

8^{VO} CONGRESO NACIONAL INGENIERÍA INFORMÁTICA / SISTEMAS DE INFORMACIÓN

VIRTUAL 2020 CONAISI



www.sanfrancisco.utn.edu.ar

05 | NOV.
06



150
ING

la Argentina celebra
su ingeniería
1870-2020



**Octavo Congreso Nacional de Ingeniería
Informática/Sistemas de información
CONAIISI**

5 y 6 de noviembre de 2020

Universidad Tecnológica Nacional, Facultad
Regional San Francisco

Memoria de Trabajos

Claudia Verino, Juan Carlos Calloni, Gabriel Cerutti, Alfonsina E. Andreatta
(Compiladores)

San Francisco, Córdoba - Argentina, Marzo de 2021

Octavo Congreso Nacional de Ingeniería Informática/Sistemas de Información: CONAIISI.
5 y 6 de noviembre de 2020 / Claudia Verino, Juan Carlos Calloni, Gabriel Cerutti,
Alfonsina E. Andreatta; compilado por Claudia Verino , Juan Carlos Calloni, Gabriel Cerutti,
Alfonsina E. Andreatta. - 1a ed. -
Ciudad Autónoma de Buenos Aires: Universidad Tecnológica Nacional.
Facultad Regional San Francisco, 2021.
Libro digital, PDF

Archivo Digital: descarga
ISBN 978-950-42-0202-8

1. Sistemas de Información. 2. Ingeniería Informática. I. Verino, Caludia, Calloni, Juan
Carlos, Cerutti, Gabriel, Andreatta, Alfonsina E. comp.
CDD 004.07

Octavo Congreso Nacional de Ingeniería Informática/Sistemas de información
CONAIISI
5 y 6 de noviembre de 2020
Universidad Tecnológica Nacional, Facultad Regional San Francisco

Memorias de trabajo

Diseño de Tapa: María Laura Vaudagna

ISBN 978-950-42-0202-8



Auspiciantes:



Validación de la Calidad en Datos Abiertos con respecto a la detección de errores ortográficos utilizando la métrica del factor Syntactic Correctness

Roxana Martínez, Christian Parkinson
Centro de Altos Estudios en Tecnología Informática (CAETI).
Universidad Abierta Interamericana,
Ciudad Autónoma de Buenos Aires, Argentina
{Roxana.Martinez, Christian.Parkinson}@uai.edu.ar

Resumen

En este trabajo se muestran distintos principios de calidad de los datos abiertos propuestos por organizaciones reconocidas a nivel internacional en la temática de gobierno abierto. Se verifica la calidad del dato a través de la métrica para la detección de errores ortográficos, utilizando el concepto del factor Syntactic Correctness. Por lo que, se presenta una sección con los trabajos relacionados orientados a la calidad de los portales de datos abiertos, y de las diversas herramientas de detección de errores ortográficos existentes. El objetivo y aporte de esta investigación fue incorporar y adaptar la herramienta más apropiada (según análisis comparativo realizado y verificado en este mismo estudio) con la inclusión de nuevas características y funciones de detección, en un desarrollo propio de software, el cual realiza un control de la calidad de los datos abiertos desde el aspecto de la detección de errores ortográficos en los datos de cada uno de los campos del tipo cadena de texto que conforman al Dataset analizado. Se tuvo en cuenta distintos criterios de análisis, que llevó al diseño de algoritmos en el prototipo para una mejor detección de falencias, como así de cantidad de ocurrencias por error. Además, se detectan los campos completos, es decir se analiza campo por campo (lo que conforma un registro), con el fin de verificar que todos fueron cargados, en caso contrario, se contabiliza como campo faltante. Este resultado, permite que un usuario utilice la herramienta de validación, y logre una mejor concientización de los aspectos que implican la calidad en los datos abiertos.

Palabras Clave: Calidad de Datos Abiertos; Herramienta de Validación; Ortografía en datos abiertos; Gobierno Abierto.

1. Introducción

Como es de público conocimiento, la innovación tecnológica facilita la creación de diversos canales que permiten que los ciudadanos tengan un mayor grado de participación pública. Es por eso que el rol de las TIC, es decir de las tecnologías de la información y la comunicación, son fundamentales en el contexto de la transformación de un Gobierno Abierto e innovador. La innovación tecnológica es fundamental en el papel de los procesos de consultas de datos públicos, abiertos y colaborativos, ésta sigue creciendo y es un aliado fundamental en la era del desarrollo y colaboración para articular la participación ciudadana en el Estado Mundial.

En las siguientes secciones se tratan los distintos conceptos y sus alcances como ser: Gobierno Abierto, Transparencia, Datos Públicos y la importancia de éstos para con el ciudadano.

1.1. Gobierno abierto

Actualmente existe un nuevo paradigma en toda la gestión pública tanto a nivel nacional como internacional, que se encuentra basado en distintos valores de transparencia de participación y de colaboración entre los distintos actores que intervienen en el gobierno y los ciudadanos. Este concepto de una política de apertura hacia el ciudadano implica la adopción de distintos procesos y una mayor participación por parte de éstos, en las actividades estatales para lograr la búsqueda de soluciones que están relacionadas con problemas públicos.

Gobierno Abierto es un “proceso de transformación del Estado que promueve la transparencia, la participación ciudadana y la colaboración como principios de diseño para una gestión pública innovadora” [1]. El “Gobierno Abierto se basa en cuatro pilares fundamentales: transparencia, rendición de cuentas, participación ciudadana y colaboración e innovación pública y

ciudadana” [2]. El concepto de transparencia va relacionado con el concepto de rendición de cuentas esto quiere decir, que un organismo público brinda información sobre las políticas que implementa, sus actividades internas, planes de acción y sus datos. El acceso a la información pública, le permite al ciudadano, tener un mejor debate público con evidencia y así, poder crear nuevas oportunidades para generar valor agregado. Por otra parte, se menciona a la colaboración, este concepto hace hincapié para trabajar en forma conjunta con los distintos ciudadanos, con las organizaciones no gubernamentales, universidades y/o empresas varias, donde todos juntos buscan soluciones innovadoras a los diversos problemas públicos que se pueden presentar generando de esa manera, una mejor calidad de vida.

La implementación del gobierno abierto en las Administraciones Públicas ha fomentado el suministro de información pública mediante la utilización de las páginas web oficiales y la creación de los portales de datos abiertos (open data) y de transparencia [3].

En la Figura 1, se muestran los distintos pilares mencionados anteriormente como ser: la transparencia participación y colaboración. Además, se pueden observar distintos elementos que forman parte del entorno del Gobierno Abierto, siendo éstos, la comunicación al ciudadano, consultas de liberación con estos, co-diseño de las políticas públicas con los ciudadanos, la organización de sociedades, metodologías de trabajo, por ejemplo: trabajo en forma horizontal, el monitoreo de las diversas políticas públicas y, sobre todo, los datos involucrados en el Gobierno Abierto.



Figura 1. Se muestra el contexto de los conceptos principales de Gobierno Abierto [1].

1.2. Datos públicos versus datos abiertos

El Estado Nacional debe incorporar las técnicas, TIC y los conocimientos necesarios para gestionar dicha

cantidad de datos con un diseño tecnológico y por sobre todo enfocado en el ciudadano promedio. Es importante tener en cuenta el uso y reutilización responsable y sustentable, por parte de actores públicos y/o privados.

Una de las grandes diferencias entre el dato público y el dato abierto, es que “el dato público es todo aquel que puede ser accedido o consultado por cualquier persona u organismo de forma libre, aunque no necesariamente sean datos digitalizados” [4], básicamente debe tener un marco regulatorio. Por el otro lado, los “datos abiertos (open data) son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen” [5]. Estos datos deben estar en formatos digitales, con un modelo estándar abierto.

Por otro lado, es importante tener en cuenta el concepto del dato sensible, es decir, “existe un conjunto de datos denominados sensibles que deben ser protegidos en todo caso, salvo que hayan sido expresamente cedidos por su titular. Son datos sensibles aquellos que, de divulgarse de manera indebida, afectarían la esfera íntima del individuo o podrían ser motivo de discurso de odio o discriminación por origen racial o étnico, estado de salud, información genética, creencias religiosas, filosóficas o morales, afiliación sindical o partidaria, opiniones políticas y preferencias sexuales” [6].

1.3. Transparencia

Todo lo que es relativo al derecho de acceso a la información pública, la participación ciudadana y la transparencia de la gestión pública, se relaciona básicamente con algunos de los principios de acceso a la información pública [7]. Uno de estos principios se enfoca en la transparencia y máxima divulgación, esto es, que toda la información en custodia o bajo control debe ser accesible a todas las personas, básicamente atender las necesidades de la sociedad y al interés público. Parte de estos principios, también hace referencia a la apertura de datos, es decir, la información debe ser accesible en distintos formatos electrónicos que de alguna forma faciliten su procesamiento y de esta manera, permitan su reutilización o su redistribución por parte de terceros.

2. Aspectos de calidad para datos abiertos

Hoy por hoy existen muchos conjuntos de datos que son brindados en los portales de datos abiertos, tanto a nivel nacional como a nivel internacional, por lo que un punto interesante a tener en cuenta, es que estos conjuntos de datos muchas veces, no poseen datos de calidad, es decir, no están en un formato correcto, existen errores ortográficos, o bien se encuentran campos incompletos. En otro de los trabajos realizados por los autores, se indica la

importancia de estos aspectos, ya que “los datos abiertos brindan una gran ventaja en la administración pública, debido a que permiten un mayor acceso a la información y un mejor conocimiento para la toma de decisiones políticas, estimulando el desarrollo de servicios innovadores y colaborativos. Los gobiernos actuales enfrentan grandes desafíos en base al contexto de este nuevo paradigma” [8].

En esta sección se tratan algunos aspectos que son considerados a la hora de determinar si un dataset, posee datos de calidad.

2.1. Aspectos de calidad según Open Data Charter

Según Open Data Charter, Carta Internacional de Datos Abiertos [9], su misión es hacer que los datos estén abiertos y disponibles gratuitamente para todas las personas y al mismo tiempo, esta organización tiene como misión proteger los derechos de las personas y las comunidades. Su visión general es que los gobiernos puedan recopilar, compartir y utilizar datos bien organizados y gestionados, para impulsar políticas y prácticas que permitan responder de manera eficaz, a los desafíos sociales económicos y ambientales más urgentes. Básicamente, trabaja en colaboración con más de 100 gobiernos y organizaciones que se esfuerzan para abrir datos basados en un conjunto de principios compartidos.

Esta organización, sostiene que “la apertura de datos gubernamentales puede fomentar la construcción de sociedades más interconectadas que satisfagan mejor las necesidades de los ciudadanos y permitan que la innovación, la justicia, la transparencia y la prosperidad prosperen, al tiempo que se garantiza la participación cívica en las decisiones públicas y la rendición de cuentas de los gobiernos” [10].

Los seis principios del Open Data Charter [9], serán la base para el acceso a los datos y para la divulgación y uso de los datos, son:

1. Abrir por defecto: Quiere decir que los datos están abiertos a todos los ciudadanos. Los gobiernos deben justificar los datos que se mantienen cerrados, por ejemplo, por razones de seguridad o protección de datos.

2. Oportuno y completo: Los datos abiertos sólo son valiosos si siguen siendo relevantes. Por otro lado, la información se publica de forma rápida y completa, lo que es fundamental para su éxito. En la medida de lo posible, los gobiernos deberían proporcionar datos en su forma original y sin modificaciones.

3. Accesible y utilizable: Garantizar que los datos sean legibles por máquina y fáciles de encontrar. Los portales son una forma de lograrlo. Pero también es importante pensar en la experiencia del usuario de quienes acceden a los datos, incluidos los formatos de archivo en los que se proporciona la información.

4. Comparable e interoperable: Cuantos más conjuntos de datos de calidad tenga acceso y más fácil sea para ellos hablar entre ellos, más valor potencial podrá obtener de ellos. Los estándares de datos cumplen un rol fundamental.

5. Mejorar la gobernanza y la participación ciudadana: Los datos abiertos tienen la capacidad de permitir que los ciudadanos (y otros en el gobierno) tengan una mejor idea de lo que están haciendo los funcionarios y los políticos. Esta transparencia puede mejorar los servicios públicos y ayudar a que los gobiernos rindan cuentas.

6. Para el desarrollo y la innovación inclusivos: los datos abiertos pueden ayudar a impulsar el desarrollo económico inclusivo. Por ejemplo, un mayor acceso a los datos puede hacer que la agricultura sea más eficiente o puede usarse para abordar el cambio climático.

2.2. Aspectos de Calidad según Open Government Data

La Open Government Data, Datos de Gobierno Abierto [11], tiene como misión brindar principios para que los datos gubernamentales, se identifiquen como datos abiertos. Los datos gubernamentales se considerarán abiertos si se hacen públicos de una manera que cumpla con los principios siguientes:

1. Completa: Todos los datos públicos están disponibles. Los datos públicos son datos que no están sujetos a limitaciones válidas de privacidad, seguridad o privilegios. Los ciudadanos tienen derecho a conocer y reutilizar el contenido de los datos que contienen los gobiernos.

2. Primaria: Los datos se recopilan en la fuente, con el mayor nivel posible de granularidad, no en forma agregada o modificada.

3. Oportuna: Los datos están disponibles tan rápido como sea necesario para preservar el valor de los datos.

4. Accesible: Los datos están disponibles para la más amplia gama de usuarios para la más amplia gama de propósitos.

5. Procesable a máquina: Los datos están razonablemente estructurados para permitir un procesamiento automatizado.

6. No discriminatorio: Los datos están disponibles para cualquier persona, sin necesidad de registrarse.

7. No propietario: Los datos están disponibles en un formato sobre el que ninguna entidad tiene control exclusivo.

8. Sin licencia: Los datos no están sujetos a ningún derecho de autor, patente, marca registrada o regulación de secreto comercial. Se pueden permitir restricciones razonables de privacidad, seguridad y privilegios.

Con respecto a los datos gubernamentales, debido a que son financiados con dinero público de los ciudadanos, la

información contenida en estos datos es pública y debe estar a disposición de cualquier ciudadano. Por lo que, estos deben estar disponibles en un formato abierto, que respete estándares técnicos y permita su reutilización y difusión.

Este trabajo trata el principio número 1, “*Completa*”. del Open Government Data, Datos de Gobierno Abierto [11], y el principio número 2, “*Oportuno y Completo*” del Open Data Charter [9], con el fin de analizar cada campo de un conjunto de datos desde su validación ortográfica.

3. Trabajos relacionados

En esta sección se analizan trabajos relacionados con la temática de calidad de datos abiertos y herramientas de detección ortográfica.

3.1. Calidad en los portales de datos abiertos

Existen trabajos [12], [13], [14], [15] [16] que se enfocan en establecer criterios y clasificaciones de niveles de calidad de los datos abiertos. Algunos escenarios de análisis se orientan a partir de la reutilización de datos abiertos y públicos. Por otra parte, algunas técnicas de medición de calidad, se basan en el concepto de la disponibilidad de los datos en los portales abiertos, con el fin de propiciar un nivel de disposición adecuado para los consumidores de estos. Además, se definen mecanismos para evaluar la madurez de un portal, a través de métricas para medir la calidad, como, por ejemplo: la trazabilidad, la completitud y la conformidad. Otros trabajos [17], orientan la evaluación de la calidad en el análisis del modelo de cinco estrellas de Berners-Lee y otros factores que ayudan a evaluar la reutilización. Además, proponen que se puede analizar la relación entre la demanda de los datos abiertos y la dispersión de distribuciones y conjuntos de datos disponibles sobre una determinada temática, por lo que, es interesante estudiar el resultado de si concentrar la información en pocos conjuntos de datos más completos y manejables, podría ayudar a mejorar la eficiencia en la publicación de éstos.

Otros trabajos [18] proponen técnicas para resolver los problemas con los sitios de gobierno de datos y para ello, desarrollan framework o técnicas que dan servicios de datos basados en la web que se estructurarán semánticamente. Este concepto en la utilización de framework, puede contemplar aspectos de: a) Proveedor de datos: Que los proveedores alojen sus datos de manera estándar para que, en el futuro, otros puedan usar / reutilizarlos de manera más semántica. El proveedor de datos puede trabajar con los formatos de conjuntos de datos estructurados como: CSV, XML, JSON, etc.; luego, que los datos estén disponibles utilizando un esquema de metadatos o un vocabulario. Para mejorar la calidad del software de los portales que realizan un tratamiento de

datos abiertos geoespaciales, existen trabajos [19], que proponen enfoques en el desarrollo y aplicación de un vocabulario base transversal que apoya la descripción de temas geoespaciales. El método de anotación resultante está integrado en una nueva plataforma de colaboración de acceso abierto que admite la difusión de datos públicos. Para ello, hacen mención a la utilización de ArcGIS Hub [20] que es una nueva plataforma de acceso abierto que admite y organiza el compromiso cívico y la colaboración directa entre los gobiernos y sus ciudadanos. ArcGIS Hub posee nuevas funcionalidades para compartir datos abiertos, catálogos de metadatos configurables, integración con registros de metadatos regionales y nacionales.

3.2. Herramientas de detección de errores ortográficos

En este apartado se describen las APIs (Interfaces de Programación de Aplicación) más populares que permiten agregar la funcionalidad de detección de errores gramaticales y ortográficos a los desarrollos.

A continuación, se muestra el trabajo realizado con una investigación de las distintas herramientas de detección de errores ortográficos para luego, en secciones posteriores, lograr implementarlas, compararlas, analizarlas y determinar cuál es la más apropiada para llevar a cabo una medición de los campos de texto que poseen errores ortográficos en los conjuntos de datos de los portales abiertos.

Si bien todas las herramientas tienen el mismo objetivo, se toma en cuenta la facilidad de integración a los proyectos, la flexibilidad de los diccionarios para ingresar nuevas palabras, la ubicación de los diccionarios (local o remoto), y los lenguajes de programación que las soportan.

En primer lugar, se describe a Hunspell [21], es uno de los componentes de detección de errores y gramaticales más populares, utilizada por aplicaciones como Google Chrome, LibreOffice, OpenOffice.org, Mozilla Firefox 3 & Thunderbird, como también en paquetes de software como macOS, InDesign, memoQ, Opera y SDL Trados. Fue Desarrollado en C++ bajo GPL/LGPL/MPL, y las principales funciones que se destacan son: la capacidad de brindar soporte para procesar caracteres Unicode, o con morfología compleja, realizar sugerencias utilizando algoritmos de similitud n-gram [22], datos de pronunciación basados en reglas y diccionarios, realizar análisis morfológico junto con sus derivadas. En cuanto a los idiomas que procesa incluye los diccionarios de MySpell (Viejo corrector ortográfico de Apache Open Office), como a su vez los de LibreOffice y extensiones, y Mozilla Add-Ons. Por último, provee interfaces para dispositivos Android (AndroidHunspellService), com además para Perl, .Net, Python, Ruby (1,2,3), Delphi, Java, etc.

En segundo lugar, se destaca Aspell.Net [23], que es un componente que provee un motor de revisión de escritura, es una API desarrollada en visual C#, basando su funcionalidad en la librería GNU Aspell mediante un puerto WIN32 API. GNU Aspell es una librería desarrollada en C++ que permite detectar los errores y realizar las sugerencias sobre posibles palabras que pueden reemplazar a la palabra que fue mal escrita. Otro elemento por destacar es que puede trabajar con documentos escritos en UTF8 sin la necesidad de utilizar un diccionario especial. Por último, soporta varios lenguajes, dado que la especificación y diccionario de cada lenguaje se encuentra en archivos independientes utilizados por el motor.

En tercer lugar, se puede mencionar a Proofreading API [24] de la empresa “LanguageTool GmbH”, se trata de un componente comercial para la validación gramática y detección de errores ortográficos. Internamente utiliza como motor de ortográfico a Hunspell, y las aplicaciones para utilizar los servicios, deben incluir una referencia a un servicio HTTP API, utilizando el formato de JSON, los usos de estos elementos estandarizados independizan a la herramienta del lenguaje de programación subyacente. Una característica interesante de este API es la validación de fechas y detección de género (en determinados lenguajes). El tiempo promedio de procesamiento es de 300 milisegundos cada 1000 caracteres. Para el uso de este API se contrata un paquete de comprobaciones diarias. Una limitación es que ofrece solamente 15 lenguajes, y debe ser utilizado en entornos conectados a la Internet.

Un servicio de similares prestaciones que se puede presentar es google-api-spelling-java [25], este componente desarrollado para Java incluye revisión, corrección y sugerencias que son procesadas en el servidor de Google. Todo el funcionamiento y responsabilidad del componente se encuentra encapsulado en una clase SpellChecker, su función principal es transformar los contenidos en a XML y traducir las respuestas del servidor en POJOs (Viejos objetos de java). El intercambio de información se realiza a través de la serialización basada en XML por medio de la estructura JAXB. Yendo a la línea de Microsoft, existe una serie de componentes desarrollados para trabajar con los lenguajes de programación suscritos al Framework .NET, que se describen a continuación.

Spellcheck es un [26] es un componente que forma parte de Framework .Net, si bien ya viene integrado dentro del IDE, para poder utilizarlo se debe agregar la referencia a la librería “PresentationFramework.dll” proporcionando funcionalidad de revisión ortográfica en tiempo real sobre los controles de ingreso de datos (Casillas de texto) provistas por el Framework.Net para proyectos de datos en WPF, como contraprestación permite únicamente 4 diccionarios: inglés, francés, español y alemán.

Otros componentes a mencionar, desarrollados para la integración en proyectos del Framework .Net, son NHunspell [27], NHunspellTextBoxExtender [28] y NetSpell [29], ambos heredan las clases de la librería mencionada anteriormente Hunspell, utiliza las de corrección y sugerencias de su base, en todos los casos son de fácil integración a los proyectos. En el caso de NHunspellTextBoxExtender y NetSpell añaden funciones a los controles del Framework, por ejemplo, el subrayado a las casillas de texto en las palabras con errores ortográficos.

NetSpell provee una mayor cantidad de clases para la integración con una mayor cantidad funciones para la configuración del diccionario, como contrapartida, maneja un conjunto de diccionarios exclusivos para el uso de la herramienta.

En la Tabla 1, se comparan todas las herramientas previamente expuestas con sus principales características.

Tabla 1. Comparativa de Herramientas de detección ortográficas

Herram.	Dicc. dist.	Open Source	Comercial	Agregar Palabras	Tecno.
Hunspell	Si	Si	No	Si	C++
Aspell.Net	No	No	No	Si	C#
Proofreading API	No	No	Si	no	C#
google-api-spelling-java	No	No	Si	No	Java
Spellcheck	No	No	No	Si	C#
NHunspell	Si	No	No	Si	C#
NHunspell TextBoxExtender	Si	No	No	Si	C#
NetSpell	No	No	No	Si	C#

Las características de la Tabla 1, son utilizadas para llevar a cabo el proceso de selección de la herramienta más apropiada para validar datos de los datasets en ese aspecto, por lo que se tiene en cuenta si: utiliza archivos de diccionario de distintos proveedores, es open source, es comercial, tiene la posibilidad de agregar palabras y su correspondiente tecnología.

4. Prototipo desarrollado y adaptado para la validación de herramientas de detección de errores ortográficos

Con el objetivo de seleccionar la herramienta más adecuada para la detección de errores en los conjuntos de datos abiertos, se desarrolló un prototipo propio para evaluar el tiempo de respuesta, sugerencias y detecciones de los principales componentes y optar por la mejor.

4.1. Proceso de Medición de errores ortográficos en textos

A continuación, se muestran las capturas de pantalla del prototipo desarrollado, en las que se visualizan los resultados arrojados para la comprobación de las herramientas puestas a prueba. En cada uno de los testeos, se verificó tiempos de respuestas promedio por palabra y tiempo de respuesta total. La frase utilizada fue “la bida es vella a pesar de los herrores”.

En la Figura 2, se visualizan las pruebas para NHunspell con sus respectivos resultados, en la Figura 3 para la herramienta NetSpell, y finalmente en la Figura 4 para Hunspell Textbox Extender.

Cabe aclarar, que se incluyeron únicamente las herramientas integrables a la plantilla de Windows Forms de Visual Studio, dado que el prototipo de la presente propuesta trabaja con esas características.

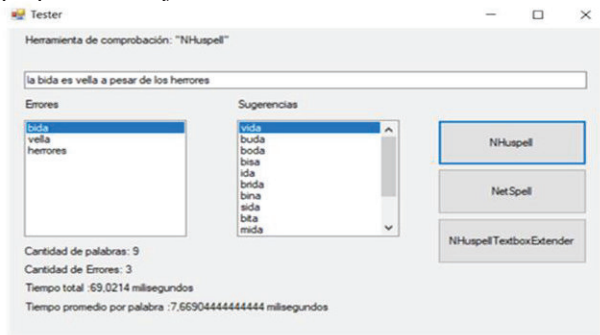


Figura 2. Comprobación de la herramienta NHunspell.

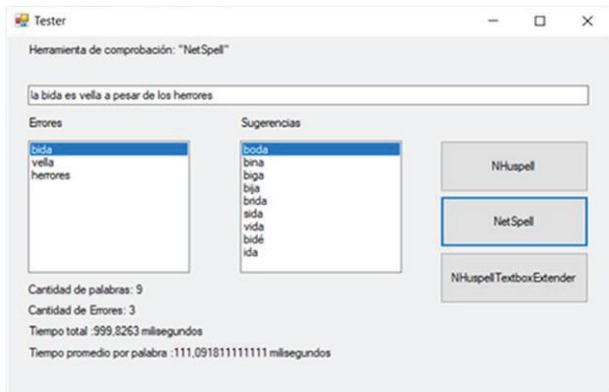


Figura 3. Comprobación de la herramienta NetSpell.

Como se puede observar en este diseño de prototipo, se presenta una casilla de texto para ingresar la frase a ser verificada por la herramienta seleccionada. Una vez que se realiza clic en el botón de alguna de las tres herramientas propuestas, se visualiza la lista de errores detectados y en una segunda lista, las sugerencias de corrección de la palabra seleccionada.

A modo de un breve resumen orientativo, se presenta la cantidad de palabras y errores, el tiempo total y el promedio por palabra expresados en milisegundos.

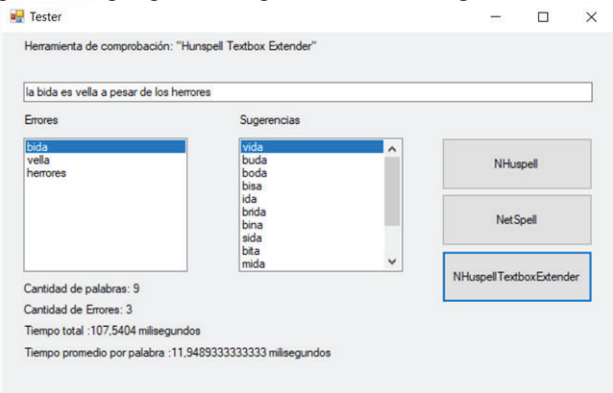


Figura 4. Comprobación de la herramienta Hunspell Textbox Extender.

En base a los resultados obtenidos, se comparan los resultados del proceso de la validación.

4.2. Validación comparativa

Los resultados obtenidos se detallan en la Tabla 2, allí se muestran los aspectos comparativos entre las herramientas, teniendo en cuenta la cantidad de palabras, el tiempo total de respuesta del texto analizado (expresado en milisegundos), el tiempo promedio por palabra (expresado en milisegundos), la cantidad de errores detectados y por ultimo, si la herramienta propone sugerencias de corrección para los errores que fueron identificados en la frase.

Tabla 2. Comparativa de tiempos de respuestas y cantidades de las herramientas validadas

Herram.	Cant Pbra	Tiemp. Rta Tot.	Tiemp. Promed Pbra	Cant Err detec	Suge renci as
NHunspell	9	69.0214 ms	7.66904 ms	3	SI
NetSpell	9	999.8263 ms	111.0918 ms	3	SI
NHunspell TextBoxExtender	9	107.5404 ms	11.94893 ms	3	SI

4.3. Resultados analizados

En base a la Tabla 2, se realiza un gráfico comparativo con los tiempos de respuestas totales, es decir, el tiempo que tarda en procesar la frase completa cargada en el prototipo.

En la Figura 5, se presenta la comparativa de tiempos, donde se observa que la herramienta NHunspell, tiene el menor tiempo de respuesta, 69,0214 milisegundos, en comparación con las otras herramientas para realizar la detección de errores ortográficos.

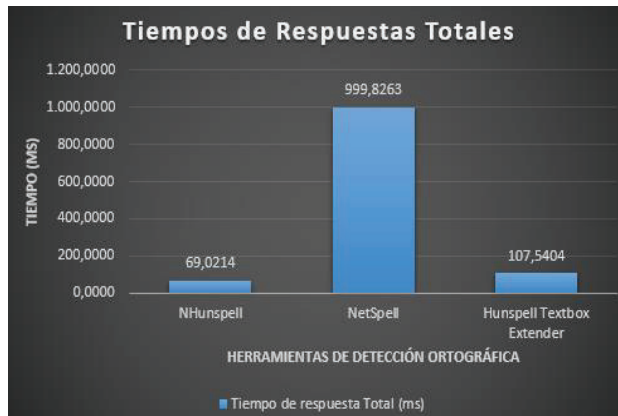


Figura 5. Comparativa de herramientas con sus tiempos de respuesta totales.

Por otra parte, en la Figura 6, se presenta la comparativa de tiempos promedio por palabras, donde se observa que la herramienta NHunspell, sigue encabezando el menor tiempo de respuesta, 7,6690 milisegundos en este caso, en comparación con las otras herramientas.



Figura 6. Comparativa de herramientas con sus tiempos promedios por palabras.

Como resultado, la herramienta NHunspell, es la que presenta mejores tiempos de performance y calidad en detección de errores ortográficos, por lo que se procede a utilizar la misma en la propuesta de un diseño de prototipo

más amplio, que implica el estudio de calidad de datos abiertos de portales de datos abiertos gubernamentales.

5. Propuesta: Herramienta de validación de calidad en datos abiertos para datasets

En base a la selección de una herramienta de detección de errores ortográficos, se propone el diseño de una herramienta propia para comprobar a través de ésta, la validación de calidad de los datos abiertos que se encuentran en un archivo de conjunto de datos, para este caso CSV, que son provenientes de portales de datos abiertos de organismos gubernamentales.

La propuesta de esta herramienta surge a partir de los errores ortográficos encontrados en algunos de los dataset que fueron publicados en diversos portales abiertos, a raíz de esta problemática es que se decide desarrollar un prototipo de herramienta de validación de calidad de datos abiertos basado en distintos aspectos, los cuales determinan si un dataset presenta calidad en los datos ofrecidos a los ciudadanos.

5.1. Alcance del prototipo

Este trabajo se enfoca en la validación del principio número 2, “Oportuno y Completo” del Open Data Charter [9], y en lo que respecta a los principios del Open Government Data, Datos de Gobierno Abierto [11], nos centraremos en el principio número 1, “Completa”, con el fin de analizar cada campo de un conjunto de datos desde su detección de errores ortográficos.

Esta herramienta realiza un análisis y verificación de la calidad de los datos abiertos desde el aspecto de la detección de errores ortográficos en los datos de cada uno de los campos del tipo cadena de texto que conforman al Dataset analizado. Además, se detectan los campos completos en el Dataset, es decir se analiza campo por campo (lo que conforma un registro), con el fin de verificar que realmente todos los campos fueron cargados con datos, en caso contrario, es decir si no existen datos registrados, se contabiliza como campo faltante. Este resultado, permite que un usuario utilice la herramienta de validación, y logre una mejor concientización de los aspectos que implican la calidad en los datos abiertos.

Cabe destacar que se brinda un análisis con la identificación de falencias ortográficas, pero no implica su modificación o alteración del archivo de datos abiertos original. Por el momento, la herramienta tiene tratamiento de archivos de datos abiertos del tipo CSV.

A continuación, se describen los aspectos de mediciones de calidad para esta propuesta, la metodología de validación implementada para la medición de errores ortográficos en textos de conjuntos de datos de portales abiertos.

5.2. Metodología implementada para la medición de calidad de datos abiertos

Una métrica debe ser válida y estar relacionada con el concepto a medir que es de interés. “Las métricas de calidad son un instrumento usado para medir cierto factor de calidad, por ejemplo, el porcentaje de datos que no tienen errores de sintaxis es una métrica del factor correctitud sintáctica” [30], o también conocida como Syntactic Correctness [31]. Esta métrica, pretende medir si los valores corresponden a valores válidos del dominio, sin importar si son reales o no. Para verificar la correctitud sintáctica se compara con un diccionario que representa el dominio del campo de texto. Como se relevó en la sección 4.3, la herramienta más oportuna para este prototipo es NHunspell.

Este estudio utiliza la dimensión de exactitud de calidad, la cual es analizada para el factor de Correctitud Sintáctica, a través de una medición booleana para saber si existen o no errores de sintaxis en los datasets de portales abiertos. La granularidad de la validación es del tipo “palabra”, por lo que se obtiene un gráfico comparativo del total de palabras del conjunto de datos y la cantidad de palabras con errores encontrados para visualizar que porcentaje fue el afectado en cada uno.

5.3. Funcionamiento del prototipo

El diseño del prototipo desarrollado tiene dos opciones: en un primer botón permite seleccionar un archivo de un Dataset, en este caso permite manipular archivos con extensión CSV. Otro de los botones es el de “Analizar Dataset de Datos Abiertos” al realizar clic, se lleva a cabo un proceso interno en el que, a través de diversos algoritmos, se realiza una comprobación de cada uno de los campos del conjunto de datos, al finalizar el proceso, se muestra el resultado que implica la visualización de una grilla con los errores ortográficos detectados en los campos de cadenas de texto, junto con información del análisis escaneado.

5.4. Aspectos técnicos utilizados en el desarrollo

A continuación, se analizan los distintos aspectos técnicos que fueron utilizados en el desarrollo de este prototipo como ser el IDE utilizado, el motor de base de datos y la elección de la herramienta más amigable para llevar a cabo la detección de distintos errores ortográficos y su incorporación en el código fuente programado.

La herramienta de validación fue realizada con el entorno de desarrollo integrado (IDE), Visual Studio Community 2019 [32], que es una herramienta completa para programar, depurar, probar e implementar soluciones en cualquier plataforma. Otro de los motivos por el cual fue seleccionada, es que posee un entorno de

programación amigable, y, además, su versión Community es gratuita. Por otra parte, existen varios foros a nivel soporte técnico de la plataforma y material de respaldo.

En cuanto al lenguaje de programación se utilizó C # que es el lenguaje de programación orientado a objetos, con ASP .NET que es el marco proporcionado por Microsoft para el desarrollo web. Puntualmente se utilizó C # junto con ASP .NET para la codificación del lado del servidor, y HTML, CSS, JS para el lado del cliente.

La base de datos utilizada fue Sql Server Express 2019 en su versión gratuita.

5.5. Incorporación de la detección de errores

Los errores ortográficos suelen ser difíciles de identificar en los conjuntos de datos, por lo que se consideró embeber un identificador de análisis en los datos de texto para detectar errores. En base al estudio del relevamiento de las distintas herramientas que existen para la detección de errores ortográficos realizada en la sección anterior, se opta la herramienta más propicia para llevar a cabo la validación de palabras con respecto a los errores ortográficos.

Cada herramienta tiene en común el criterio de búsqueda, para realizar las correcciones ortográficas se realizan dos técnicas conocidas como análisis de n-gramas y la búsqueda en diccionario. La primera consiste en descomponer las palabras en diferentes secuencias de elementos, es decir, fonemas, letras, sílabas, palabras, dependiendo de la aplicación que se le vaya a realizar. A partir de esta composición, la técnica que se utiliza para realizar la corrección en un lenguaje consiste en determinar la frecuencia de ocurrencia de estos n-gramas, cuando se ingresa una palabra, se descompone y se compara con los n-gramas, en caso de no encontrar algún n-grama correspondiente, o bien, algunos con muy baja probabilidad de ocurrencia, se asume haber detectado un error tentativo.

Para complementar la técnica de n-gramas se agrega la búsqueda en diccionarios, que consiste en localizar palabras de texto en un listado, esto es, entre más palabras se encuentren cargadas en el diccionario, mayor será el peso del archivo y su tratamiento.

La búsqueda de palabras de forma secuencial resulta ser un mecanismo bastante ineficiente, razón por la cual se utilizan 2 mecanismos:

1. **Árboles binarios:** se calcula que da 1.000.000 de palabras y se realizan 20 iteraciones para encontrar un elemento.

2. **Tablas de Hash:** consiste en una matriz asociativa que, dependiendo de la eficacia del algoritmo para realizar la búsqueda, el número de iteraciones podría llegar a 1.

Partiendo de la premisa que una verbo tiene varias formas de conjugarse, afijos, también los sustantivos tienen número y género, tener un diccionario con cada uno de los vocablos conlleva a tener un diccionario de gran peso, por lo cual, se divide en dos archivos, el primero con extensión “.dic” almacena todas formas básicas de las palabras, el segundo archivo almacena las formas y afijos y tiene su extensión es “.aff”.

Para el desarrollo de la presente propuesta se eligió el componente “NHunspell”, la elección se vio basada en la sección 4.3 de este trabajo. El prototipo está desarrollado con el lenguaje de programación, Visual C#, por lo cual su integración se limita únicamente al agregado de una referencia. Por otro lado, las cadenas de texto a evaluar, no están en ningún control (casilla de texto), razón por la cual, no requiere funcionalidades adicionales de subrayado, resultando este componente más liviano.

Otro de los puntos a destacar, es que, al ser una librería gratuita, permite reducir la dependencia de contratos con entidades externas para el libre uso de esta propuesta. Por último, el hecho de que los diccionarios y la librería realizan la comprobación de forma local (on site), mejora la performance entre la consulta de búsqueda y la respuesta, sin depender del estado de red, es decir, se prioriza todos los elementos necesarios para resolver la tarea de manera local (on site), reduciendo las dependencias de terceros.

6. Resultados de la Validación

Se realizó un estudio comparativo de 4 conjuntos de datos extraídos de portales abiertos gubernamentales.

Los conjuntos de datos son: a) Dataset de Traslados COVID-19 [33], tipo de archivo CSV, listado con información de los traslados realizados por el Gobierno de la Ciudad en relación a COVID-19; b) Dataset de organizaciones [34], tipo de archivo CSV, el cual tiene los datos abiertos de organizaciones vinculadas con actividades científico tecnológicas o bien en las que se desarrollan actividades de I+D y/o innovación; c) Dataset del operativo DetectAR [35], tipo de archivo CSV, este conjunto de datos, tiene la ubicación de los operativos DetectAR para testeo de COVID-19 en los barrios de la Ciudad de Buenos Aires; d) Dataset de Padrón de Operadores Orgánicos Certificados [36], tipo de archivo CSV, listado de operadores (productores primarios, elaboradores y comercializadores) orgánicos certificados.

En las Figuras 7, se muestra un caso de ejemplo con los resultados arrojados por el prototipo desarrollado para el dataset [33]. Como se puede observar, se muestra un resumen del proceso realizado con: el tiempo de la duración del proceso, cantidad de: registros analizados, columnas, campos sin datos, palabras con errores detectados y, el total de palabras encontradas en el dataset.

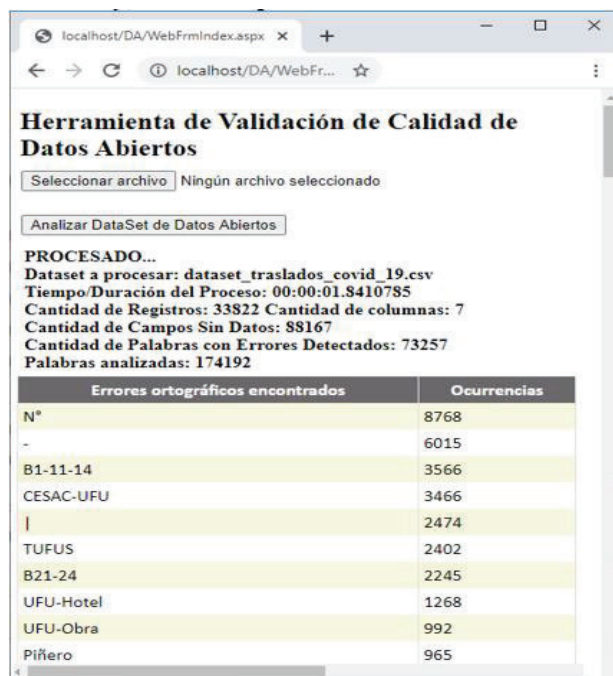


Figura 7. Pantalla con el resumen del resultado final del análisis del dataset Traslados COVID-19 [33].

Según los resultados obtenidos, se realiza una comparativa entre las palabras con y sin errores ortográficos de los datasets, con el fin de identificar el porcentaje para la métrica del factor correctitud sintáctica. Esta métrica, pretende medir si los valores corresponden a valores válidos del dominio, sin importar si son valores reales o no, como se explicó en la sección 5.2.

En la Figura 11, se muestran los resultados comparativos entre los 4 datasets y sus correspondientes cantidades de palabras analizadas y detectadas con errores ortográficos para cada uno de éstos.

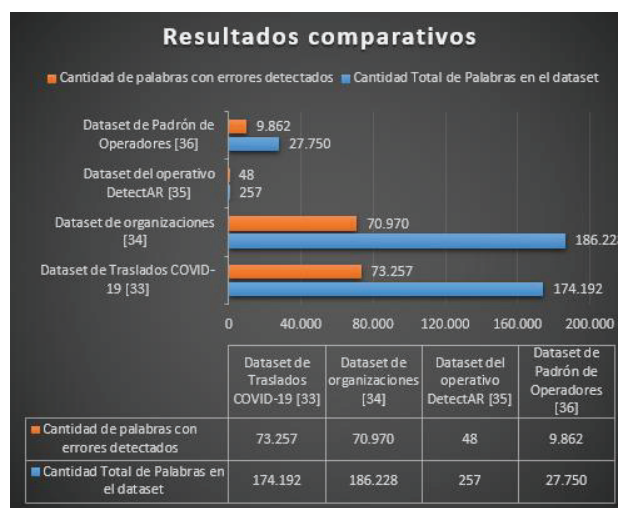


Figura 11. Gráfico comparativo de los resultados arrojados por el prototipo de validación de datos abiertos.

En la Figura 12, se muestran los porcentajes discriminados por dataset. Estos porcentajes, representan las palabras con errores, dentro del total de palabras del dataset.

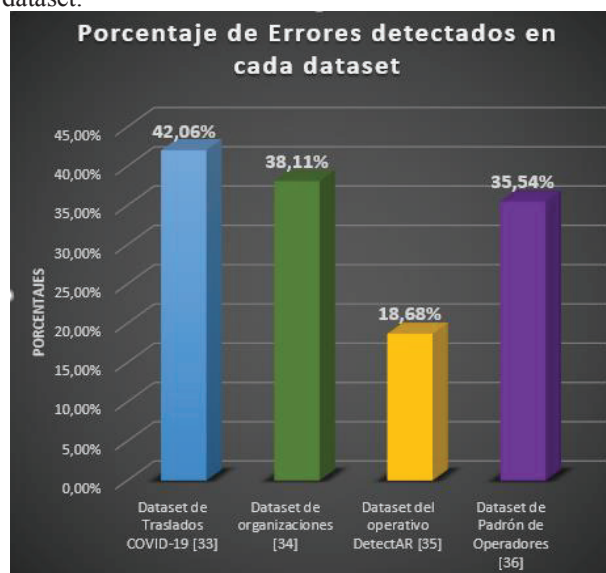


Figura 12. Porcentaje de errores detectados en cada dataset.

Como se muestra en la Figura 12, para estos casos, existe menos de un 50% de errores encontrados en cada uno de los conjuntos de datos (datasets), pero los valores de los porcentajes arrojados siguen siendo elevados, por lo que, es necesario prestar especial atención, ya que es importante mantener la calidad del dato completo, para una mejor identificación y procesamiento de la información que es reutilizada de los portales de datos abiertos gubernamentales. Este tipo de errores de precisión (ortográficos) son aquellos que afectan la identificación [37], por ejemplo, si en un registro bibliográfico se escribe el apellido Zábato, y un usuario quiere solicitar Sábado, obtendrá una respuesta negativa; si en cambio se registra Güir_aldes no se podrá recuperar directamente.

7. Líneas Futuras

Este trabajo forma parte de una tesis doctoral la cual se enfoca en aspectos y métricas de calidad de datos abiertos, que se analizan en los distintos dataset de los portales de gobierno abierto. Por lo que, como parte de los trabajos futuros se pretende continuar con el relevamiento y estudio de más aspectos que son esenciales para la medición de calidad en los conjuntos de datos abiertos. Otro de los puntos para continuar trabajando, es la ampliación de la herramienta de validación propuesta, incorporando más criterios y principios mencionados en este trabajo por las distintas organizaciones mundiales que inculcan la cultura de gobierno abierto.

8. Conclusiones

En estos tiempos, los datos son un recurso indispensable para cualquier actividad de gestión pública, por lo que es necesario mantenerlos actualizados, claros y completos. Si bien se puede determinar que cada vez son más los gobiernos que se embarcan en este nuevo paradigma de concepto de datos abiertos, todavía falta un largo camino por recorrer. Distintos trabajos relacionados en esta temática, se están enfocando en la calidad de los datos abiertos de los portales de entidades gubernamentales, pero aún sigue siendo necesario, reforzar y comprender la importancia del dato, es decir, si un dato tiene calidad puede ser mejor utilizado, manipulado y distribuido a los ciudadanos para varios fines, permitiendo a su vez, una mayor participación ciudadana y transparencia en organismos estatales. Este trabajo se enfoca en difundir este paradigma, y así colaborar con el cumplimiento y la divulgación de los distintos criterios y/o principios que proponen entidades mundialmente reconocidas.

Con la verificación del prototipo desarrollado, se pudo comprobar que es posible incorporar métricas de calidad en la medición de estos datos abiertos, y que estas mediciones finales realizados en datasets, permiten dejar en evidencia que existen varios portales de datos abiertos, que ofrecen conjuntos de datos con errores ortográficos, cómo así campos incompletos. El gran problema que conlleva que los datos se presenten con errores ortográficos, supondrá una pérdida de información y, por consiguiente, una pérdida de la objetividad de lo que se está analizando como resultado.

9. Referencias Bibliográficas

- [1] Ministerio de modernización - presidencia de la Nación. Kit de Gobierno Abierto – Transparencia. Disponible en: https://www.argentina.gob.ar/sites/default/files/1._kit_gobierno_abierto_-_transparencia.pdf al 18 de agosto 2020.
- [2] Arroyo Chacón, J. (2017). La Innovación Abierta Como Pilar Del Gobierno Abierto (Open Innovation as a Pillar of Open Government). Revista Enfoques, 15(27), 13-41.
- [3] Curto Rodríguez, R. (2017). Datos abiertos y rendición de cuentas en las comunidades autónomas españolas. Niveles de divulgación y determinantes.
- [4] Secretaría de Modernización - Presidencia de la Nación. Paquete de Apertura de Datos de la República Argentina. Disponible en: <https://datosgobar.github.io/paquete-apertura-datos/guia-subnacionales/#1-que-son-los-datos-abiertos> al 18 de agosto 2020.
- [5] Aleixandre-Benavent, R., Sapena, A. F., & Peset, F. (2019). Compartir los recursos útiles para la investigación: datos abiertos (open data). Educación Médica.

- [6] Atenas, J., Belbis, J., Pane, J., Casanueva, J. M., & Bonina, C. (2019). Estado abierto a través de Datos abiertos.
- [7] Argentina.gob.ar. Principios de acceso a la información pública. Disponible en: <https://www.argentina.gob.ar/aaip/accesoalainformacion/principios> al 18 de agosto 2020.
- [8] Martínez, R., Rodríguez, R., Vera, P., & Parkinson, C. (2020). Análisis del procedimiento para la solicitud de información pública y tiempos de respuesta. Caso de Estudio: Ciudad Autónoma de Buenos Aires (Argentina). 49º Jornadas Argentinas de Informática, SIE 2020 - Simposio de Informática en el Estado.
- [9] ODC. Open Data Charter - La carta Internacional de Datos Abiertos. Disponible en: <https://opendatacharter.net/> al 18 de agosto 2020.
- [10] Open Data Charter - La carta Internacional de Datos Abiertos. Principios. Disponible en: <https://opendatacharter.net/principles/#timely-and-comprehensive> al 18 de agosto 2020.
- [11] Open Government Data. The 8 Principles of Open Government Data (OpenGovData.org). Disponible en: <https://opengovdata.org/> al 18 de agosto 2020.
- [12] Oviedo Blanco, E. (2016). Modelo de madurez para portales de datos abiertos e incorporación a la norma técnica nacional de Costa Rica. Repositorio Institucional de la Universidad de Alicante, Tesis Doctoral, Disponible en: <http://rua.ua.es/dspace/handle/10045/54316>, al 18 de agosto 2020.
- [13] Beltrán, L., Estefán, N., & Mahecha Moyano, J. F. (2017). Prototipo de software para la evaluación de la calidad de datos abiertos. Tesis de grado, Repositorio Institucional de la Universidad Católica de Colombia, Disponible en: <https://repository.ucatolica.edu.co/jspui/bitstream/10983/14642/1/Data%20Quality%20in%20OpenData.pdf>, al 18 de agosto 2020.
- [14] Ibanez Gonzalez, L., Millard, I., Glaser, H., & Simperl, E. (2019). An assessment of adoption and quality of linked data in European open government data.
- [15] Rodríguez Rojas, L. A. (2017). Metamodelo para integración de datos abiertos aplicado a inteligencia de negocios. Tesis de Doctoral, Repositorio Institucional de la Universidad de Oviedo, Disponible en: <http://digibuo.uniovi.es/dspace/handle/10651/44552>, al 18 de agosto 2020.
- [16] Arizo, I. (2016). Métricas basadas en datos. Tesis de Maestría en Gestión de la Información, Universitat Politècnica de València.
- [17] Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Herederó, C. (2018). Indicadores de calidad de datos abiertos: el caso del portal de datos abiertos de Barcelona. El profesional de la información (EPI), 27(2), 375-382.
- [18] Nayek, J. K. (2018). Evaluation of Open Data Government Sites: A Comparative Study. Library Philosophy & Practice. (e-journal). 1781. Disponible en: <https://digitalcommons.unl.edu/libphilprac/1781>, al 18 de agosto 2020.
- [19] Lafia, S., Turner, A., & Kuhn, W. (2018). Improving discovery of open civic data. In 10th International Conference on Geographic Information Science, (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- [20] ArcGIS Hub. ArcGIS Hub. Disponible en: <http://hub.arcgis.com/> al 01 de junio 2020.
- [21] HunsPELL. Disponible en: <http://hunsPELL.github.io/> al 19 de agosto 2020.
- [22] Kondrak, G. (2005). N-gram similarity and distance. In International symposium on string processing and information retrieval, (págs. 115-126). Springer, Berlin, Heidelberg.
- [23] aspell-net. Disponible en: <http://aspell-net.sourceforge.net/> al 19 de agosto 2020.
- [24] LanguageTool GmbH. Proofreading API. Disponible en: <https://languagetoolplus.com/> al 19 de agosto 2020.
- [25] Google, google-api-spelling-java. Disponible en: <https://code.google.com/archive/p/google-api-spelling-java/> al 19 de agosto 2020.
- [26] Google, google-api-spelling-java. Disponible en: <https://docs.microsoft.com/es-es/dotnet/api/system.windows.controls.spellcheck?view=netcore-3.1> al 19 de agosto 2020.
- [27] NHunsPELL. Disponible en: <https://www.codeproject.com/Articles/33658/NHunsPELL-HunsPELL-for-the-NET-platform> al 19 de agosto 2020.
- [28] NetSPELL. Disponible en: <https://sourceforge.net/projects/netspell/> al 19 de agosto 2020.
- [29] NetSPELL. Disponible en: <https://sourceforge.net/projects/netspell/> al 19 de agosto 2020.
- [30] Martirena, E. (2008). Medición de la calidad de datos: un enfoque parametrizable. Tesis de posgrado: Magister en Informática, Universidad de la República (Uruguay). Facultad de Ingeniería. Instituto de Computación – PEDECIBA.
- [31] Batini, C., & Scannapieco, M. (2016). Data and information quality. Cham, Switzerland: Springer International Publishing. Google Scholar, 43.
- [32] Microsoft. IDE de Visual Studio 2019. Disponible en: <https://visualstudio.microsoft.com/es/vs/> al 17 de agosto 2020.
- [33] Buenos Aires Data. Dataset Traslados COVID-19. Disponible en: <https://data.buenosaires.gov.ar/dataset/traslados-covid-19/archivo/48970b40-5d73-4995-a69a-54124a1ed22c> al 18 de agosto 2020.
- [34] Datos Argentina. Organizaciones. Disponible en: https://datos.gob.ar/dataset/mincyt-instituciones-sistema-nacional-ciencia-tecnologia-argentino/archivo/mincyt_6b4d32b3-b28a-47cb-bdc9-321c36cd0195 al 18 de agosto 2020.
- [35] Buenos Aires Data. Dataset Operativo DetectAR. Disponible en:

<https://data.buenosaires.gob.ar/dataset/operativo-detectar> al 18 de agosto 2020.

[36] Datos Argentina. Padrón de Operadores Orgánicos Certificados. Disponible en:

https://datos.gob.ar/dataset/agroindustria-organicos---padron-operadores-certificados/archivo/agroindustria_6ab26b25-c323-46b9-8cef-2de43c439e0f al 18 de agosto 2020.

[37] Martínez Tamayo, A. M., Mangiaterra, N. E., Pisarello, R. Z., Stubbs, E. A., & Cap, A. S. (2002). Errores de precisión y de consistencia en la catalogación descriptiva. Información, cultura y sociedad, 7.