su influencia en un resultado confiable a partir de la aplicación de métodos de inteligencia artificial.

Palabras clave: datos abiertos; limpieza de datos, plataformas de datos abiertos.

ABSTRACT

Open data promotes the development of society based on a technological and legal framework that contributes to transparency in the management of government and society. The publication of data allows scientists and companies to innovate based on local problems and generate improvements in the services and management of various sectors through the use of Artificial Intelligence and Big Data methods. This research addresses the main concepts associated with open data, the relevance of its adoption as a government practice and its influence on the academy-society relationship in terms of social transformation. The main characteristics of the platforms for the publication of open data and the requirements of the data to be published are exposed. The analysis of the fundamental criteria identified for the selection of tools for the preparation and publication of the data made it possible to propose an adequate solution for the Cuban environment. The development of a framework through the customization of systems for the publication of open data based on standards, which allow Cuban organizations to publish data and the consequent contribution to the development of data analysis solutions. Possibilities for future development are identified around the adoption of standards for the quality of open data and its influence on a reliable result from the application of Artificial Intelligence methods.

Keywords: *data cleansing, open data; open data platforms.*

INTRODUCCIÓN

Los datos son el recurso más estratégico para la digitalización en la actualidad, un reciente artículo así lo aborda (Hartmann & Henkel, 2020). Resalta que es cada vez mayor el interés corporativo en el análisis de los datos y apuestan por el uso intensivo de métodos de ciencias de datos para la Transformación Digital de su entorno. En este esfuerzo promueven la aper-

tura de sus datos para tener variantes de soluciones y utilizar la creatividad de la comunidad científica.

Los datos abiertos son datos a los que cualquiera puede acceder, usar o compartir. Cuando las grandes empresas o gobiernos divulgan datos no personales, permite a las pequeñas empresas, ciudadanos e investigadores desarrollar recursos que hacen mejoras cruciales en sus comunidades (Pfenninger, *et al.,* 2017). Los datos abiertos del gobierno son datos producidos o encargados por el gobierno o entidades controladas por este, datos que están abiertos como se define en la Definición Abierta (*Open Knowledge Foundation,* s. f.), es decir, pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona (*European Data Portal,* s. f.).

La publicación de datos es una actividad que varias instituciones gubernamentales están poniendo en práctica con el objetivo de hacer que sus datos sean visibles y reutilizables por la comunidad. En un artículo reciente sobre la conceptualización de los datos abiertos en la transformación digital (Pascual, *et al.,* 2020) se reconocen las principales fuentes de datos abiertos y las iniciativas que desde el reconocimiento del valor de los datos han surgido para moderar la publicación de datos a nivel global.

Este es el poder de los datos abiertos: el poder de encontrar problemas en entornos complicados, y posiblemente incluso de evitar que surjan. Estudios de casos han demostrado la influencia positiva a nivel social de publicar datos de actividades públicas para la prevención de problemas y para la ejecución de análisis correlacionales entre diversos sectores que comúnmente no colaboran en sus dinámicas habituales (Ruijer, *et al.,* 2017). Esta situación también está presente en organizaciones que realizan actividades de investigación, en la que sus investigadores tienen la necesidad de publicar sus trabajos de investigación, así como los datos resultantes de estos trabajos con el objetivo de que esta información sea reutilizada por la comunidad científica (Mueller-Langer & Andreoli-Versbach, 2018).

La implementación de datos abiertos, a nivel gubernamental resulta ser una buena práctica de transparencia en la actividad política hacia los ciudadanos, los autores (Afful-Dadzie & Afful-Dadzie, 2017) exploran los temas fundamentales alrededor de este fenómeno y su influencia en la libertad de investigación. Varios gobiernos y organizaciones a nivel mundial publican datos, por ejemplo, las cifras de gasto para un departamento gubernamental, las lecturas de temperatura de varias estaciones meteorológicas, o incluso publican problemas para promover soluciones de análisis de datos que mejoran sus procesos desde la comunidad científica internacional. Safarov y colaboradores (2017) y Ruijer y colaboradores (2020) muestran un análisis de los principales países que fomentan el uso de datos abiertos gubernamentales y los principales efectos de su utilización por los usuarios de acuerdo a su rol social.

Algunos ejemplos de plataformas para la publicación de datos son:

- Portal de datos abiertos de Europa[1] tiene 7054 *datasets* y se pueden encontrar más de 100 relacionados con la COVID-19. Diversos trabajos (Alamo, *et al.,* 2020; Cheng, *et al.,*

[1] https://data.europa.eu

2020) han manifestado el fuerte movimiento de apertura de los datos relativos a la COVID-19 que han permitido un avance acelerado de las investigaciones médicas y sociales alrededor de la enfermedad.

- Iniciativa de datos abiertos del Gobierno de España (Gobierno de España, s. f.);
- Portal de datos del Gobierno de Estados Unidos (Krishnamurthy & Awazu, 2016);
- Datos abiertos en México (Quintanilla & Gil-García, 2016). Particularmente México ha mantenido la publicación diaria de estadísticas epidemiológicas que han permitido el desarrollo de numerosas investigaciones relativas al comportamiento de la COVID-19 (Salud, s. f.) (Medel-Ramírez & Medel-Lopez, 2020; Melin, *et al.,* 2020).

Un estudio de la CEPAL (Naser & Rosales, 2016) refleja las principales iniciativas en América Latina y el Caribe, valoran su estado en función del uso de los datos abiertos, abordan el uso de plataformas de código abierto y las normativas establecidas por los gobiernos para la publicación de datos. El informe de la Organización de las Naciones Unidas tiene la función de promover la colaboración y establece algunos indicadores para medir el avance los países del área en este sentido.

La transformación digital de la sociedad cubana ha sentado las bases para el crecimiento de los datos en todas las esferas y sectores del país. En sectores claves tales como la salud, el transporte, las estadísticas, la banca y las telecomunicaciones existen grandes volúmenes de datos que están siendo acumulados. Existen algunos esfuerzos y soluciones de análisis de datos que permiten la mejora de procesos y toma de decisiones.

La coherente interacción entre los científicos y el gobierno es clave en la gestión gubernamental. Particularmente Bermúdez y Jover (2020) exponen el impacto positivo la gestión de los problemas en el enfrentamiento a la COVID-19, a partir de la colaboración interdisciplinaria y la ejecución de varios proyectos, que requieren el uso de métodos y técnicas de inteligencia artificial y *Big Data* para la construcción de modelos predictivos, análisis de dinámicas de la población y visualización científica de los datos. Con el desarrollo de la tecnología han crecido las brechas en el acceso a la información para el desarrollo de investigaciones básicas y aplicadas. Ha sido un reclamo del gobierno cubano la aplicación práctica de las investigaciones para el desarrollo del país. Sin embargo, en muchas áreas del conocimiento es más accesible la información de otros sectores y países que la propia de Cuba, debido en gran medida a la apertura de los datos que existe en otros países.

En Cuba a través del portal de la Oficina Nacional de Estadística e Información (ONEI) (ONEI, s. f.) se publican datos de diferentes ámbitos sociales, dígase: agricultura, salud, deporte, educación, seguridad social, entre otros. La recuperación y análisis de esos datos mediante la utilización herramientas informáticas se torna compleja debido a la heterogeneidad de los formatos utilizados para su publicación. De igual modo, se imposibilita su integración ya que cada uno de los dominios tiene su propia estructura. Este tipo de soluciones están limitadas por el bajo nivel de acceso de la comunidad científica cubana a los datos gubernamentales o de empresas públicas claves para el desarrollo de los sectores estratégicos para el desarrollo del país.

En la recientemente publicación "Síntesis de la Estrategia Económico-Social para el impulso de la economía y el enfrentamiento a la crisis mundial provocada por la COVID-19", que atempera y prioriza las acciones para el cumplimiento de los lineamientos del VII congreso del PCC (Escandell-Sosa, 2016), se plantea para el sector de las telecomunicaciones como una de las acciones claves: "Implementar los 9 proyectos iniciales vinculados con el control de epidemias, la planificación del transporte, la evasión fiscal y los estados de opinión, entre otros, con el uso de herramientas informáticas para el manejo de grandes volúmenes de datos, *Big Data*, con la incorporación del Centro de Sistemas Complejo y *Big Data* de la Universidad de La Habana" (Ministerio de Economía y Planificación (MEP), 2020).

Es transversal a todos los proyectos mencionados el uso intensivo de los datos que utiliza métodos de inteligencia artificial, la explotación de infraestructuras de alto procesamiento de datos y el trabajo coordinado de las entidades del gobierno que gestionan los procesos, con las universidades y empresas del sector del *software*. Si se toma como punto de referencia los elementos anteriormente descritos, se puede plantear como contribución fundamental de la investigación:

- Desarrollar un marco de trabajo mediante la personalización de sistemas para la publicación de datos abiertos basados en estándares, que permitan a las organizaciones cubanas la publicación de datos y la consiguiente contribución al desarrollo de soluciones de análisis de datos.
- Desarrollar una plataforma que soporte los procesos de limpieza y extracción de datos y un marco común de publicación que permita desarrollar investigaciones en el campo de la minería datos y el *Big Data*.

METODOLOGÍA

El estudio fue realizado en dos etapas, la primera consistió en una revisión de herramientas y tecnologías, se sintetizaron las principales características y funcionalidades deseables en este tipo de solución. La segunda etapa consistió en la evaluación experimental de la herramienta identificada como más completa de acuerdo al entorno y los requisitos identificados. El análisis realizado parte de la evaluación de la calidad de datos, los principales métodos de limpieza y anonimización de datos hasta la evaluación de las herramientas para el soporte de estos procesos, imprescindibles para la publicación de los datos.

Calidad de datos

La publicación de los datos no se puede tomar a la ligera, es necesario que se haga bajo estándares que garanticen la calidad y permitan la interoperabilidad de estos. Los datos abiertos de alta calidad son una condición previa para analizarlos, reutilizarlos y garantizar su valor. Es importante evaluar la calidad de los datos para verificar su exactitud, confiabilidad y aptitud para el uso, es decir, su capacidad de prestar adecuadamente una función o servicio. En conclusión, los datos deben permitir transformarse, georreferenciar, analizar, reutilizar, visualizar y agregar sin necesidad de depurarse previamente (Abella, *et al.*, 2018).

En los trabajos del área de Calidad de Datos existe un núcleo de dimensiones que es compartido por la mayoría de las propuestas, las principales dimensiones que forman parte este son: confidencialidad, relevancia, actualidad, trazabilidad, conformidad, exactitud, completitud, consistencia, precisión, portabilidad, credibilidad, comprensibilidad, accesibilidad, disponibilidad y unicidad (Abdullah & Arshah, 2018; Hassine & Clément, 2020; Nasr, *et al.,* 2020). Con el fin de garantizar estas dimensiones, múltiples herramientas, tanto de limpieza como de publicación de datos, son empleadas. Usualmente se garantizan mediante tareas de transformación de los datos que son responsabilidad del usuario que los publica. Este se apoya en guías de publicación otorgadas por las entidades que despliegan las herramientas y los diferentes métodos que implementan las mismas.

Herramientas de limpieza de datos

La limpieza de datos es una tarea clave, en el artículo de Sadiq & Indulska (2017) se resalta la importancia de la calidad sobre la cantidad de datos. Un extenso volumen de datos sin la calidad requerida complejiza su análisis y consume más recursos computacionales. Las herramientas de limpieza de datos intentan resolver todos los problemas derivados de datos sucios mediante variados métodos. Como son:

- la conversión y normalización de funciones que transforman y estandarizan formatos de datos heterogéneos,
- la limpieza de propósito-especial que limpia campos específicos a través del uso de diccionarios para buscar sinónimos,
- la limpieza de dominio-independiente que aplica algoritmos de emparejamiento de campos a los campos equivalentes de las fuentes diferentes para decidir cómo emparejarlos,
- la limpieza basada en reglas que está basada en un conjunto de “reglas de negocio” que especifican las reglas en las cuales dos valores de diferentes fuentes se corresponden.

Trabajos recientes de revisión (Abdullah & Arshah, 2018; Henriques, *et al.,* 2020; L’Heureux, *et al.,* 2017) plantean entre los desafíos en el avance de dig data son: la limpieza y calidad de los datos, así como la influencia de este proceso en la privacidad cuando se realiza en entornos de datos abiertos. La manera clásica de comenzar la limpieza de datos es hacer un cómputo de los registros duplicados que se deben eliminar, o ejecutar una comparación total de todos los registros. Algunos de las criterios fundamentales para la revisión de este tipo de herramientas son: funcionalidades principales, plataforma, escalabilidad, nivel de habilidad necesario, curva de aprendizaje y extensibilidad (Samson Oni, *et al.,* 2019). Particularmente en este estudio se ha adicionado su posibilidad de integración a partir de APIs y que sean de código abierto como un aspecto esencial.

Algunos análisis comparativos (Petrova-Antonova & Tancheva, 2020; Samson Oni, *et al.,* 2019) caracterizan y comparan las herramientas *Trifacta, OpenRefine* y otras básicas como *R, Python.* En el presente artículo se han analizado otras en función de valorar algunas menos mencionadas en la literatura y que pudiera resultar de interés en cuanto a sus funcionalida-

des. La tabla 1 expone en resumen la comparación de las herramientas *Trifacta* (Trifacta, s. f.), *Paxata* (Paxata, s. f.), *Alteryx* (Alteryx, s. f.) y *OpenRefine* (*OpenRefine*, s. f.).

Tabla 1. Comparación de herramientas para el tratamiento de calidad de datos.

Aplicación	Curva de aprendizaje	Código abierto	API de desarrollo	Gratuita	Extensiones
Trifacta	Baja	No	No	tiene limitaciones de uso gratis	No
Paxata	Alta	No	No	tiene limitaciones de uso gratis	No
Alteryx	Baja	No	No	tiene limitaciones de uso gratis	No
OpenRefine	Baja	Sí	Sí	Sí	Sí

Se determinó que la herramienta *OpenRefine* satisface las necesidades planteadas para el tratamiento de la calidad de datos y es la más completa a emplear y adaptar al marco de trabajo, que se realiza debido a su API de desarrollo, su código abierto y además posee una extensión que la vincula directamente con CKAN. *OpenRefine* es el paradigma creado por Google que posee implementadas las principales funcionalidades necesarias para garantizar las dimensiones de calidad de datos anteriormente abordadas. Y a diferencia de los otros sistemas estudiados posee una amplia comunidad que la respalda y desarrolla complementos para aumentar así su accionar.

Sistemas para la publicación de datos abiertos

Las plataformas de publicación de datos abiertos son piezas de software que facilitan la publicación y gestión de datos en la web. Para los editores, una plataforma de datos abiertos proporciona una vía para publicar datos. Las plataformas guían a los editores a través de procesos de publicación de datos, y ofrecen a los usuarios consistencia y facilidad de acceso a datos abiertos desde cualquier parte del mundo. Existen varias plataformas de datos abiertos que son utilizadas por entidades gubernamentales e institutos privados[2]. La tabla 2 resume las plataformas que han sido ampliamente utilizadas en el campo de los datos abiertos: CKAN (CKAN, s. f.), DKAN (DKAN, s. f.), JUNAR (JUNAR, s. f.) y SOCRATA (Socrata, s. f.). Estas fueron estudiadas teniendo en cuenta criterios como que posean código abierto, con el objetivo de realizar modificaciones en este, en caso de que sea necesario, como se realiza la gestión de datos, para garantizar que sea factible adaptar la misma a las condiciones de nuestra red nacional y a la solución que se pretende llegar en este proyecto, así como los lenguajes de programación que emplea la comunidad que le brinda soporte y si es flexible a cambios.

Las herramientas estudiadas en su totalidad presentan las funcionalidades indispensables para una correcta publicación de datos abiertos al público. Después de interactuar con las mismas y realizar la comparación anterior, se definió que CKAN es la solución adecuada para personalizar en el entorno cubano. Los criterios para su selección ante DKAN es primeramente debido a la posibilidad de integración con herramientas como *OpenRefine*, su desarrollo en

[2] http://bit.ly/2XwQDMz ,http://bit.ly/3bCljEf , http://bit.ly/3bBjZS2

Tabla 2. Comparativa de las plataformas de datos abiertos

Producto	Creador	Tipo	Gestión de datos	Soporte/ Comunidad	Flexibilidad ante cambios
CKAN	Open Knowledge Foundation	Código Abierto (cloud hosting available)	Local o Federado	Python developer community	Sí
DKAN	Nuams	Código Abierto (cloud hosting available)	Local o Federado	Drupal developer community	Sí
JUNAR	JUNAR	SaaS	Local	Vendor	No
SOCRATA	SOCRATA	SaaS	Local o Federado	Vendor	No

Python posibilita una mayor contribución de la comunidad científica que mayoritariamente utiliza Python en sus rutinas habituales. CKAN ha sido utilizado en un amplio número de portales institucionales o gubernamentales de probada funcionalidad y aceptación por parte de los usuarios. Un estudio comparativo reciente (Milic, *et al.,* 2018) muestra como CKAN tiene un mejor manejo de los metadatos y mayor cantidad de funcionalidades expuestas a través de su API.

Aspectos éticos en la publicación de datos

Diversos trabajos genéricos y sectorizados han abordado los problemas o dilemas éticos en la publicación de los datos. En un ecosistema de datos es clave la gobernanza en el acceso a los *datasets* y en este proceso no se puede perder de vista aspectos éticos tales como: privacidad, responsabilidad, propiedad, accesibilidad y motivación (Rantanen, *et al.,* 2019). La vigencia está soportada en por diversas razones, particularmente está relacionado con el Gobierno Electrónico (Ronzhyn & Wimmer, 2019):

- La relación entre el gobierno y los ciudadanos es desigual: el ciudadano es dependiente y vulnerable.
- Las TIC tienen un efecto sobre los valores públicos y su potencial transformador también debe verse en esta dimensión.
- El panorama de la esfera pública es diferente del ámbito privado, ya que los objetivos finales de las organizaciones involucradas son muy diferentes.

Diversos autores abordan el tratamiento de los datos abiertos como un elemento básico en la transparencia de la gestión del gobierno, a pesar de los dilemas éticos con los que hay que lidiar coherentemente con la tecnología para impulsar el desarrollo y la influencia del gobierno en la transformación de la sociedad (Flasher, 2019; Hardy & Maurushat, 2017; Kassen, 2017; Ruijer & Meijer, 2020; Safarov *et al.,* 2017).

Los elementos mencionados avalan la factibilidad del desarrollo de soluciones de este tipo en Cuba, debido a sus aspiraciones de construir una sociedad cada vez más en línea y conectada con su gobierno a través de las tecnologías de la información. Las herramientas seleccionadas son de código abierto por lo que se minimizan los riesgos en su adopción ante cualquier cambio en su forma de licencia y las limitaciones que tiene Cuba por el bloqueo.

RESULTADOS Y DISCUSIÓN

La solución propuesta consiste en la configuración de un marco de trabajo tecnológico que contribuya al preprocesamiento de los datos y su publicación. Varios autores como Henriques y colaboradores (2020) y L'Heureux y colaboradores (2017) han identificado que las tareas de limpieza de datos aún requiere de grandes esfuerzos en la primera fase de cualquier proyecto de análisis de datos. La integración de las herramientas OpenRefine y CKAN cubrirá los aspectos de preparación de los datos y su publicación. Normativas y políticas deben ser diseñadas como complemento para garantizar cuestiones metodológicas, confidencialidad y la ética de los usuarios de la solución.

CKAN y OpenRefine tienen una arquitectura modular (componentes) a través de la cual se pueden añadir funcionalidades a la medida de los usuarios. Por lo tanto, se propone la arquitectura basada en componentes para el desarrollo (Figura 1).

Se utiliza además el estilo arquitectónico en capas, el cual brinda un cierto aislamiento entre las distintas capas de la aplicación, de forma tal que cualquier cambio o modificación que ocurra en una de ellas no afecte al resto. Permite el desarrollo en paralelo, o sea que se puede ir trabajar en diferentes capas de forma separada, lo cual posibilita agilizar el desarrollo.

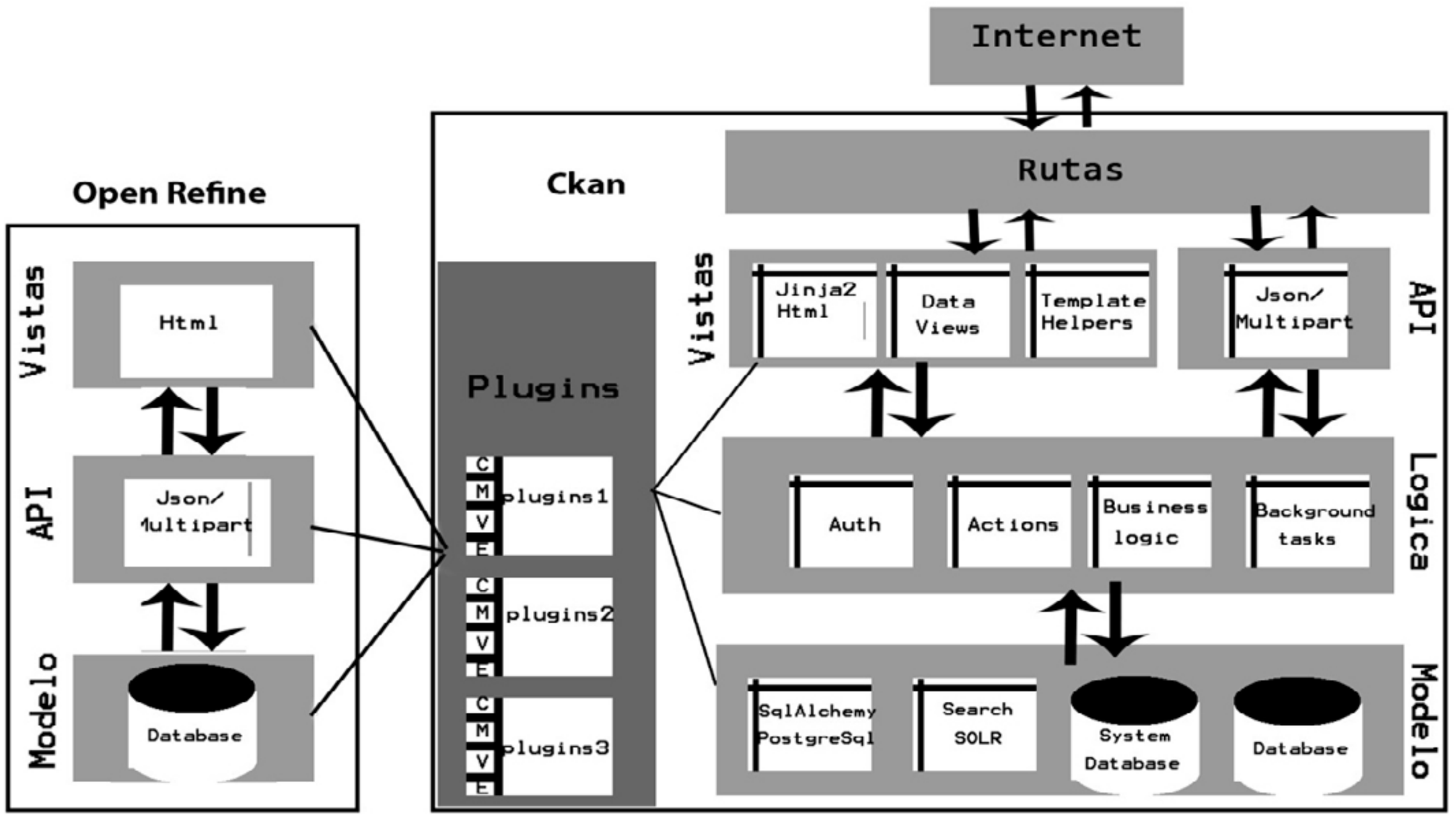

Figura 1. Diseño arquitectónico de la solución propuesta. Fuente: Adaptado de CKAN (s. f.-b).

Los requerimientos para el desarrollo del marco de trabajo se han identificado a partir de la observación de las funcionalidades de soluciones tecnológicas y en la revisión de fuentes bibliográficas en *Google Académico* encontradas bajo los términos de búsqueda *open government data initiatives* y *open data platforms.* Al realizar un análisis de las características funcionales de las herramientas observadas se identificaron 18 requisitos funcio-

nales prioritarios que debe tener el marco de trabajo. Los requisitos abordan la gestión de usuarios y grupos, el flujo de publicación y el proceso de visualización y descarga. El análisis de bibliográfico por su parte fue particularmente útil en la identificación de requisitos no funcionales de privacidad, legales, seguridad y almacenamiento como cuestiones críticas en este tipo de solución.

La comunidad de CKAN formada por un amplio grupo de desarrolladores, establece las pautas para el desarrollo de pruebas. En su última versión la documentación oficial incluye pasos detallados para el desarrollo de pruebas en todos los niveles de la aplicación (CKAN, s. f.-c). Varios trabajos previos han comprobado la calidad de CKAN a través de pruebas específicas de carga y rendimiento a sus diferentes componentes que validan su calidad (Winn, et al, 2013). Particularmente en el trabajo Herrera-Cubides y colaboradores (2019) se hace una profunda evaluación de los criterios: nivel de confianza, calidad, vinculación y usabilidad de los *Linked Open Data* de varias instancias de CKAN. Se concluye que funcionalmente CKAN satisface los requerimientos de calidad y funcionales esperados, se identifica que los principales problemas son relativos a la calidad de los datos.

Un estudio de caso realiza una comparación en cuanto a la experiencia de usuarios y usabilidad a partir de la opinión de expertos, que resalta la calidad de CKAN en este aspecto (Akyürek, *et al.,* 2018). Por su parte, Herrera-Cubides y colaboradores (2017) describen la realización de pruebas al API de servicios de CKAN y concluyen que "los servicios ofrecidos por CKAN, para la consulta y descarga de la data, son servicios consistentes, que ofrecen peticiones y respuestas en formato JSON, sin restricciones al público".

Uno de los estudios revisados realiza la valoración más completa de varias herramientas de gestión de *Open Data* en varios de los aspectos resalta el comportamiento de CKAN sobre otras alternativas a pesar de estar en un estado de desarrollo inicial (Str\aale & Lindén, 2014). El estudio destaca el soporte a múltiples formatos de datos, el rendimiento de su API y las capacidades de visualización de los datos.

Dada la revisión bibliográfica acerca la calidad de CKAN en varias características, las pruebas realizadas en la investigación se realizaron en una instancia local de la aplicación para comprobar la satisfacción de los requerimientos derivados del estudio inicial. La instalación se hizo a nivel de laboratorio basado en la guía de la documentación oficial. Se comprueba que la aplicación satisface los requerimientos y que es factible dada la documentación disponible desarrollar cambios menores para la personalización de la solución en cuanto al diseño de la interfaz de usuarios y otras funcionalidades.

Como trabajo futuro, se ha identificado la definición de políticas y normativas que regulen el proceso de publicación de datos y la integración de CKAN con tecnologías y plataformas de análisis de datos tales como *Apache Spark, Hadoop* u otras para automatizar los procesos de importación de los datasets para su análisis o en función de realizar tareas de verificación de la calidad de estos. Herramientas como *Dataverse* (Dataverse.org, s. f.) y *OpenDataSoft* (Opendatasoft, s. f.) han evolucionado a partir del uso en entornos o países específicos y pueden ser valoradas en otros estudios de este tipo.

La calidad de los datos abiertos es un área de trabajo en desarrollo, numerosos estándares surgen para habilitar aún más la colaboración entre ecosistemas digitales (Rudmark, 2020). Ante la madurez de herramientas como CKAN el problema consiste en garantizar que los datasets públicos sean útiles para "aprender" y ese es otro de los retos que plantea las recientes investigaciones acerca de Trust AI (Gillath, *et al.,* 2020) y Xplanaible AI (Das & Rad, 2020), áreas claves para la búsqueda de la transformación digital coherente, colaborativa y con ética científica.

CONCLUSIONES

A partir del estudio realizado sobre los criterios de calidad requeridos para la publicación de conjuntos de datos y los sistemas homólogos se determinó la necesidad de personalizar un sistema que integrara dos herramientas libres, de código abierto, CKAN y *OpenRefine.* De esta manera se definió una propuesta de solución de acuerdo a las necesidades existentes.
La implementación de un sistema para la publicación de datos abiertos en Cuba que utilice las herramientas y tecnologías estudiadas permite sentar las bases para la solución de problemáticas sociales y empresariales a partir de la colaboración con la academia. La solución planteada contribuye a la gestión del conocimiento y al soporte a investigaciones en el área de la minería de datos, el aprendizaje automatizado y el *Big Data* sobre problemas del contexto cubano, así como la solución a proyectos ya identificados en la estrategia de desarrollo del país.

REFERENCIAS

Abdullah, M. Z., & Arshah, R. A. (2018). A Review of Data Quality Assessment: Data Quality Dimensions from Users Perspective. *Advanced Science Letters, 24*(10), 7824-7829. https://doi.org/doi:10.1166/asl.2018.13025

Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2018). Open data quality metrics: Barcelona open data portal case/INDICADORES DE CALIDAD DE DATOS ABIERTOS DE BARCELONA. *El Profesional de la Informacion, 27*(2), 375-383.

Afful-Dadzie, E., & Afful-Dadzie, A. (2017). Liberation of public data: Exploring central themes in open government data and freedom of information research. *International Journal of Information Management, 37*(6), 664-672. https://doi.org/10.1016/j.ijinfomgt.2017.05.009

Akyürek, H., Scholl, C., Stodden, R., Siebenlist, T., & Mainka, A. (2018). Maturity and usability of open data in North Rhine-Westphalia. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age,* 1-10.

Alamo, T., Reina, D. G., Mammarella, M., & Abella, A. (2020). Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic. *Electronics, 9*(5), 827. https://doi.org/10.3390/electronics9050827

Alteryx. (s. f.). Alteryx. Recuperado 7 de noviembre de 2020, de https://www.alteryx.com/

Bermúdez, M. D.-C., & Jover, J. N. (2020). Gestión gubernamental y ciencia cubana en el enfrentamiento a la COVID-19. *Anales de la Academia de Ciencias de Cuba, 10*(2), 881.

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020). COVID-19 Government Response Event Dataset (CoronaNet v.1.0). *Nature Human Behaviour, 4*(7), 756-768. https://doi.org/10.1038/s41562-020-0909-7

CKAN. (s. f.-a). *CKAN*. Ckan. Recuperado 7 de noviembre de 2020, de https://ckan.org/

CKAN. (s. f.-b). *CKAN code architecture.* Recuperado 20 de enero de 2021, de https://docs.ckan.org/en/2.9/contributing/architecture.html

CKAN. (s. f.-c). *Testing CKAN — CKAN 2.10.0a documentation.* Testing CKAN — CKAN 2.10.0a documentation. Recuperado 21 de enero de 2021, de https://ckan.readthedocs.io/en/latest/contributing/test.html

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371.*

Dataverse. (s. f.). *The Dataverse Project—Dataverse.org.* Recuperado 19 de enero de 2021, de https://dataverse.org/

DKAN. (s. f.). *DKAN Open Data Platform.* Recuperado 7 de noviembre de 2020, de http://getkdkan.org/

Escandell-Sosa, V. E. (2016). Lineamientos de la Política Económica y Social del Partido y la Revolución aprobados en el VI Congreso del Partido Comunista de Cuba: Una visión desde la Economía Política. *Anuario Facultad de Ciencias Económicas y Empresariales, 3,* 51-60.

European Data Portal. (s. f.). *What is open data | European Data Portal.* Recuperado 23 de septiembre de 2020, de https://www.europeandataportal.eu/en/training/what-open-data

Flasher, R. (2019). Sunshine to Government—Opportunities for Engagement with Government Data. *Journal of Emerging Technologies in Accounting, 17*(1), 57-62. https://doi.org/10.2308/jeta-52654

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2020). Attachment and trust in artificial intelligence. *Computers in Human Behavior, 115,* 106607.

Gobierno de España. (s. f.). *Datos.gob.es.* Recuperado 18 de enero de 2021, de https://datos.gob.es/

Hardy, K., & Maurushat, A. (2017). Opening up government data for Big Data analysis and public benefit. *Computer Law & Security Review, 33*(1), 30-37. https://doi.org/10.1016/j.clsr.2016.11.003

Hartmann, P., & Henkel, J. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries, 6*(3), 359-381.

Hassine, S. B., & Clément, D. (2020). Open Data Quality Dimensions and Metrics: State of the Art and Applied Use Cases. En W. Abramowicz & G. Klein (Eds.), *Business Information Systems Workshops* (pp. 311-323). Springer International Publishing.

Henriques, A. C. V., Meirelles, F. de S., & Cunha, M. A. V. C. da. (2020). Big data analytics: Achievements, challenges, and research trends. *Independent Journal of Management & Production, 11*(4), 1201-1222.

Herrera-Cubides, J. F., Gaona-García, P. A., Montenegro-Marín, C. E., Varón-Capera, Á., et al. (2019). Confidence level evaluation of LOD resources on CKAN instances. *Visión electrónica, 13*(2).

Herrera-Cubides, J. F., Gaona-García, P. A., & Orjuela, K. G. (2017). A view of the web of data. Case study: Use of services CKAN. *Ingeniería, 22*(1), 46-64.

JUNAR. (s. f.). *Junar Data Platform.* Recuperado 7 de noviembre de 2020, de https://www.junar.com/

Kassen, M. (2017). Understanding transparency of government from a Nordic perspective: Open government and open data movement as a multidimensional collaborative phenomenon in Sweden. *Journal of Global Information Technology Management, 20*(4), 236-275. https://doi.org/10.1080/1097198X.2017.1388696

Krishnamurthy, R., & Awazu, Y. (2016). Liberating data for public value: The case of Data. Gov. *International Journal of Information Management, 36*(4), 668-672.

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access, 5,* 7776-7797. https://doi.org/10.1109/ACCESS.2017.2696365

Medel-Ramírez, C., & Medel-Lopez, H. (2020). *Data Mining for the Study of the Epidemic (SARS-CoV-2) COVID-19: Algorithm for the Identification of Patients (SARS-CoV-2) COVID 19 in Mexico.* Available at SSRN 3619549. http://dx.doi.org/10.2139/ssrn.3619549

Melin, P., Monica, J. C., Sanchez, D., & Castillo, O. (2020). Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of Mexico. *Healthcare, 8*(2), 181.

Milic, P., Veljkovic, N., & Stoimenov, L. (2018). Comparative analysis of metadata models on e-government open data platforms. *IEEE Transactions on Emerging Topics in Computing,* 1-1. https://doi.org/10.1109/TETC.2018.2815591

Ministerio de Economía y Planificación. (2020). *Cuba y su desafío económico y social* [Guvernamental]. Ministerio de Economía y Planificación de Cuba. Recuperado de https://www.mep.gob.cu/es/documento/cuba-y-su-desafio-economico-y-social

Mueller-Langer, F., & Andreoli-Versbach, P. (2018). Open access to research data: Strategic delay and the ambiguous welfare effects of mandatory data disclosure. *Information Economics and Policy, 42,* 20-34.

Naser, A., & Rosales, D. (2016). Panorama regional de los datos abiertos: Avances y desafíos en América Latina y el Caribe. Serie Gestión Pública. *Serie Gestión Pública,* 86, 125.

Nasr, M., Shaaban, E., & Gabr, M. I. (2020). Data Quality Dimensions. En A. Z. Ghalwash, N. El Khameesy, D. A. Magdi, & A. Joshi (Eds.), *Internet of Things—Applications and Future* (pp. 201-218). Springer Singapore.

Open Knowledge Foundation. (s. f.). *The Open Definition—Open Definition—Defining Open in Open Data, Open Content and Open Knowledge.* Recuperado 11 de enero de 2021, de https://opendefinition.org/

Opendatasoft. (s. f.). *Opendatasoft—Make your data bright.* Recuperado 19 de enero de 2021, de https://www.opendatasoft.com/es/

OpenRefine. (s. f.). *OpenRefine.* Recuperado 7 de noviembre de 2020, de https://openrefine.org/

Pascual, A. F. R., Sánchez, C. S., y Borreguero, J. M. R. (2020). Los datos abiertos: Definición técnica de un concepto clave para la Transformación Digital: Open data: technical definition of a key concept for Digital Transformation. *Revista Cubana de Transformación Digital, 1*(2), 7-22.

Paxata. (s. f.). *Paxata.* Recuperado 7 de noviembre de 2020, de https://www.paxata.com/

Petrova-Antonova, D., & Tancheva, R. (2020). Data Cleaning: A Case Study with OpenRefine and Trifacta Wrangler. En M. Shepperd, F. Brito e Abreu, A. Rodrigues da Silva, & R. Pérez-Castillo (Eds.), *Quality of Information and Communications Technology* (pp. 32-40). Springer International Publishing. https://doi.org/10.1007/978-3-030-58793-2_3

Pfenninger, S., DeCarolis, J., Hirth, L., Quoilin, S., & Staffell, I. (2017). The importance of open data and software: Is energy research lagging behind? *Energy Policy, 101,* 211-215.

Quintanilla, G., y Gil-García, J. R. (2016). Gobierno abierto y datos vinculados: Conceptos, experiencias y lecciones con base en el caso mexicano. *Revista del clad Reforma y Democracia, 65,* 69-102.

Rantanen, M. M., Hyrynsalmi, S., & Hyrynsalmi, S. M. (2019). Towards Ethical Data Ecosystems: A Literature Study. *2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC),* 1-9. https://doi.org/10.1109/ICE.2019.8792599

Ronzhyn, A., & Wimmer, M. A. (2019). Literature review of ethical concerns in the use of disruptive technologies in government 3.0. *The Thirteenth International Conference on Digital Society and eGovernments, ICDS,* 85-92.

Rudmark, D. (2020). Open Data Standards: Vertical Industry Standards to Unlock Digital Ecosystems. *53rd Hawaii International Conference on System Sciences.* EE.UU.

Ruijer, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly, 34*(1), 45-52.

Ruijer, E., Grimmelikhuijsen, S., van den Berg, J., & Meijer, A. (2020). Open data work: Understanding open data usage from a practice lens. *International Review of Administrative Sciences, 86*(1), 3-19.

Ruijer, E., & Meijer, A. (2020). Open Government Data as an Innovation Process: Lessons from a Living Lab Experiment. *Public Performance & Management Review, 43*(3), 613-635. https://doi.org/10.1080/15309576.2019.1568884

Sadiq, S., & Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management, 37*(3), 150-154. https://doi.org/10.1016/j.ijinfomgt.2017.01.003

Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity, 22*(1), 1-24.

Salud, S. de. (s. f.). *Datos Abiertos Dirección General de Epidemiología.* gob.mx. Recuperado 18 de enero de 2021, de http://www.gob.mx/salud/documentos/datos-abiertos-152127

Samson Oni, Zhiyuan Chen, Susan Hoban, & Onimi Jademi. (2019). A Comparative Study of Data Cleaning Tools. *International Journal of Data Warehousing and Mining (IJDWM), 15*(4), 48-65. https://doi.org/10.4018/IJDWM.2019100103

Socrata. (s. f.). Socrata. Recuperado 7 de noviembre de 2020, de https://www.tylertech.com/products/socrata

Str\aale, J., & Lindén, H. (2014). *An evaluation of platforms for open government data.*

Trifacta. (s. f.). *Trifacta.* Recuperado 7 de noviembre de 2020, de https://www.trifacta.com/

Winn, J. & others. (2013). *Open data and the academy: An evaluation of CKAN for research data management.*