



Calidad de Datos en Linked Data

Proyecto Final de la Carrera Ingeniería en Informática

Alumnos

Alderete, Facundo

de la Puerta Echeverría, María

Romarión, Germán Rodrigo

Tutor

Vaisman, Alejandro Ariel

Índice

Capítulo 1	5
Introducción	5
1.1 Historia de la Web	5
1.2 Tecnologías de la Web Semántica	6
1.2.1 RDF	6
Capítulo 2	9
Linked Data	9
2.1 ¿Qué es Linked Data?	9
2.2 Los cuatro principios	9
2.3 La tecnología detrás de Linked Data	10
Capítulo 3	13
Calidad de datos	13
3.1 ¿Qué es la calidad de datos?	13
3.2 Calidad de datos en la Web Semántica y Linked Data	13
3.2.1 Calidad de la fuente de datos	14
3.2.2 Calidad del documento	15
3.3 Fórmulas para cada métrica	15
3.3.1 Métricas para la calidad de la fuente de datos	16
3.3.2 Métricas para la calidad del documento	17
3.4 Fórmulas generales	19
3.4.1 Fórmula para la calidad del endpoint	19
3.4.2 Fórmula para la calidad del documento	19
3.4.3 Fórmula para la calidad global	20
Capítulo 4	20
Metodología de evaluación de calidad de datos	20

4.1	Introducción	20
4.2	Data Quality Assessment	21
4.2.1	Registración	21
4.2.2	Ingreso a la aplicación y roles	22
4.2.3	Edición de perfil	25
4.2.4	Búsqueda de documentos	26
4.2.5	Evaluación de un documento	27
4.2.6	Mis evaluaciones	35
4.2.7	Reportes	36
4.2.8	Usuario administrador	41
	Referencias	44

Capítulo 1

Introducción

1.1 Historia de la Web

En 1989, Tim Berners-Lee, un científico inglés, inventó la World Wide Web (WWW) haciendo realidad su visión de un sistema de información global hipervinculada. Hacia fines de 1990 presentó las primeras versiones de Hypertext Transfer Protocol (HTTP), Hypertext Markup Language (HTML), el primer Web browser y HTML editor y el primer Web server software.

“El proyecto World Wide Web (WWW) tiene como objetivo permitir que se creen enlaces a cualquier información en cualquier lugar... El proyecto WWW se inició para permitir que los científicos compartan datos, noticias y documentación. Estamos muy interesados en la difusión de la web en otras áreas y en la posibilidad de tener servidores para otros datos. Colaboradores bienvenidos!” (Tim Berners-Lee, alt. Hypertext, 1991).

Desde ese entonces, la Web permanece como un espacio de información distribuida que provee una gran cantidad de conocimiento en muchos formatos. El intercambio de información en la Web es posible únicamente aceptando un estándar de formato para los datos. La característica clave de HTML es que los enlaces se denotan usando *dictated mark-up* que le permite a las máquinas entenderlos sin asistencia humana.

Los buscadores clásicos se tornaron insuficientes para manejar la cantidad de contenido en Web que crece a gran velocidad. Para solucionar esto, las aplicaciones Web empezaron a usar paradigmas para organizar y buscar información. Un ejemplo popular es el *tagging*, que se basa en palabras clave (“*tags*”) que los usuarios asignan a los recursos. Este enfoque fue popular para estructurar contenido que no es en su mayoría texto, como imágenes, videos, etcétera. Otro ejemplo que se usó para mejorar las búsquedas es el ranking de usuarios.

Resumiendo, hay una clara tendencia hacia agregar más estructura a los recursos Web. Sin embargo, muchos de los ejemplos mencionados no siguen un estándar común, lo que

dificulta la explotación de esta información de otra forma que no sea utilizando los típicos buscadores de internet.

1.2 Tecnologías de la Web Semántica

La Web Semántica fue concebida como una extensión de la World Wide Web que permite a las computadoras buscar, combinar y procesar contenido Web de forma inteligente basado en el significado que tiene este contenido para los humanos. Al no tener una computadora inteligencia artificial al nivel humano, esto sólo puede lograrse si el significado deseado de los recursos Web es explícitamente especificado en un formato procesable por computadoras. No basta con guardar los datos en una sintaxis procesable por una máquina (de hecho, toda página HTML lo es), sino que se requiere que estos datos estén dotados de semánticas formales que especifiquen claramente qué conclusiones deben ser sacadas de la información recolectada. La Web Semántica se basa en estándares que promueven tipos de datos comunes y protocolos de intercambio en la Web.

El propósito final de la Web Semántica es permitir a las máquinas acceder a más información que requeriría el uso de mucha atención y tiempo humano.

1.2.1 RDF

Para el intercambio de datos en la Web se utiliza el *Resource Description Framework* (RDF), un lenguaje formal para describir información estructurada. El objetivo de RDF es permitir a las aplicaciones intercambiar datos en la Web preservando su significado original. A diferencia de HTML y XML, la principal intención no es mostrar documentos de forma correcta, sino permitir el procesamiento y la recombinación de la información contenida en ellos.



Figura 1.1: Un grafo simple RDF que describe la relación entre un libro y su editor

Un **documento RDF** describe un grafo dirigido donde tanto los nodos como la arista cuentan con identificadores para diferenciarlos. La información en XML se codifica en estructuras de tipo árbol. Los árboles son adecuados para organizar información en documentos electrónicos, donde con frecuencia nos encontramos con estructuras estrictamente

jerárquicas. Además, la información en árboles puede ser recolectada directamente y procesada de forma eficiente. Nos preguntamos entonces por qué RDF utiliza grafos. La respuesta gira en torno a la concepción de RDF; no fue ideado con el propósito de estructurar documentos, sino para describir relaciones entre objetos de interés, llamados comúnmente “recursos” en RDF.

El grafo de la Figura 1.1 podría usarse para describir que el libro fue publicado por CRC Press, si interpretamos las etiquetas. Esta relación entre libro y editor es información que no representa una jerarquía en un sentido obvio. Otro motivo es que RDF fue pensado como un lenguaje de descripción para datos en la WWW y otras redes electrónicas. La información en estos ambientes se suele almacenar y manejar de manera descentralizada y por esto es muy fácil combinar datos RDF de muchas fuentes. Por ejemplo, el grafo de la Figura 1.1 podría ser combinado con otros grafos de <http://semantic-web-book.org>, lo que llevaría a un grafo más grande que podría proveer nueva información interesante.

RDF usa URIs como nombres o identificadores, para distinguir recursos entre sí. Las URIs son una generalización de las URLs. Cada URL es a su vez una URI válida y puede ser utilizada como identificador en un documento RDF.

Como se ve en la Figura 1.1 tanto los nodos como las aristas en grafos RDF se etiquetan con URIs para distinguirlas de otros recursos. Esta regla tiene dos posibles excepciones:

- RDF permite codificar valores de datos que no son URIs
- RDF tiene nodos que no llevan nombre, llamados *blank nodes*

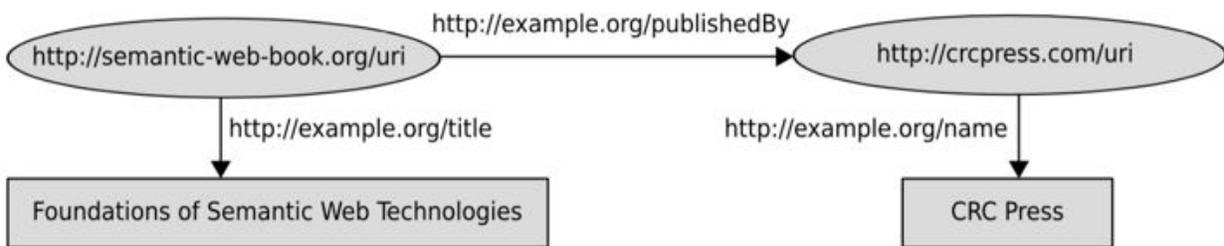


Figura 1.2: Un grafo RDF con literales para describir valores de datos.

Los valores de datos en RDF son representados por los llamados “literales”. Estos son nombres reservados para recursos RDF de un tipo de dato concreto. El valor de cada literal es

descrito por una secuencia de caracteres. En representaciones de grafos RDF las cajas rectangulares son usadas para distinguir literales de URIs, como se observa en la Figura 1.2. Los literales no pueden ser origen de una arista, esto quiere decir que nunca podemos hacer una declaración directa sobre literales. Tampoco se permite etiquetar aristas con literales.

Vistos de esta forma, los literales son meramente un arreglo de caracteres. Se requieren otros tipos de datos para aplicaciones prácticas, como para denotar números o puntos en el tiempo, por ejemplo. Los tipos de datos son de gran utilidad a la hora de interpretar un valor dado (para realizar un ordenamiento de valores de datos, por ejemplo). Es por esto que RDF permite que los literales lleven asociado un cierto tipo de dato. Cada tipo de dato es identificado unívocamente por una URI que puede ser escogida de forma arbitraria. Sin embargo, en la práctica es útil referirse a URIs ampliamente conocidas y soportadas por muchas herramientas de *software*. Por este motivo, RDF sugiere el uso de XML Schema.

```
<rdf:Description rdf:about="http://www.w3.org/TR/rdf-primer">
  <ex:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    RDF Primer
  </ex:title>
  <ex:publicationDate
    rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
    2004-02-10
  </ex:publicationDate>
</rdf:Description>
```

Figura 1.3: Representación de un documento RDF en formato RDF/XML utilizando atributos XML.

La Figura 1.3 muestra cómo puede agregarse información sobre los tipos de datos a un grafo RDF. Se ve que se especifica el título y la fecha de publicación y el tipo de dato con su URI correspondiente entre paréntesis angulares con el atributo XML "rdf:datatype:".

Los grafos proveen una forma meramente ilustrativa de representar RDF. Se puede representar un grafo como un conjunto de aristas donde cada una tiene un punto de partida, una etiqueta y un punto de llegada (sujeto, predicado y objeto respectivamente). A este conjunto se lo llama Terna RDF.

Capítulo 2

Linked Data

2.1 ¿Qué es Linked Data?

La Web Semántica es una Web de datos de cualquier tipo que pueda concebirse, tales como fechas, títulos, números, etcétera. La colección de tecnologías de la Web Semántica provee un entorno en el que se puede consultar esos datos y sacar conclusiones sobre los mismos.

Los conjuntos de datos publicados cubren una diversa gama de dominios como la geografía, las ciencias y los medios de comunicación, por nombrar algunos. Sin embargo, para lograr que la Web de datos sea una realidad, es importante tener una gran cantidad de datos en la Web disponibles en un formato estándar, alcanzable y manejable por herramientas de la Web Semántica. Más aún, no sólo se necesita tener acceso a los datos sino que además deben estar disponibles las relaciones entre los mismos para poder crear una Web de datos y no sólo una mera colección de conjuntos de datos. La colección de conjuntos de datos interrelacionados en la Web se denomina Linked Data.

Para lograr crear Linked Data, las tecnologías deben estar disponibles en un formato común que permita realizar consultas.

2.2 Los cuatro principios

En 2006, Tim Berners-Lee esbozó los cuatro principios de Linked Data, con el objetivo de señalar reglas que hagan que la Web crezca. Estos son:

1. Utilizar URIs para identificar los recursos publicados en la Web
2. Aprovechar el HTTP de la URI para que la gente pueda localizar y consultar estos recursos.

3. Proporcionar información útil acerca del recurso cuando la URI haya sido desreferenciada.
4. Incluir enlaces a otras URIs relacionadas con los datos contenidos en el recurso, de forma que se potencie el descubrimiento de información en la Web.

Si bien los puntos anteriormente mencionados son llamados reglas o principios, representan además expectativas de comportamiento, ya que el incumplimiento de ellos no es destructivo, sino que se pierde una oportunidad de interconectar datos correctamente. Esto a su vez limita las formas en que los datos pueden ser reusados de maneras inesperadas. Es este tipo de reutilización de la información lo que agrega valor a la Web.

El primer principio, identificar recursos a través de URIs, es un concepto clave de la Web Semántica. Si no se utiliza el conjunto de símbolos URI universal, entonces no es considerado Web Semántica.

La segunda regla, utilizar el HTTP de la URI, también es un concepto ampliamente comprendido. La única desviación conocida con relación a este principio es la tendencia constante de las personas de inventar nuevos esquemas URI. Generalmente, esto se relaciona con no querer comprometerse con el Domain Name System (DNS) establecido.

El tercer principio, el de proveer información en la Web contra una URI, es seguido por la mayoría de las ontologías pero no así por la mayoría de los conjuntos de datos.

La cuarta y última regla, la de incluir enlaces a otros datos, es necesaria para conectar los datos que existen a una Web; una Web ilimitada en donde uno puede encontrar todo tipo de información, tal como se concibió con la Web hipervinculada.

En la Web hipervinculada, está mal visto no generar enlaces a datos externos relacionados. El valor de su propia información es en gran medida función de los enlaces que tiene, así como también el valor inherente de la información dentro de la página web.

2.3 La tecnología detrás de Linked Data

Linked Data se basa en dos tecnologías que son fundamentales en la Web: *Uniform Resource Identifiers* (URIs) y *HyperText Transfer Protocol* (HTTP). Mientras los *Uniform Resource Locators* (URLs) se popularizaron en la forma de direcciones para documentos y

otras entidades que pueden ser localizadas en la Web, las URIs proveen una forma más genérica de identificar cualquier entidad que existe en el mundo.

Cuando las entidades están definidas por URIs que utilizan el esquema *http://*, estas pueden ser buscadas desreferenciando la URI a través del protocolo HTTP. De esa forma, este protocolo provee un mecanismo simple y universal de obtener recursos que pueden ser serializados en forma de un flujo de bytes.

Las URIs y HTTP son complementados por RDF: mientras que HTML provee un medio para estructurar y enlazar documentos en la Web, RDF provee un modelo genérico y basado en grafos con el que se puede estructurar y enlazar datos que describan cosas en el mundo.

El modelo RDF codifica datos en la forma de ternas “sujeto, predicado y objeto”. El sujeto y el objeto son URIs, cada una identificando un recurso (o una URI) y un *string*. El predicado indica cómo un sujeto y un objeto están relacionados, y también es representado por una URI. Por ejemplo, una terna RDF puede afirmar cómo dos personas, Ana y María, ambas identificadas por una URI, están relacionadas por el hecho de que Ana conoce a María. De la misma forma, una terna RDF puede relacionar a Juan con un artículo *X* en cierta base de datos declarando que Juan es autor de *X*. Dos recursos enlazados de esta forma, pueden ser extraídos de distintas fuentes de datos en la Web, permitiendo que un dato en un conjunto de datos pueda ser enlazado con otro en un conjunto diferente, y de esta forma surge una Web de Datos.

```
Subject: http://dig.csail.mit.edu/data#DIG
Predicate: http://xmlns.com/foaf/0.1/member
Object: http://www.w3.org/People/Berners-Lee/card#i

Subject: http://data.linkedmdb.org/resource/film/77
Predicate: http://www.w3.org/2002/07/owl#sameAs
Object: http://dbpedia.org/resource/Pulp_Fiction_%28film%29
```

Figura 2.1: Ejemplos de enlaces RDF

Los enlaces RDF toman la forma de ternas RDF cuando el sujeto de la terna es una referencia URI en un conjunto de datos y el objeto de la terna es una referencia URI en otro. La Figura 2.1 muestra dos ejemplos de enlaces RDF. El primero declara que un recurso identificado por la URI *http://www.w3.org/People/Berners-Lee/card#i* es miembro de otro

recurso llamado <http://dig.csail.mit.edu/data#DIG>. Cuando la URI del sujeto es desreferenciada a través del protocolo HTTP, el servidor *dig.csail.mit.edu* contesta con una descripción RDF del recurso. Cuando la URI del objeto es desreferenciada, el servidor de W3 provee una descripción de Tim Berners-Lee. Si se desreferencia el predicado se obtiene una definición del tipo de enlace *member*, descrito en RDF y utilizando *RDF Vocabulary Definition Language* (RDFS). El enlace RDF del segundo ejemplo conecta la descripción de la película Pulp Fiction de la base de datos de Linked Movie con aquella en la base de datos *DBpedia*¹.

Los estándares *RDF Vocabulary Definition Language* (RDFS) y *Web Ontology Language* (OWL) proporcionan una base para la creación de vocabularios que pueden ser usados para describir entidades en el mundo, y cómo estas se relacionan.

Los vocabularios son colecciones de clases y propiedades. Estos son expresados en RDF, usando condiciones de RDFS y OWL, que proveen diferentes grados de expresividad para modelar dominios. Cualquier persona es libre de publicar vocabularios en la Web de Datos, que a su vez pueden ser conectados por ternas RDF que enlacen clases y propiedades en un vocabulario a aquellas en otro, definiendo mapeos entre vocabularios relacionados.

Usando URIs para identificar recursos, el protocolo HTTP como mecanismo de recuperación y el modelo RDF para representar descripciones de recursos, Linked Data se basa directamente en la arquitectura general de la Web. De esta forma, la Web de Datos puede ser vista como una capa adicional que está estrechamente entrelazada a la Web de Documentos y tiene muchas de las mismas propiedades:

- La Web de Datos es genérica y puede contener todo tipo de datos.
- Cualquiera puede publicar datos a la Web de Datos.
- Aquellos que publican datos no se ven limitados por la elección de vocabulario en el cual representar datos.
- Las entidades se conectan mediante enlaces RDF, creando un grafo global de datos que extiende a las fuentes de datos y permite el descubrimiento de nuevas fuentes.

¹ *DBpedia* es un proyecto cuyo objetivo es extraer contenido estructurado de la información creada a partir del proyecto Wikipedia. *DBpedia* permite a sus usuarios realizar consultas de forma semántica sobre relaciones y propiedades asociadas a recursos de Wikipedia.

Capítulo 3

Calidad de datos

3.1 ¿Qué es la calidad de datos?

Existen muchas definiciones de lo que implica que ciertos datos sean de calidad. Algunas de ellas son:

1. Grado de excelencia mostrada por los datos en relación a la representación del escenario real.
2. El estado de completitud, validez, consistencia, atemporalidad y exactitud que hace a los datos apropiados para un uso específico.
3. Los procesos y tecnologías involucradas para garantizar la conformidad de los valores de los datos a los valores del negocio y a criterios de aceptación.

La definición de calidad de datos de la ISO 9000 es el grado al que un conjunto de características de los datos cumple requerimientos.

3.2 Calidad de datos en la Web Semántica y *Linked Data*

El creciente tamaño y disponibilidad de datos publicados en la Web como *Linked Data* (LD) hace que su calidad sea un desafío clave en muchas aplicaciones. Los principios de calidad de datos son vistos como esenciales para asegurar que los datos son adecuados para su uso previsto en toma de decisiones, operaciones y planificación.

El valor real de los datos se observa cuando estos se utilizan, por ende, la calidad se relaciona directamente con la habilidad de satisfacer las necesidades continuas del usuario. Muchos de los problemas en calidad de datos se dan por interpretaciones equivocadas o problemas en la semántica de los datos. Incluso conjuntos de datos con problemas de calidad pueden ser útiles para ciertas aplicaciones, por ejemplo, la calidad de datos de *DBpedia* es perfectamente suficiente para enriquecer una búsqueda en la Web sobre hechos o sugerencias

sobre información general. Sin embargo, si quisiese desarrollar una aplicación para uso médico probablemente la calidad de datos de *DBpedia* sea insuficiente.

Asegurar la calidad es un desafío ya que los datos subyacentes provienen de un conjunto de fuentes de datos anónimas y en constante desarrollo. La calidad de datos es un área que ha sido investigada desde mucho antes de la aparición de LD, algunas cuestiones de calidad son únicas de LD y otras ya han sido investigadas previamente. A pesar de que la calidad en LD es un concepto esencial, existen pocos esfuerzos en marcha para estandarizar cómo la calidad de datos debe ser implementada. Tampoco existe un consenso sobre cómo deben ser definidas las dimensiones y métricas de calidad de datos.

Se han podido presentar algunos enfoques que definen terminologías comunes relacionadas a la calidad de datos, diferentes dimensiones con sus definiciones y métricas cualitativas y cuantitativas para esas dimensiones. La evaluación de la calidad de datos involucra la medición de dimensiones o criterios de calidad que son relevantes para el consumidor. Las dimensiones pueden ser consideradas como las características de un conjunto de datos. Una métrica, medida o indicador de calidad de datos es un procedimiento para medir una dimensión de calidad de datos. Estas métricas son heurísticas que son diseñadas para una situación específica de evaluación. Se puede computar un puntaje de evaluación de estos indicadores de calidad.

Evaluaremos la calidad desde dos puntos de vista:

- Calidad de la fuente de datos
- Calidad del documento

Cada una de estas categorías contiene distintos atributos sobre los cuales se puede medir la calidad. A continuación se describirán los atributos correspondientes a cada una.

3.2.1 Calidad de la fuente de datos

En esta perspectiva se considera la disponibilidad de los datos y la credibilidad de la fuente de datos.

Los atributos con los cuales medimos la calidad de la fuente de datos son:

- **Disponibilidad:** se define como el grado al que los datos (o una porción de ellos) están presentes, alcanzables y listos para su uso.
- **Licencias:** datos que carecen de una licencia explícita son una posible responsabilidad legal y no deja en claro a los consumidores cuales son sus condiciones de uso. Por lo tanto, es importante aquellos que publiquen datos hagan explícitos los términos bajo los cuales el conjunto de datos puede ser utilizado.

3.2.2 Calidad del documento

Consideraremos atributos de calidad que se relacionen directamente con el recurso o documento que se esté evaluando.

Los atributos con los cuales medimos la calidad del documento son:

- **Validez sintáctica:** la validez sintáctica es definida como el grado en que un documento RDF se ajusta a la especificación del formato de serialización. La serialización es el proceso de trasladar una estructura de datos u objeto a un formato que pueda ser almacenado y luego reconstruido. En esta métrica se consideran errores de sintaxis en RDF/XML, tipos de datos de literales mal formados y literales incompatibles con los valores posibles para el tipo de dato.
- **Precisión semántica:** la precisión semántica se define como el grado en que los valores de los datos correctamente representan hechos del mundo real.
- **Correctitud:** Este criterio determina si los enlaces a sitios Web externos contienen información relacionada con el recurso en cuestión.

3.3 Fórmulas para cada métrica

Luego de haber mencionado las métricas relevantes para medir la calidad en Linked Data plantearemos en esta sección las fórmulas que usaremos para cada una de ellas para obtener un valor numérico que represente la calidad.

3.3.1 Métricas para la calidad de la fuente de datos

Disponibilidad

Para poder medir el grado de disponibilidad guardamos las respuestas obtenidas cada vez que se realiza una consulta al *endpoint* a lo largo de la evaluación. De esta forma se plantea una fórmula que tiene en cuenta la cantidad de solicitudes exitosas y las totales:

$$e_0 = 1 - \frac{\text{solicitudes}_{erroneas}}{\text{solicitudes}_{totales}}$$

Donde S_e son solicitudes erróneas y S_t son solicitudes totales realizadas. Definimos una solicitud errónea cuando el pedido obtiene una respuesta por parte del servidor con código 5xx.

Licencias

Tal como mencionamos en la sección 3.2.1, es importante explicitar bajo qué licencia se publica un conjunto de datos para comprender sus condiciones de uso. La propiedad `dcterms:licence` debe ser utilizada para indicar bajo qué licencia se publica una fuente de datos y realizando una consulta mediante *SPARQL* podemos saber si existe una licencia para el conjunto de datos. La fórmula que planteamos para esta métrica es binaria y se basa en la existencia o no de esta propiedad:

$$e_1 = \begin{cases} 1 & \text{si existe licencia} \\ 0 & \text{si no existe licencia} \end{cases}$$

Métrica	Descripción	Fórmula	Tipo de evaluación
Disponibilidad	Verificar si el servidor responde a una <i>query</i> SPARQL.	$e_0 = 1 - \frac{\text{solicitudes}_{erroneas}}{\text{solicitudes}_{totales}}$	Automático
Licencias	Determinar si existe una licencia para el conjunto de datos	$e_1 = \begin{cases} 1 & \text{si existe licencia} \\ 0 & \text{si no existe licencia} \end{cases}$	Automático

Tabla 1: Métricas de calidad relacionadas a la fuente de datos

3.3.2 Métricas para la calidad del documento

Validez sintáctica

Para evaluar la validez sintáctica de un documento tomaremos dos tipos de errores. Los distintos tipos de errores sintácticos que tendremos en consideración:

- **Tipo de dato incorrectamente extraído (e_2):** tipo de dato de un literal que está incorrectamente mapeado. Ejemplo: foaf:name, غیبی@en (es incorrecto ya que el tipo de dato no se encuentra en inglés).
- **Valor del objeto extraído de forma incompleta (e_3):** parte de los datos no han podido ser extraídos de forma completa, debido a algún tipo de error durante el proceso. Ejemplo: dbpprop:dateOfBirth "3"^^.

Las métricas que usaremos para evaluar el nivel de calidad en cuanto a validez sintáctica de un documento, son las siguientes:

$$e_2 = \frac{datos_{incorrectos}}{p}$$

$$e_3 = \frac{valores_{incompletos}}{p}$$

Donde $datos_{incorrectos}$ y $objetos_{incompletos}$ representan la cantidad de errores de “tipo de dato incorrectamente extraído” y “valor del objeto extraído de forma incompleta” respectivamente, y p es la cantidad de propiedades presentes en el documento.

Precisión semántica

Diremos que un valor es semánticamente correcto cuando representa el estado correcto de un objeto. Por ejemplo, si observamos un recurso que trata sobre el atentado al *World Trade Center*, entonces la fecha de ocurrencia debería ser 11 de septiembre de 2001 y no otra diferente.

Planteamos como métrica la siguiente fórmula:

$$e_4 = \frac{\text{valores}_{\text{incorrectos}}}{p}$$

Donde $\text{valores}_{\text{incorrectos}}$ indica la cantidad de propiedades que contienen valores semánticamente incorrectos y p es la cantidad de propiedades presentes en el documento.

Correctitud

Tomaremos como base que un enlace incorrecto contiene información que no tiene un grado de relación obvio con el documento, por ejemplo, no menciona en su contenido al recurso en cuestión o datos relacionados con el mismo.

Para evaluar la calidad se define una fórmula de la siguiente manera:

$$e_5 = \frac{\text{enlaces}_{\text{incorrectos}}}{\text{enlaces}_{\text{presentes}}}$$

Donde $\text{enlaces}_{\text{incorrectos}}$ indica la cantidad de enlaces que contienen información irrelevante al documento y $\text{enlace}_{\text{presentes}}$ es la cantidad de propiedades que contienen enlaces presentes en el documento.

Métrica	Descripción	Fórmula	Tipo de evaluación
Validez sintáctica	Tipo de dato de un literal mal construido	$e_2 = \frac{\text{datos}_{\text{incorrectos}}}{p}$	Manual
	Valor del objeto extraído en forma incompleta	$e_3 = \frac{\text{valores}_{\text{incompletos}}}{p}$	
Precisión semántica	Cantidad de valores incorrectos asociados a cada propiedad de un documento sobre el total de valores en el documento	$e_4 = \frac{\text{valores}_{\text{incorrectos}}}{p}$	Manual
Correctitud	Determina si los enlaces a sitios Web externos contienen información relacionada con el documento en cuestión.	$e_5 = \frac{\text{enlaces}_{\text{incorrectos}}}{\text{enlaces}_{\text{presentes}}}$	Manual

Tabla 2: Métricas de calidad relacionadas a los documentos

3.4 Fórmulas generales

3.4.1 Fórmula para la calidad del endpoint

Para obtener un valor de calidad del endpoint planteamos una fórmula que toma en consideración tanto la disponibilidad como la presencia de licencias. El valor obtenido nos otorgará información sobre la calidad del endpoint sobre el que se esté realizando evaluaciones de documentos. La fórmula es la siguiente:

$$Q_{fuentededatos} = \frac{e_0 + e_1}{2}$$

3.4.2 Fórmula para la calidad del documento

Luego de haber obtenido el valor para cada métrica individual del documento definimos una fórmula que logra obtener un valor general para este. La fórmula plantea un promedio ponderado, es decir, tomando los valores de cada evaluación de error individuales estos se multiplican por pesos determinados. Luego se divide por la suma de los valores de evaluación:

$$Q_{documento} = 1 - (w_1e_2 + w_2e_3 + w_3e_4 + w_4e_5)$$

Donde w_i indica el peso asignado a cada tipo de error y e_i es la cantidad de veces que se halló el error en las propiedades del documento.

De esta fórmula se obtiene q que indica un valor para el nivel de calidad del documento. q tomará un valor entre 0 y 1 y según este tomaremos una escala de calificación de la siguiente forma:

$Q_{documento}$	Calificación
0.9 - 1	A
0.8 - 0.89	B
0.7 - 0.79	C
0.6 - 0.69	D

0 - 0.59	F
----------	---

Tabla 3: Calificación según valor de $Q_{documento}$

Dependiendo de la calificación del documento podremos sacar una conclusión sobre su calidad.

3.4.3 Fórmula para la calidad global

Llamaremos calidad global a aquella que contempla tanto a la calidad de los documentos como a la calidad del endpoint del cual provienen esos documentos. De esta forma se puede obtener una métrica general que sea relevante para el usuario. Planteamos la siguiente expresión:

$$Q_{global} = \frac{Q_{fuentededatos} + \frac{\sum_{i=1}^q Q_{documento_i}}{q}}{2}$$

Donde $Q_{fuentededatos}$ indica la calidad del endpoint dependiendo de su disponibilidad y presencia de licencias, $Q_{documento}$ es la fórmula que evalúa la calidad de los documentos y q es la cantidad de documentos evaluados. Mediante esta fórmula se propone entonces calcular un promedio de la calidad de los documentos evaluados.

Capítulo 4

Metodología de evaluación de calidad de datos

4.1 Introducción

Una metodología de evaluación de calidad de datos se define como el proceso de determinar si una porción de datos cumple con la información que el consumidor requiere para cierto uso específico.

El proceso que escogimos para evaluar la calidad de datos involucra tanto a la fuente de datos en sí como a los distintos documentos que la componen, tal como indicamos en las fórmulas presentadas en el capítulo anterior.

Para asistirnos con la tarea, hemos desarrollado un programa al que llamamos “*Data Quality Assessment*” que permite a usuarios visualizar diferentes documentos pertenecientes a un conjunto de datos y efectuar una evaluación en base a las métricas existentes en cada propiedad de un documento. Además, el programa da la posibilidad de correr chequeos de forma automática sobre la fuente de datos.

Los datos recolectados de la fuente de datos y de los documentos son luego procesados mediante las fórmulas planteadas en la sección previa para presentarle al usuario los resultados en una página de estadísticas y reportes.

A continuación explicaremos con mayor nivel de detalle los aspectos de nuestro programa y algunos casos de uso.

4.2 Data Quality Assessment

En esta sección se explica de forma detallada el uso de la aplicación **Data Quality Assessment**. Para acceder, ingrese a <https://itba-data-quality-assessment.herokuapp.com/>.

4.2.1 Registración

La aplicación da la opción al usuario de registrarse proveyendo un nombre de usuario y una contraseña.

Registrar Usuario

Nombre completo

Nombre de usuario

Contraseña

Confirmar contraseña

[Registrar](#)

[¿Ya estás registrado? ¡inicia sesión ahora!](#)

Imagen 4.1: Pantalla de registración de usuario.

Al registrarse luego podrá ingresar a la aplicación para realizar evaluaciones sobre fuentes de datos y sus documentos.

4.2.2 Ingreso a la aplicación y roles

Data Quality Assessment

Iniciar sesión

Nombre de usuario

Contraseña

Endpoint

[Ingresar](#)

[Ingresar como invitado](#)

[¿Aún no eres un usuario? ¡Regístrate ahora!](#)

Acerca del endpoint

SNOMED CT es el producto terminológico de salud clínica más completo y preciso del mundo, propiedad de SNOMED International y distribuido por todo el mundo. SNOMED CT se ha desarrollado en colaboración para garantizar que cumpla con las diversas necesidades y expectativas de los médicos de todo el mundo y ahora se acepta como un lenguaje global común para los términos de salud. Los pacientes y los profesionales de

Imagen 4.2: Pantalla de inicio de sesión. En rojo se marca la opción para “Ingresar como invitado”.

Definimos roles dentro de la aplicación que permiten al usuario ingresar creando un usuario y contraseña como fue explicado anteriormente o ingresar como invitado como puede

verse en la **Imagen 4.2**. Al utilizar esta opción no es necesario registrarse pero solo se poseerán permisos de lectura sobre evaluaciones hechas por otros usuarios.

Al crear un usuario para el ingreso se asigna el rol de “Evaluador” por defecto. Este rol permite realizar evaluaciones sobre documentos de forma ilimitada y ver las evaluaciones que hayan realizado otros usuarios.

Por último, existe el rol de “Administrador” que además de poseer permisos de lectura y escritura globales también posee permisos sobre los usuarios de la plataforma.

Rol	Permisos	Característica
Administrador	W/R global	Escritura y lectura de todas las evaluaciones realizadas en el programa y permisos sobre usuarios
Evaluador	W documentos y R global	Escritura y lectura de evaluaciones propias y de otros usuarios
Invitado	R global	Lectura de evaluaciones globales de los documentos

Tabla 4.1: Roles de usuarios

Al iniciar sesión también se da la posibilidad de elegir el *endpoint* que se quiere utilizar para realizar las evaluaciones. Según el *endpoint* que se elija se podrán evaluar los documentos pertenecientes al mismo. Para saber la naturaleza de los datos que se encuentran en cada *endpoint* se puede leer un resumen del mismo debajo.

The screenshot shows the 'Data Quality Assessment' login interface. It features a title 'Data Quality Assessment' at the top. Below it is a section titled 'Iniciar sesión' (Log in). This section contains three input fields: 'Nombre de usuario' (Username) with the value 'mariadelapuerta', 'Contraseña' (Password) with masked characters, and 'Endpoint' with a dropdown menu showing 'geo.linkeddata [dataset/municipios]'. There are two buttons: 'Ingresar' (Log in) in blue and 'Ingresar como invitado' (Log in as guest) in grey. A link '¿Aún no eres un usuario? ¡Regístrate ahora!' (Not a user? Register now!) is also present. Below the login section is a section titled 'Acercas del endpoint' (About the endpoint) with a paragraph of text explaining the project's goal of adding spatial dimensions to the Web of Data / Web Semantics.

Imagen 4.3: Pantalla de inicio de sesión y selección de *endpoint*

Para finalizar sesión o cambiar de usuario debe hacerse clic en el nombre del usuario ubicado en la parte derecha de la barra de navegación y seleccionar la opción “Cerrar sesión”.

The screenshot shows the 'Data Quality Assessment' dashboard. At the top left is the title 'Data Quality Assessment'. At the top right is a navigation bar with links for 'Reportes', 'Mis Evaluaciones', and the user name 'German Rodrigo Romarion'. In the top right corner, there is a dropdown menu with two options: 'Editar perfil' and 'Cerrar sesión', with the latter circled in red. The main content area has the title 'Data Quality Assessment' and a search bar with the placeholder text 'Buscar documento en Bioontology'. Below the search bar are two buttons: 'Buscar documento' (blue) and 'Obtener documento aleatorio' (grey). At the bottom of the page, there is a footer with the text 'Facundo Alderete - María de la Puerta - Germán Romarion'.

Imagen 4.3: Enlace a “Cerrar sesión”, marcado en rojo.

4.2.3 Edición de perfil

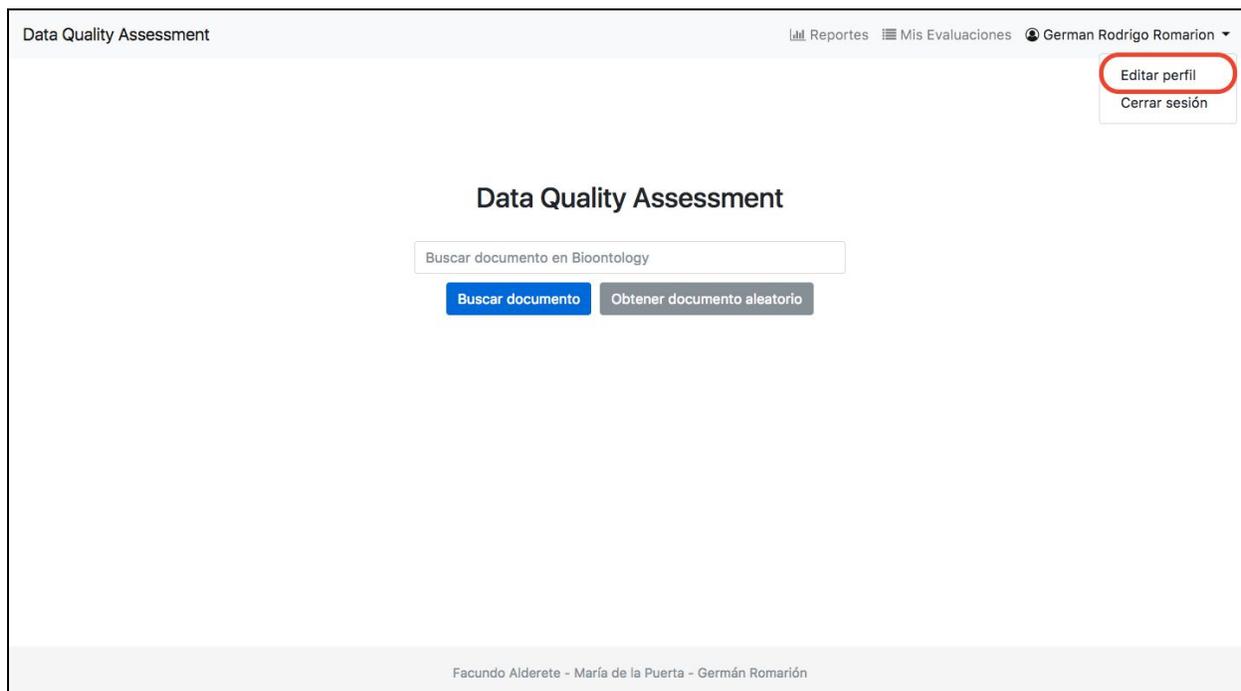


Imagen 4.4: Enlace a “Editar perfil”, marcado en rojo.

El usuario puede editar ciertos aspectos de su perfil haciendo clic en su nombre, ubicado en la parte derecha de la barra de navegación y seleccionando la opción “Editar perfil”, como se muestra en la **Imagen 4.4**.

En la pantalla de edición de perfil el usuario podrá editar su nombre completo, y actualizar su contraseña.

Data Quality Assessment

Reportes Mis Evaluaciones German Rodrigo Romarion

Editar perfil

Nombre completo

Password

Confirmar password

Guardar cambios Cancelar

Imagen 4.5: Pantalla de edición de perfil de usuario.

4.2.4 Búsqueda de documentos

Al ingresar a la aplicación se observa un buscador donde se puede ingresar el nombre del documento que se quiere evaluar.

Data Quality Assessment

Reportes Mis Evaluaciones María de la Puerta Echeverría

Data Quality Assessment

Buscar documento Obtener documento aleatorio

Imagen 4.6: Pantalla de búsqueda de documentos

Se cuenta también con una opción para obtener un documento aleatorio. En la **Imagen 4.6** se observa la pantalla de búsqueda.

Al realizar la búsqueda, se observa una página de resultados en forma de listado donde se lee el nombre del documento y su enlace dentro del *endpoint*.

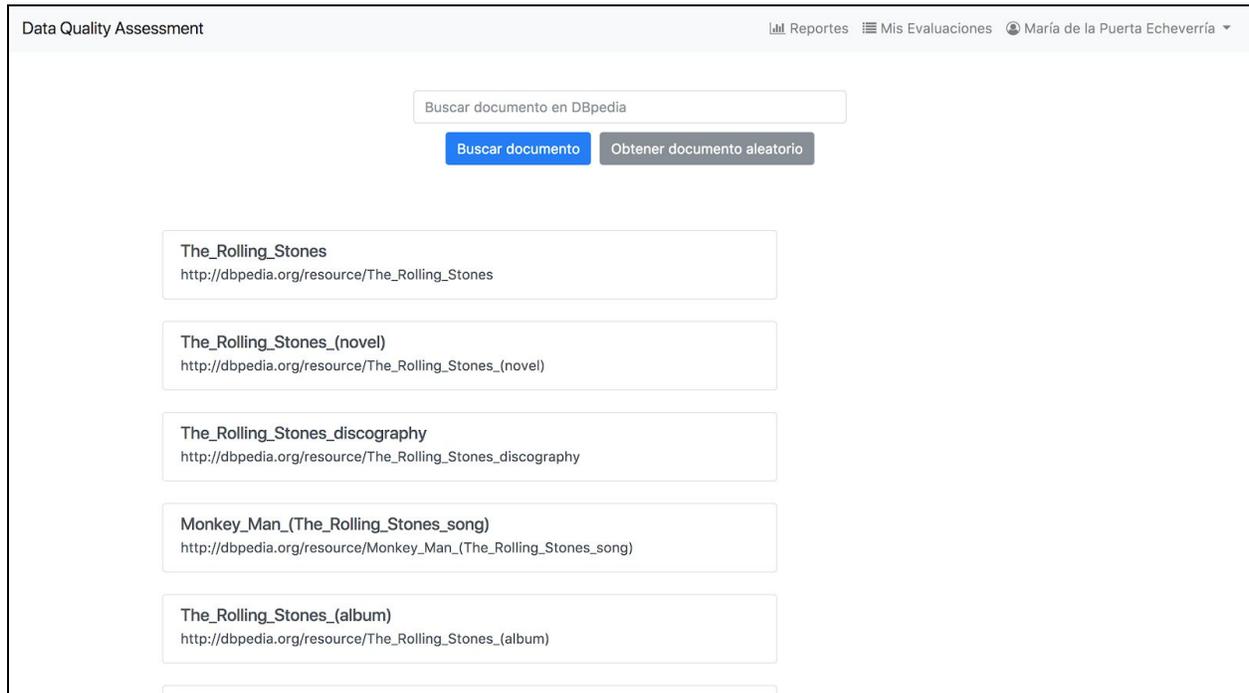


Imagen 4.7: Listado de resultados de búsqueda

Para acceder al documento se debe hacer clic en el recuadro del resultado que se quiere evaluar. En el caso de la **Imagen 4.7**, si queremos evaluar el documento “Rolling Stones”, haríamos clic en el primer recuadro.

4.2.5 Evaluación de un documento

Al elegir un documento para evaluar se observa una pantalla con el nombre del documento, un listado de predicados y objetos para ese documento y tres botones que ejecutan distintas acciones. El objetivo de esta sección es que el usuario lea los objetos y decida si estos contienen errores.

Data Quality Assessment		Reportes	Mis Evaluaciones	María de la Puerta Echeverría
http://dbpedia.org/ontology/activeYearsStartYear	"1962" (@type = http://www.w3.org/2001/XMLSchema#gYear)			
http://dbpedia.org/ontology/alias	"The Stones" (@lang = en)			
http://dbpedia.org/ontology/alias	"The Stones, Los Rolling" (@lang = en)			
http://dbpedia.org/ontology/associatedBand	http://dbpedia.org/resource/Billy_Preston			
http://dbpedia.org/ontology/associatedBand	http://dbpedia.org/resource/The_New_Barbarians_(band)			
http://dbpedia.org/ontology/associatedBand	http://dbpedia.org/resource/John_Mayall_&_the_Blue...			
http://dbpedia.org/ontology/associatedBand	http://dbpedia.org/resource/Bill_Wyman's_Rhythm_Ki...			
http://dbpedia.org/ontology/associatedBand	http://dbpedia.org/resource/Faces_(band)			
http://dbpedia.org/ontology/associatedMusicalArtist	http://dbpedia.org/resource/Billy_Preston			
http://dbpedia.org/ontology/associatedMusicalArtist	http://dbpedia.org/resource/The_New_Barbarians_(band)			
http://dbpedia.org/ontology/associatedMusicalArtist	http://dbpedia.org/resource/John_Mayall_&_the_Blue...			
http://dbpedia.org/ontology/associatedMusicalArtist	http://dbpedia.org/resource/Bill_Wyman's_Rhythm_Ki...			
http://dbpedia.org/ontology/associatedMusicalArtist	http://dbpedia.org/resource/Faces_(band)			
http://dbpedia.org/ontology/background	"group_or_band" (@lang =)			
http://dbpedia.org/ontology/bandMember	http://dbpedia.org/resource/Mick_Jagger			

Imagen 4.8: Listado de predicados y objetos. En rojo se marcan los íconos para reportar errores.

En caso que se encuentren errores, los mismos se pueden reportar haciendo clic en el ícono en forma de lápiz a la derecha del objeto, tal como se muestra en la **Imagen 4.8**. Cabe aclarar que si el usuario ingresó a la aplicación como invitado, entonces no le será posible realizar evaluaciones sobre un documento, y por lo tanto, no verá el ícono de lápiz junto a cada una de sus propiedades.

Al hacer clic en el ícono en forma de lápiz, la aplicación llevará al usuario a una pantalla donde podrá asignarle al objeto uno de los siguientes cuatro errores:

- Tipo de dato incorrectamente extraído
- Valor del objeto extraído de forma incompleta
- Objeto semánticamente incorrecto
- Enlace externo incorrecto

The screenshot shows a web interface for reporting data quality errors. The title is 'Ingresar nuevo error'. The form contains the following fields and content:

- Documento:** http://dbpedia.org/resource/Shawn_Barker
- Predicado:** <http://dbpedia.org/ontology/abstract>
- Objeto:** "Shaun Barker (born 19 September 1982) is an English footballer who plays as a defender for Burton Albion." (@lang = en)
- Tipo de error:** A dropdown menu with the selected option 'Tipo de dato incorrectamente extraído'.
- Descripcion del error:** Tipo de dato de un literal que está incorrectamente mapeado.
- Ejemplo:** "foaf:description, Questa è una descrizione di una risorsa in Italiano@es" es incorrecto ya que el tipo de dato no se encuentra en español
- Comentarios:** An empty text input field.
- Buttons:** 'Crear' (blue) and 'Volver' (grey).

Imagen 4.9: Pantalla para reportar errores.

Cada uno de los tipos de errores viene acompañado de una breve explicación sobre el mismo y un ejemplo ilustrativo que ayudará a que el usuario elija la mejor opción.

Es importante aclarar que la opción "Enlace externo incorrecto" aparecerá en el listado de errores únicamente si la propiedad en cuestión es un enlace externo o si contiene al menos un enlace externo.

Además, el usuario puede dejar un comentario respecto al error que acaba de seleccionar. De esta forma, puede dejar en claro el motivo por el cual marcó dicho error sobre la propiedad.

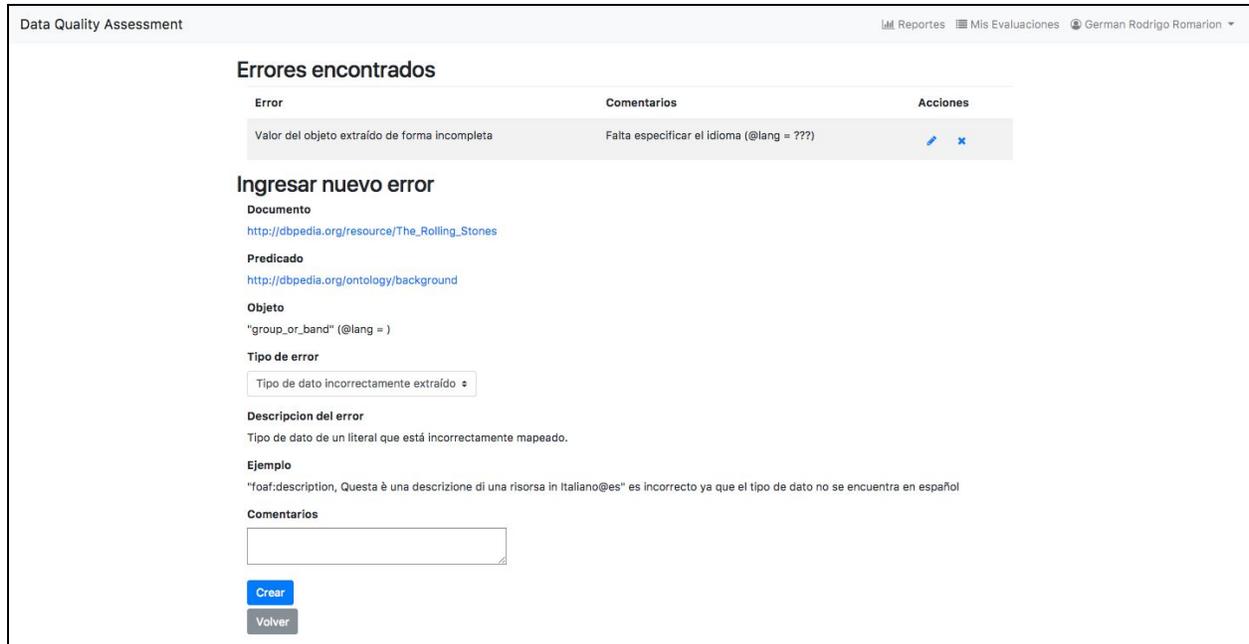


Imagen 4.10: Pantalla para reportar errores. En la parte superior se ven los errores encontrados.

En la **Imagen 4.10** puede verse la disposición de la pantalla, una vez que se reportó un error. En la parte superior se puede visualizar una tabla que contiene los errores hallados por el usuario, y comentarios que haya agregado. Si hace clic en el ícono en forma de lápiz, el usuario puede editar tanto el tipo de error, como los comentarios que haya hecho. Si hace clic en la cruz, se eliminará el error reportado.

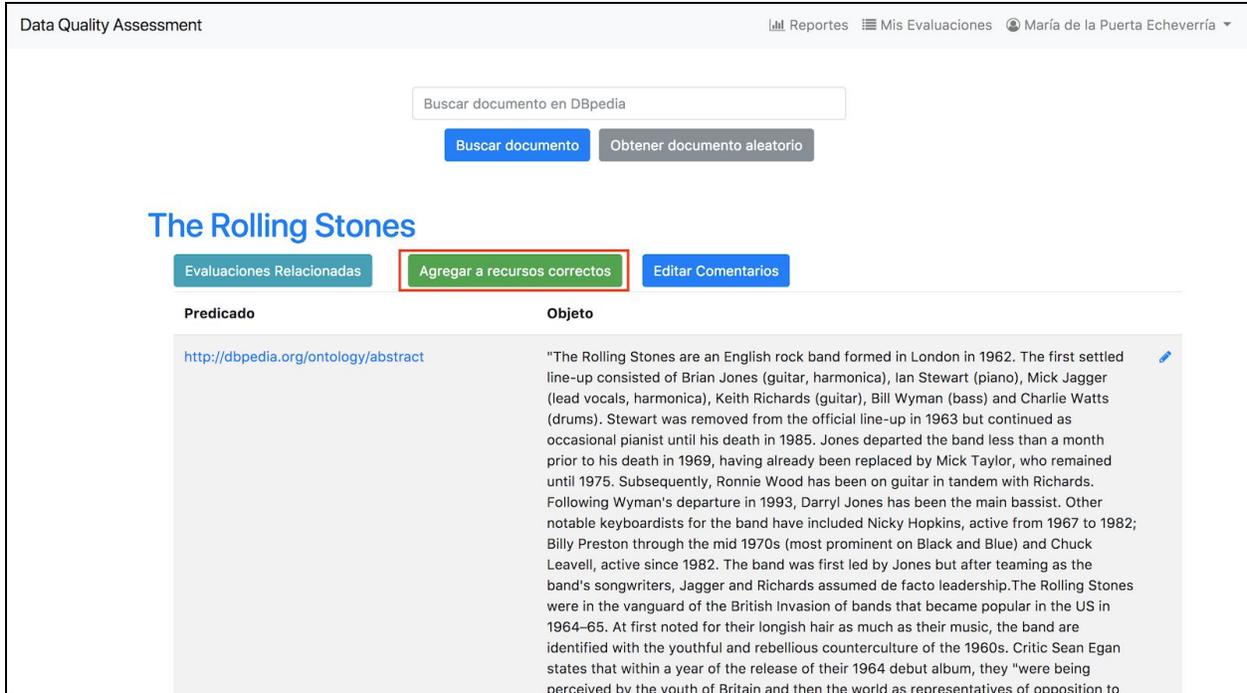


Imagen 4.11: Página de un documento. En rojo se muestra la opción para marcar al documento como correcto.

En caso que no se hayan encontrado errores se puede marcar a ese documento como correcto utilizando el botón “Agregar a documentos correctos” al inicio del listado. Se puede también dejar un comentario sobre ese documento haciendo clic en “Editar comentarios”.

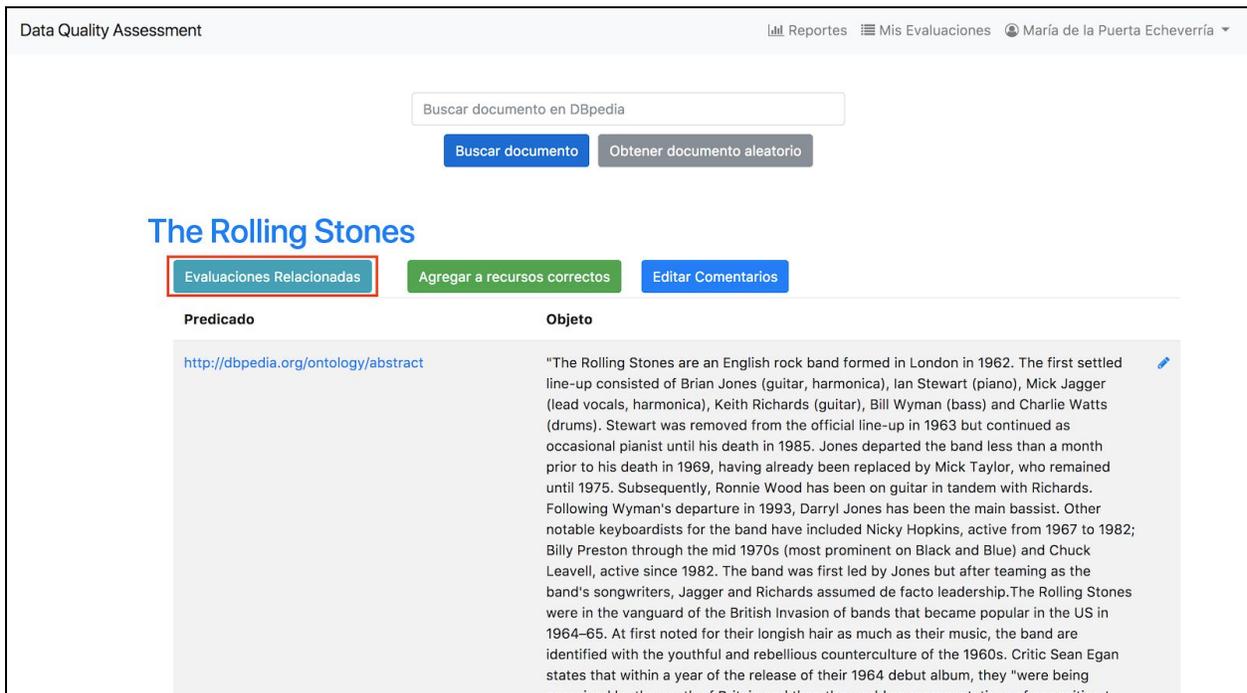


Imagen 4.12: Página de un documento. En rojo se muestra la opción para ver evaluaciones relacionadas.

Si el usuario quiere ver cómo otros usuarios evaluaron ese mismo documento puede hacerlo haciendo clic en “Evaluaciones relacionadas”, como se ve en la **Imagen 4.12**.

Data Quality Assessment Reportes Mis Evaluaciones María de la Puerta Echeverría

Evaluaciones relacionadas a
http://dbpedia.org/resource/The_Rolling_Stones

Usuario	Puntaje	Fecha de evaluación	
faculaderete2	0.999	11/12/2017 02:27	👁
germanroma	1	13/02/2018 15:56	👁

Imagen 4.13: Listado de evaluaciones relacionadas al documento “The Rolling Stones”.

Allí podrá ver en un listado los usuarios que realizaron una evaluación y el puntaje que se le asignó al documento según esa evaluación, como puede verse en la **Imagen 4.13**.

En el ejemplo de la **Imagen 4.13** se observan dos evaluaciones relacionadas al documento *The Rolling Stones*: el usuario **faculaderete2** señaló al menos un error para dicho documento, lo cual le otorgó un puntaje de 0.999 al mismo según las fórmulas planteadas, mientras que el usuario **germanroma** marcó al documento como correcto, lo cual le da el puntaje máximo que se puede obtener, es decir 1.

Data Quality Assessment		Reportes	Mis Evaluaciones	German Rodrigo Romarion
http://dbpedia.org/ontology/wikiPageRevisionLink	http://en.wikipedia.org/w/index.php?title=Edgar_Allan_Poe&oldid=775082344			
http://dbpedia.org/ontology/wikiPageRevisionLink	http://en.wikipedia.org/w/index.php?title=Edgar_Allan_Poe&oldid=819198131			
http://dbpedia.org/ontology/wikiPageRevisionLink	http://en.wikipedia.org/w/index.php?title=Edgar_Allan_Poe&oldid=801507915			
http://dbpedia.org/property/almaMater	http://dbpedia.org/resource/United_States_Military...			
http://dbpedia.org/property/almaMater	http://dbpedia.org/resource/University_of_Virginia			
http://dbpedia.org/property/b	"no" (@lang = en)			
http://dbpedia.org/property/birthDate	"1809-01-19" (@type = http://www.w3.org/2001/XMLSchema#date)			
http://dbpedia.org/property/birthName	"Edgar Poe" (@lang = en) 			
http://dbpedia.org/property/birthPlace	"Boston, Massachusetts, U.S." (@lang = en)			
http://dbpedia.org/property/bot	"InternetArchiveBot" (@lang = en)			
http://dbpedia.org/property/caption	"1849" (@type = http://www.w3.org/2001/XMLSchema#integer)			
http://dbpedia.org/property/date	"January 2018" (@lang = en)			
http://dbpedia.org/property/deathDate	"1849-10-07" (@type = http://www.w3.org/2001/XMLSchema#date)			
http://dbpedia.org/property/deathPlace	"Baltimore, Maryland, U.S." (@lang = en)			
http://dbpedia.org/property/fixAttempted	"yes" (@lang = en)			

Imagen 4.14: Pantalla del documento. Se marca en rojo la propiedad que cuenta con al menos un error.

Al hacer clic sobre el ícono en forma de ojo, la aplicación redirigirá al usuario a la pantalla del documento en cuestión, pero esta vez en modo lectura, es decir que no podrá realizar evaluaciones sino solamente ver las evaluaciones que realizó otro usuario sobre dicho documento. En la **Imagen 4.14** se observa que solo las propiedades del documento que cuentan con evaluaciones tienen el ícono en forma de ojo junto a ellas.

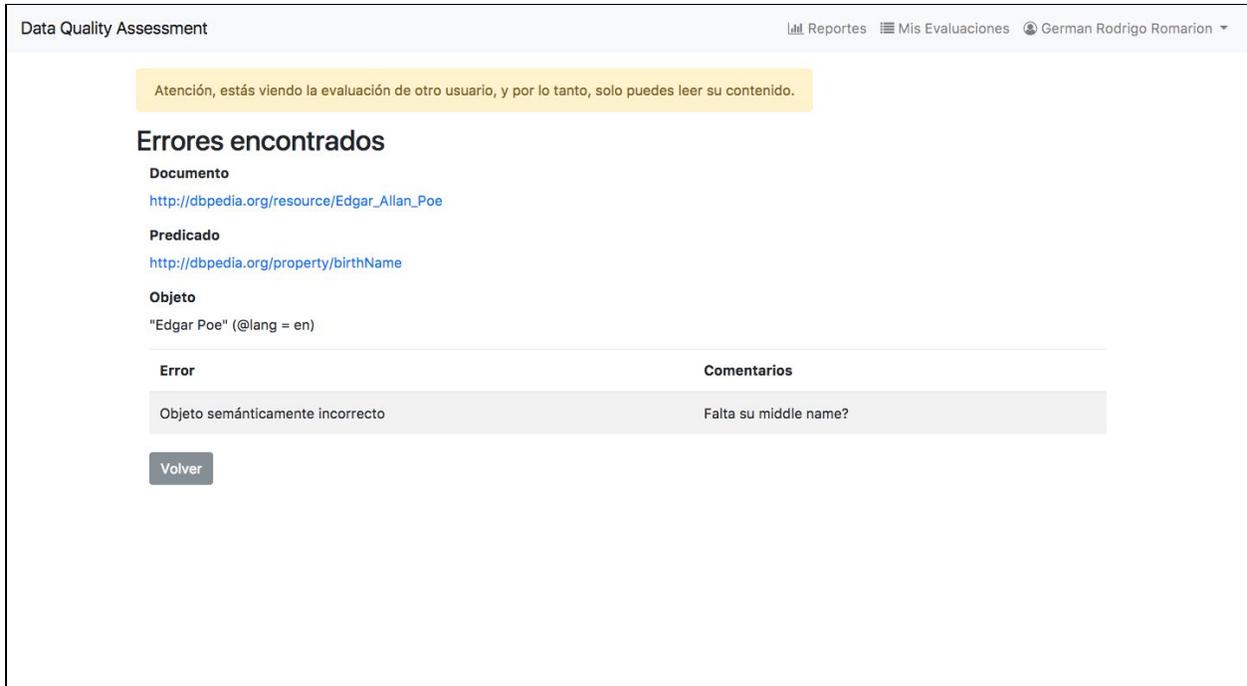


Imagen 4.15: Pantalla de errores encontrados por otro usuario.

Al hacer clic en el ícono en forma de ojo sobre la propiedad evaluada, la aplicación llevará al usuario a una pantalla con el listado de errores marcados por el otro usuario, como puede verse en la **Imagen 4.15**.

El mensaje de advertencia que se lee en la aplicación le recuerda al usuario que está viendo la evaluación de otro usuario, y que por lo tanto no puede editar su contenido.

4.2.6 Mis evaluaciones

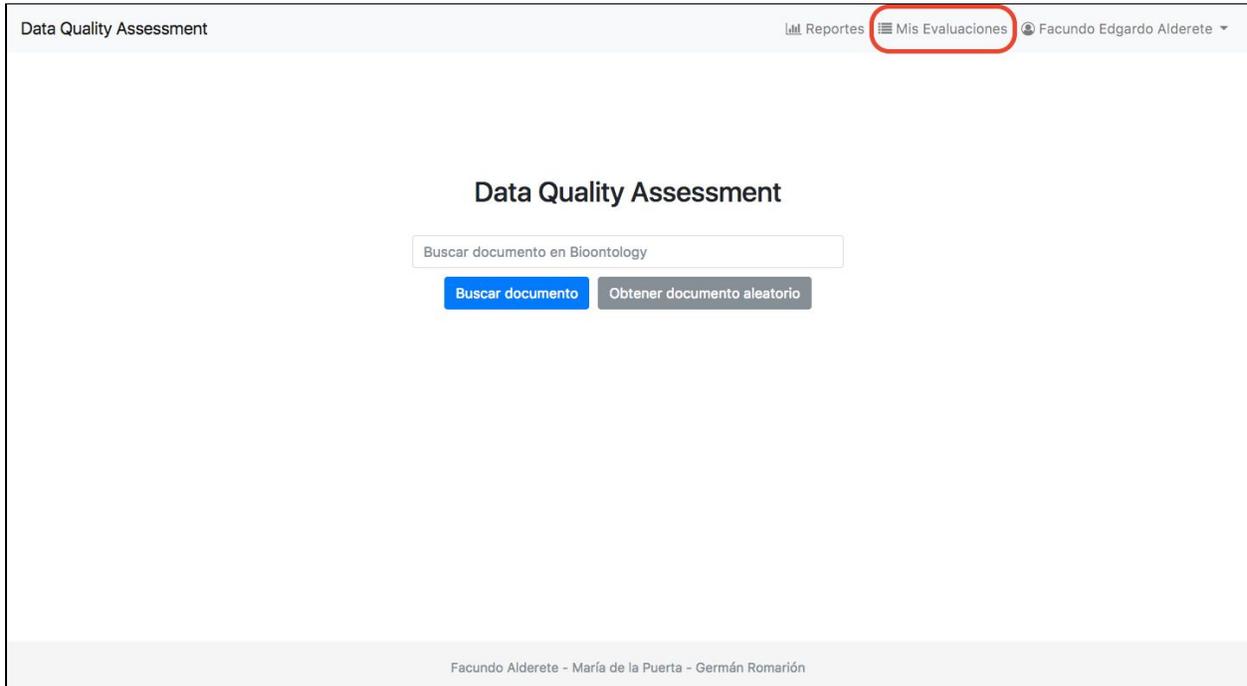


Imagen 4.16: enlace a “Mis Evaluaciones”, marcado en rojo.

El usuario puede acceder al historial de evaluaciones que ha realizado hasta el momento haciendo clic en el enlace “Mis Evaluaciones”, como puede verse en la **Imagen 4.16**.

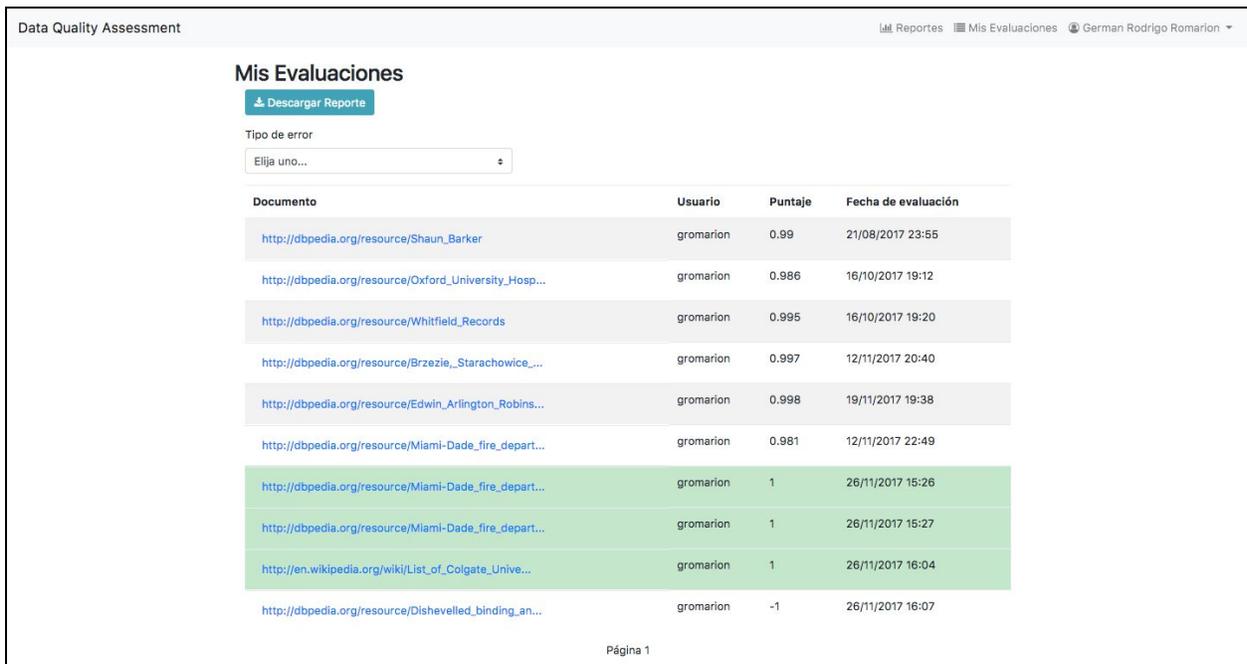
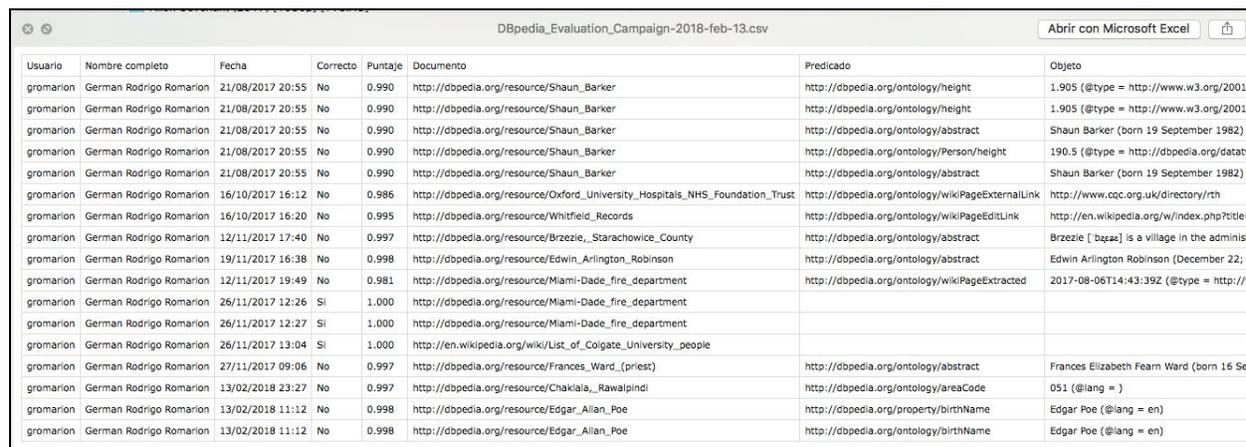


Imagen 4.17: Pantalla donde se listan las evaluaciones del usuario.

Allí se podrá visualizar un listado con El usuario puede acceder al historial de evaluaciones que ha realizado hasta el momento haciendo clic en el enlace “Mis Evaluaciones”, ubicado en la parte superior derecha de la barra de navegación, como puede verse en la **Imagen 4.17**.

Además de poder visualizar y acceder a las evaluaciones que el usuario realizó anteriormente, también puede descargar un reporte en formato CSV (*Column Separated Values*) que contiene la siguiente información:

- Endpoint
- Usuario
- Fecha
- Correctitud
- URL del documento
- Predicado
- Objeto
- Tipo de error encontrado



Usuario	Nombre completo	Fecha	Correcto	Puntaje	Documento	Predicado	Objeto
gromarion	German Rodrigo Romarion	21/08/2017 20:55	No	0.990	http://dbpedia.org/resource/Shاون_Barker	http://dbpedia.org/ontology/height	1.905 (@type = http://www.w3.org/2001/
gromarion	German Rodrigo Romarion	21/08/2017 20:55	No	0.990	http://dbpedia.org/resource/Shاون_Barker	http://dbpedia.org/ontology/height	1.905 (@type = http://www.w3.org/2001/
gromarion	German Rodrigo Romarion	21/08/2017 20:55	No	0.990	http://dbpedia.org/resource/Shاون_Barker	http://dbpedia.org/ontology/abstract	Shaun Barker (born 19 September 1982) is
gromarion	German Rodrigo Romarion	21/08/2017 20:55	No	0.990	http://dbpedia.org/resource/Shاون_Barker	http://dbpedia.org/ontology/Person/height	190.5 (@type = http://dbpedia.org/datat
gromarion	German Rodrigo Romarion	21/08/2017 20:55	No	0.990	http://dbpedia.org/resource/Shاون_Barker	http://dbpedia.org/ontology/abstract	Shaun Barker (born 19 September 1982) is
gromarion	German Rodrigo Romarion	16/10/2017 16:12	No	0.986	http://dbpedia.org/resource/Oxford_University_Hospitals_NHS_Foundation_Trust	http://dbpedia.org/ontology/wikiPageExternalLink	http://www.ccc.org.uk/directory/rth
gromarion	German Rodrigo Romarion	16/10/2017 16:20	No	0.995	http://dbpedia.org/resource/Whitfield_Records	http://dbpedia.org/ontology/wikiPageEditLink	http://en.wikipedia.org/w/index.php?title=
gromarion	German Rodrigo Romarion	12/11/2017 17:40	No	0.997	http://dbpedia.org/resource/Brzezine,_Starachowice_County	http://dbpedia.org/ontology/abstract	Brzezine [bʒɛzɛ] is a village in the administ
gromarion	German Rodrigo Romarion	19/11/2017 16:38	No	0.998	http://dbpedia.org/resource/Edwin_Arlington_Robinson	http://dbpedia.org/ontology/abstract	Edwin Arlington Robinson (December 22; 1
gromarion	German Rodrigo Romarion	12/11/2017 19:49	No	0.981	http://dbpedia.org/resource/Miami-Dade_fire_department	http://dbpedia.org/ontology/wikiPageExtracted	2017-08-06T14:43:39Z (@type = http://w
gromarion	German Rodrigo Romarion	26/11/2017 12:26	Si	1.000	http://dbpedia.org/resource/Miami-Dade_fire_department		
gromarion	German Rodrigo Romarion	26/11/2017 12:27	Si	1.000	http://dbpedia.org/resource/Miami-Dade_fire_department		
gromarion	German Rodrigo Romarion	26/11/2017 13:04	Si	1.000	http://en.wikipedia.org/wiki/List_of_Colgate_University_people		
gromarion	German Rodrigo Romarion	27/11/2017 09:06	No	0.997	http://dbpedia.org/resource/Frances_Ward_(priest)	http://dbpedia.org/ontology/abstract	Frances Elizabeth Fearn Ward (born 16 Sep
gromarion	German Rodrigo Romarion	13/02/2018 23:27	No	0.997	http://dbpedia.org/resource/Chakiala,_Rawalpindi	http://dbpedia.org/ontology/areaCode	051 (@lang =)
gromarion	German Rodrigo Romarion	13/02/2018 11:12	No	0.998	http://dbpedia.org/resource/Edgar_Allan_Poe	http://dbpedia.org/property/birthName	Edgar Poe (@lang = en)
gromarion	German Rodrigo Romarion	13/02/2018 11:12	No	0.998	http://dbpedia.org/resource/Edgar_Allan_Poe	http://dbpedia.org/ontology/birthName	Edgar Poe (@lang = en)

Imagen 4.18: Ejemplo de reporte sobre “Mis Evaluaciones”.

En la **Imagen 4.18** se puede ver un ejemplo del contenido del reporte descargado por un usuario.

4.2.7 Reportes

Todos los tipos de usuario (Administrador, Evaluador e Invitado) tienen acceso a la pantalla de reportes de cada *endpoint*.

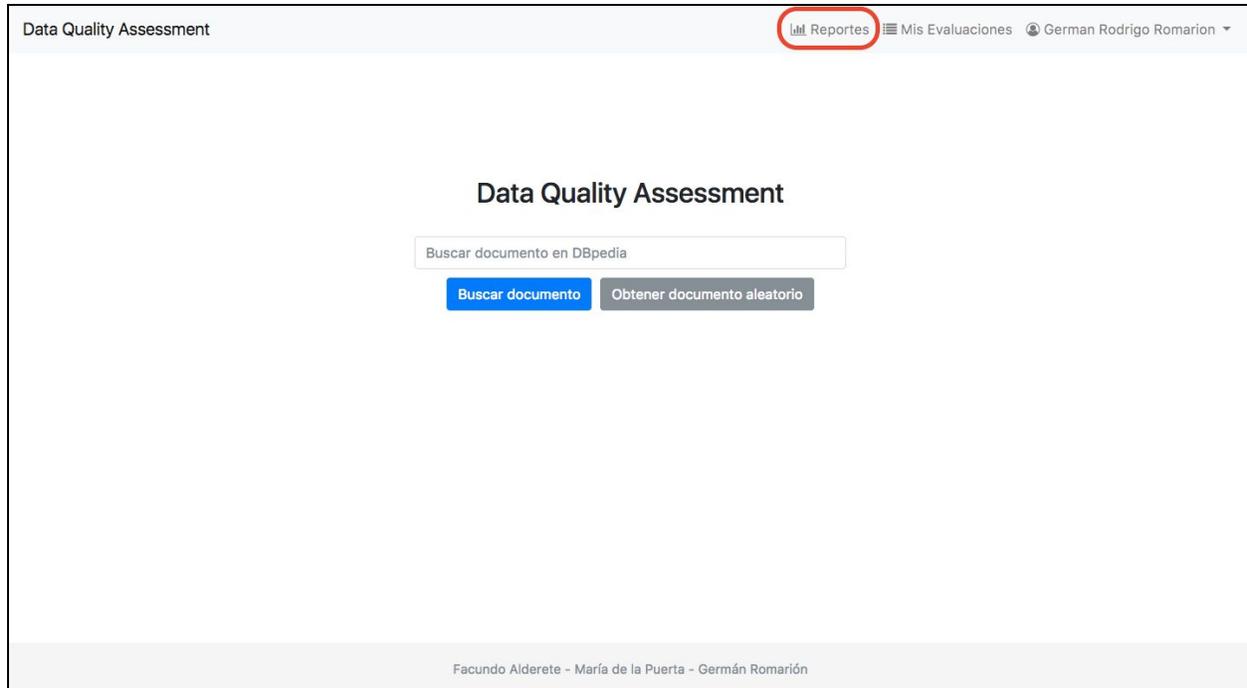


Imagen 4.19: Enlace a la pantalla “Reportes”, marcado en rojo.

Al hacer clic en “Reportes”, tal como se muestra en la **Imagen 4.19**, la aplicación llevará al usuario a la pantalla de reportes generales del endpoint, donde se recolectan las siguientes estadísticas:

- Calificación global
- Estadísticas del endpoint
- Calidad de los documentos

La **Calificación global** promedia los resultados obtenidos en las secciones **Estadísticas del endpoint** y **Calidad de los documentos**. En esta sección se muestra un valor entre 0 y 100, acompañado de una letra que representa dicho valor (Ver sección 3.4.2).

Las **Estadísticas del endpoint** muestran tres métricas puntuales:

- Estado del servidor:
 - **NORMAL** indica que el servidor se encuentra operacional
 - **CAÍDO** indica que no se encuentra disponible
- Disponibilidad: muestra el porcentaje de tiempo que el servidor contestó solicitudes de forma exitosa

- Licencias: indica si el servidor cuenta con licencias para el uso de sus datos

La **Calidad de los documentos** muestra las siguientes métricas en la pantalla:

- Total de documentos evaluados: cantidad de documentos evaluados por todos los usuarios que iniciaron sesión en dicho *endpoint*.
- Calidad promedio de los documentos
- Porcentaje de documentos que presentan el error **Tipo de dato incorrectamente extraído**
- Porcentaje de documentos que presentan el error **Valor del objeto extraído de forma incompleta**
- Porcentaje de documentos que presentan el error **Objeto semánticamente incorrecto**
- Porcentaje de documentos que presentan el error **Enlace externo incorrecto**

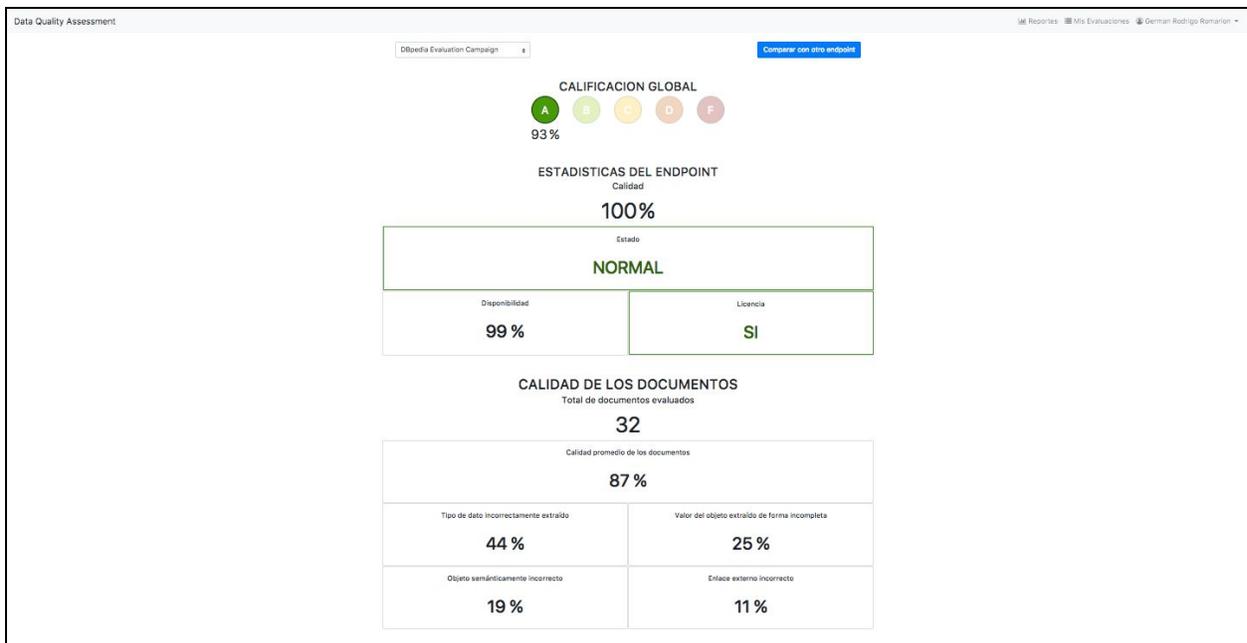


Imagen 4.20: Pantalla de reportes.

Es posible seleccionar otro endpoint por medio del *dropdown* que se encuentra en la parte superior izquierda de la pantalla.

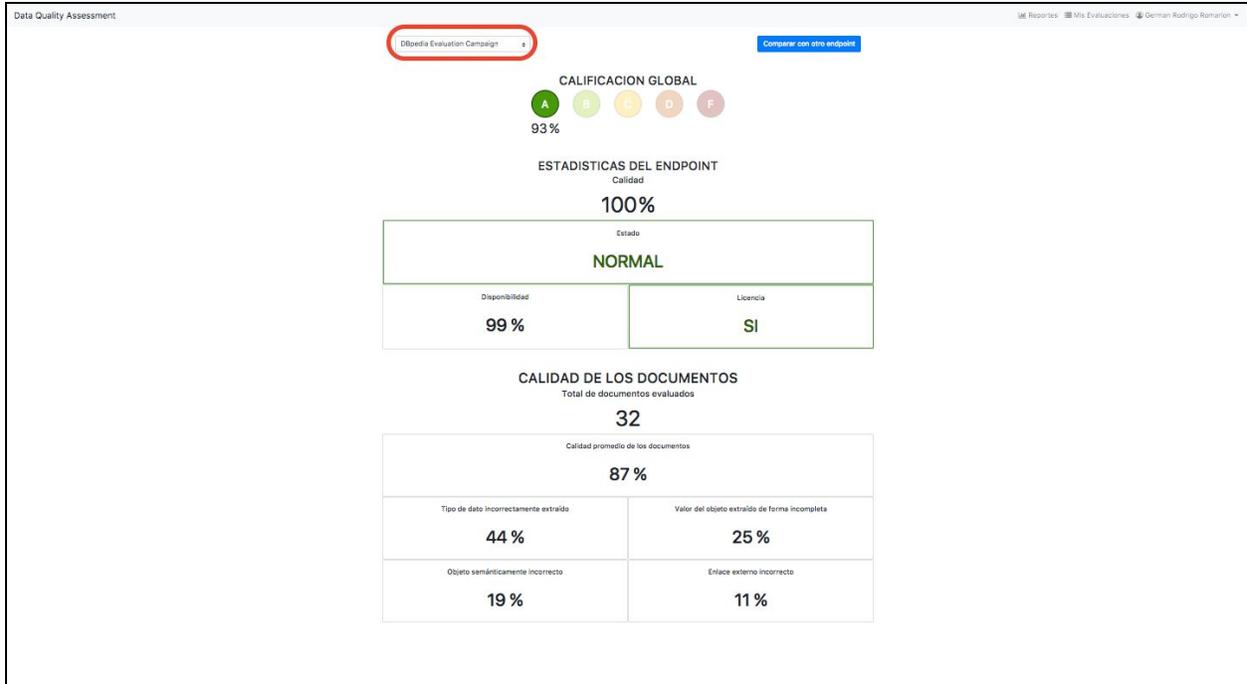


Imagen 4.21: Dropdown de selección de endpoints, marcado en rojo.

El usuario cuenta además con la posibilidad de comparar lado a lado la calidad de dos endpoints, haciendo clic en el botón “Comparar con otro endpoint”, ubicado en la parte superior derecha de la pantalla.

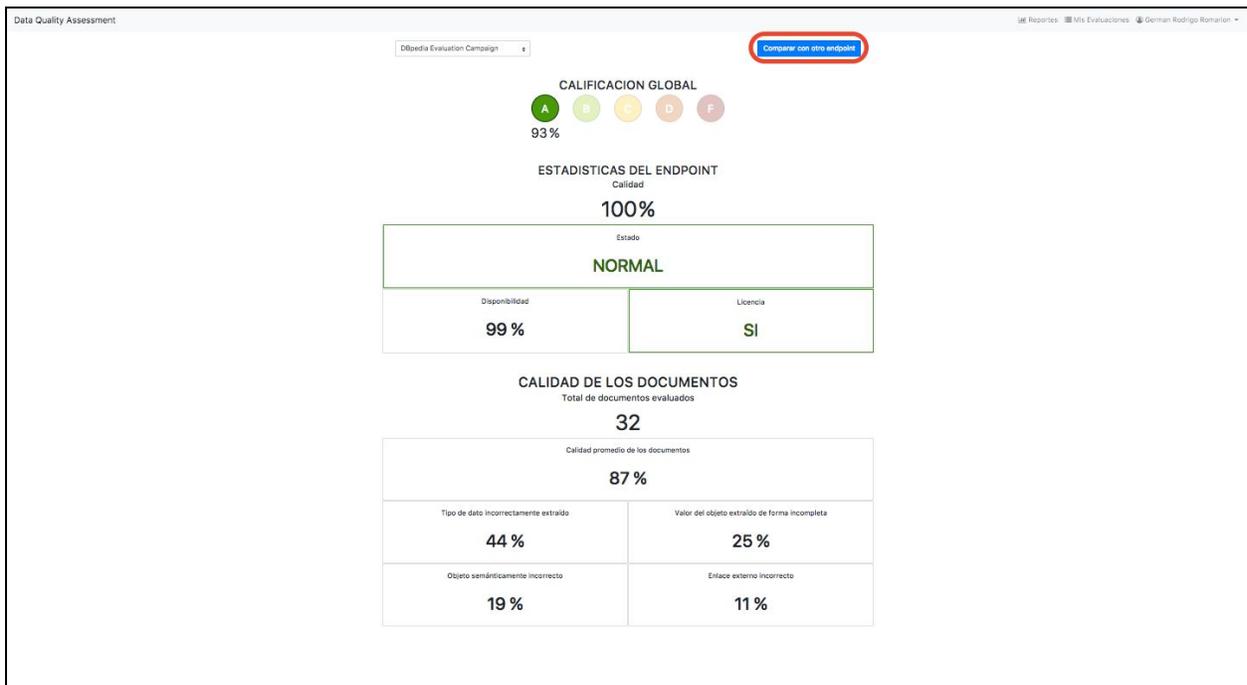


Imagen 4.22: Opción para comparar calidad de dos endpoints, marcado en rojo.

El usuario será entonces redirigido al menú de selección de endpoints a comparar.

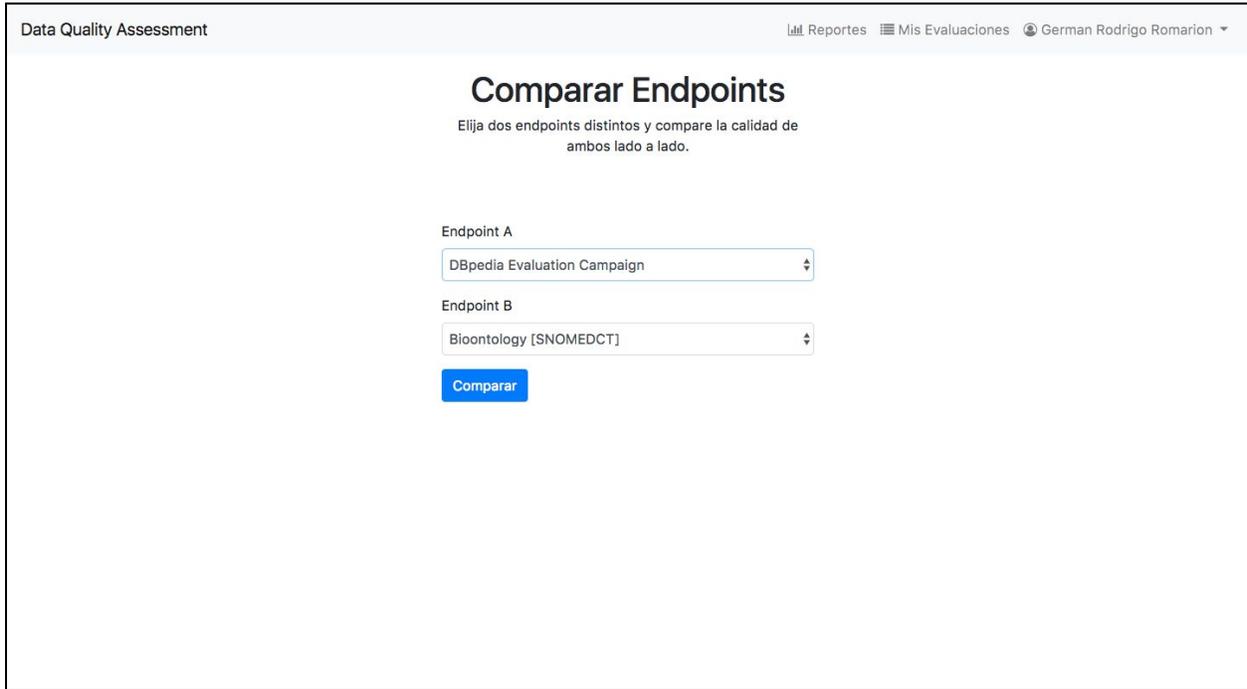


Imagen 4.23: Pantalla de menú de selección de endpoints a ser comparados.

Una vez que el usuario eligió los endpoints a comparar, la aplicación mostrará en pantalla el mismo esquema de estadísticas mencionado previamente, para ambos endpoints.

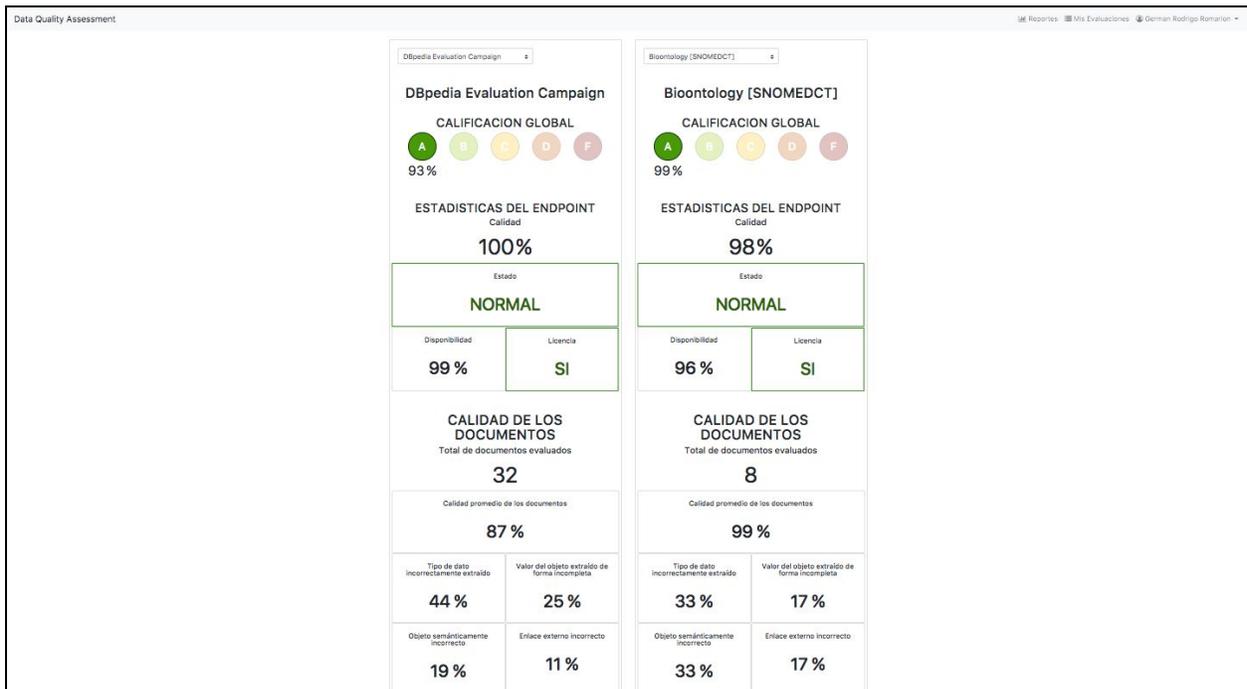


Imagen 4.24: Pantalla de comparación de calidad de endpoints.

4.2.8 Usuario administrador

Como su rol lo indica, este tipo de usuarios pueden crear y modificar endpoints, además de poder dar de baja a usuarios del sistema.

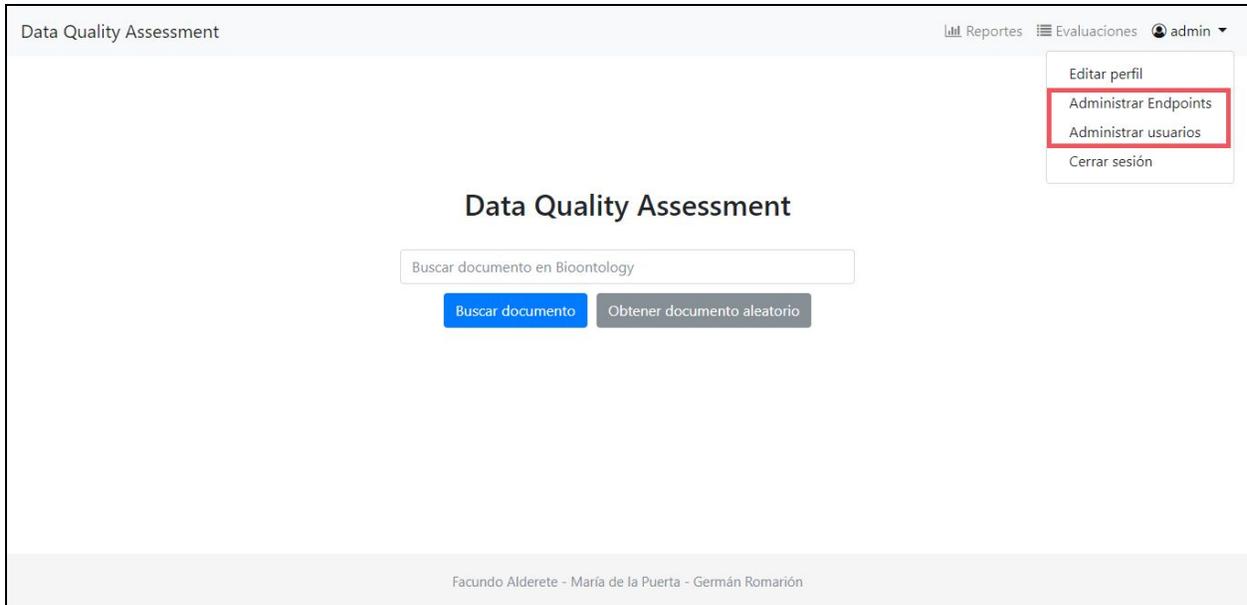


Imagen 4.25: opciones extra con las que cuenta el administrador

En la **Imagen 4.25** se pueden ver dos opciones extra con las que cuenta el usuario administrador en el menú desplegable del usuario: **Administrar endpoints** y **Administrar usuarios**.

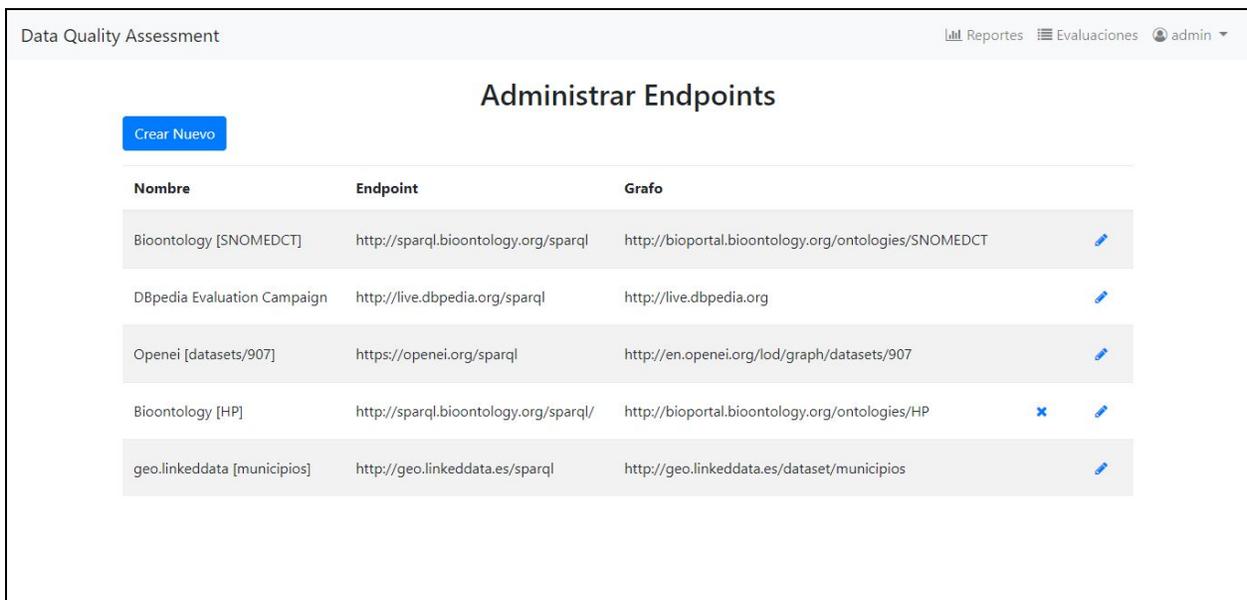


Imagen 4.26: pantalla de administración de endpoints

La primera opción llevará al usuario a una pantalla donde se mostrará un listado con los endpoints cargados en el sistema. Para cada uno de los endpoints que figuran en el listado, el usuario podrá editar sus configuraciones, haciendo clic en el ícono de lápiz, como se muestra en la **Imagen 4.26**. Sólo aquellos endpoints que no cuenten con ninguna evaluación podrán ser eliminados del sistema, haciendo clic en el ícono de cruz.



Imagen 4.27: pantalla de administración de usuarios

Por otra parte, la segunda opción llevará al usuario a una pantalla donde se mostrará un listado con todos los usuarios registrados en el sistema. El administrador podrá ver sus nombres completos, nombres de usuario y fecha de registración. Además, contará con la posibilidad de eliminar a un usuario del sistema, haciendo clic en el link que dice **Borrar**, como se muestra en la **Imagen 4.27**.

The screenshot shows the 'Evaluaciones' page in the Data Quality Assessment system. The page title is 'Data Quality Assessment' and the user is logged in as 'admin'. The 'Evaluaciones' menu item is highlighted in red. Below the title, there is a 'Descargar Reporte' button and a 'Tipo de error' dropdown menu. The main content is a table with the following data:

Documento	Usuario	Puntaje	Fecha de evaluación
http://dbpedia.org/resource/Shawn_Barker	gromarion	0.99	21/08/2017 23:55
http://dbpedia.org/resource/Oxford_University_Hosp...	gromarion	0.986	16/10/2017 19:12
http://dbpedia.org/resource/Whitfield_Records	gromarion	0.995	16/10/2017 19:20
http://dbpedia.org/resource/Brzezine,_Starachowice_...	gromarion	0.997	12/11/2017 20:40
http://dbpedia.org/resource/Edwin_Arlington_Robins...	gromarion	0.998	19/11/2017 19:38

Imagen 4.28: Pantalla de Evaluaciones de todos los usuarios registrados en el sistema.

La opción **Mis evaluaciones**, común para los evaluadores del sistema, se reemplaza por **Evaluaciones**. Tras hacer clic en dicha opción, el sistema llevará al usuario a una pantalla donde se visualiza un listado con las evaluaciones de todos los usuarios registrados en el sistema, para el endpoint en el cual el usuario administrador inició sesión, como puede verse en la **Imagen 4.28**.

Referencias

Quality Assessment for Linked Data: A Survey

Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer.

Crowdsourcing Linked Data Quality Assessment

Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, Jens Lehmann.

creativecommons.org

"... Creative Commons provides free, easy-to-use copyright licenses to make a simple and standardized way to give the public permission to share and use your creative work—on conditions of your choice."

Linked Data SPARQL endpoints

- <http://dbpedia.org/>
- <http://www.lexvo.org/>
- <http://id.loc.gov/authorities/subjects.html>
- <http://www.linked-brain-data.org/sparqlsearch.jsp?link=link0>