Importación y Almacenamiento de Datos Educativas de PISA 2012 a un Manejador de Bases de Datos Relacional

Jesús Donaldo Osornio Hernández, Luis Alejandro Herrera León, Gabriel López Morteo, Rafael Resendiz Ramirez *

Instituto de Ingeniería, UABC, Mexicali, Baja California, México

Abstract

En este estudio se presenta el procedimiento que se realiza en la adopción de uso de datos abiertos, debido a que esto se ha convertido en una tendencia a nivel mundial. Por ello, se presenta la experiencia en la importación de las bases de datos de PISA 2012 a una base de datos relacional, bases que la OCDE pone a disposición del público en general en formatos CSV, SPSS y SAS. Para ello, se desarrollaron programas escritos con lenguaje de script PERL y AWK, utilizando el manejador de bases de datos de código abierto POSTGRESQL. Se reportan los problemas encontrados durante el proceso de almacenamiento, así como las opciones encontradas para su solución.

Palabras clave: PISA, datos educativos, bases de datos, DBMS, metadatos.

1. Introducción

Actualmente, existe una fuerte tendencia para la adopción del uso de los datos abiertos por parte de diversos organismos nacionales, así como de instituciones del ámbito público y privado a nivel internacional. Los datos abiertos son definidos por la Open Knowledge International como aquellos: "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" (Open Knowledge International). Las instituciones de gobierno tanto de México, como de Estados Unidos de América promovieron nuevas políticas para fomentar el uso de datos abiertos entre la sociedad civil, brindando el acceso a la sociedad, lo cual concomitantemente, fomenta la transparencia.

Estas políticas, derivaron en la creación de dos portales para proporcionar tales datos al público, http://datos.gob.mx para México y http://data.gov para Estados Unidos. Ambos portales, suministran información sobre distintas áreas: economía, salud, energía, medio ambiente y educación. En concordancia como resultado del acceso, se pueden obtener los datos en distintos formatos según los requerimientos del solicitante, ya sea un un experto en computación o en aquella área en la que se necesiten dichos datos.

En el campo de la educación se puede observar la presentación de los datos abiertos por parte de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), la cual pone a disposición del público en general, los resultados por cohorte generacional de PISA (por sus siglas en inglés, Programme for International Student Assessment).

Estos dos portales, son auxiliados por otras instancias en su labor de presentarse como fuentes de datos abiertos, en el área educativa, la Secretaría de Educación Pública (SEP), cuenta con el portal del Sistema Nacional de Registro del Servicio Profesional Docente (http://servicioprofesionaldocente.sep.gob.mx/) donde se encuentran disponibles, los resultados, las estadísticas de evaluaciones de desempeño y diagnóstico a docentes de los niveles básico y media superior, así como la evaluación del desempeño de personal con función de director y la evaluación del desempeño de los evaluadores.

De esta forma, se tiene como finalidad describir como la modalidad de distribución de los datos tiene diversas implicaciones negativas directas para los usuarios de los datos, tal como se muestra en el análisis de resultados de PISA 2012.

1.1. Planteamiento del problema

Los metadatos de PISA se encuentran disponibles para dos herramientas estadísticas: Statistical Product and Service Solutions (SPSS) y Statistical Analysis System (SAS). Este tipo de herramientas permite analizar, interpretar, representar y organizar los datos.

Sin embargo, uno de los problemas con este tipo de herramientas es la ineficiencia que existe en la manera en que se

almacena la información. Por ejemplo, si se desea estudiar la actitud de los estudiantes hacia su propio aprendizaje, por lo regular se realizan diversos estudios, a diferentes tipos de escuelas, por ejemplo, en el caso del almacenamiento de datos de un proyecto que conste de una encuesta de 15 preguntas, que se hayan realizado a los alumnos de los seis grados de primaria de cinco escuelas distintas, cuya población promedio sea de 500 estudiantes por grado, así como 25 variables adicionales con los datos característicos por estudiante, se debe considerar que el archivo de datos tendrá una extensión de 6 * 5* 500 = 15,000 estudiantes o renglones, mientras que tendrá 15 + 25 = 40 columnas o variables (preguntas más datos característicos).

Uno de los problemas que suscita el archivo de datos derivado del ejemplo anterior al ser generado por medio de herramientas estadísticas es que se produce una tabla bidimensional simple. Pues, cada renglón corresponde a una observación y cada columna a una variable. Esta representación funciona para el análisis de datos pero es una manera ineficiente de almacenar información redundante.

En cambio, en un ambiente de base de datos se cuenta con una serie de tablas bidimensionales enlazadas entre sí por un índice o una llave, la cual permite que las tablas complejas y redundantes sean construidas al requerirse. Este sistema aplicado en el ejemplo facilita la obtención de diversas tablas de información, una para los estudiantes, otra para los profesores, otra más para las escuelas y la última para las respuestas. Las tablas se combinarán en un sola tabla cuando se ejecute una consulta que así lo requiriera.

1.2. Antecedentes

El Programa para la Evaluación Internacional de Estudiantes (PISA), es una evaluación aplicada internacionalmente que se encarga de medir las habilidades de lectura, matemáticas y literatura científica de los estudiantes de 15 años de edad o un poco mayores a esa edad.

La evaluación PISA tiene sus inicios en el año 2000, es aplicada cada tres años, es coordinada por la OCDE, organización intergubernamental de países industrializados, fundada en 1961 que "…agrupa a 34 países miembros y su misión es promover políticas que mejoren el bienestar económico y social de las personas alrededor del mundo." (Organización para la Cooperación y el Desarrollo Económicos, 2010).

México ha participado en todas las aplicaciones de PISA desde la primera en el año 2000 hasta la última en el año 2015 (National Center for education Statistics, 2015).

Un resumen ejecutivo de los resultados de las evaluaciones PISA se encuentran disponibles en la página web de la OCDE en la sección de Resultados (Organización para la Cooperación y el Desarrollo Económicos, 2015), donde es posible descargar las bases de datos de cada año en el que se aplicó la evaluación PISA. Los resultados de las evaluaciones son publicados en la página de la OCDE un año después de su aplicación.

La información de los resultados de PISA en México es difundida ampliamente por el Instituto Nacional para la Evaluación de la Educación (INEE) a través de su página web y de forma impresa, haciéndola llegar a las secundarias y bachilleratos del país (Instituto Nacional para la Evaluación de la Educación, 2013).

No obstante, si por alguna razón se requiere trabajar con las bases de datos de PISA, éstas se deben descargar de la página de la OCDE en alguno de los siguientes formatos: en archivos valores separados por coma (CSV por sus siglas en inglés Comma Separated Values), donde cada archivo contiene los datos para cada una de las cinco bases de datos de las aplicaciones realizadas, así como otro archivo en texto con las instrucciones y metadatos de cada examen, es decir, archivos para el software estadístico SPSS de IBM, o bien, archivos para el software estadístico SAS de la empresa SAS Institute Inc.

Esto conlleva a concluir que para trabajar con las bases de datos, se tienen dos opciones, la primera consiste en adquirir las licencias de cualquiera de los softwares propietarios, así como la capacitación correspondiente en el uso de estas plataformas, o bien, la segunda opción que requiere el desarrollo de software a la medida para operar los archivos de texto.

En cualquiera de estos casos, el nivel de especialización del usuario es muy alto, añadiendo a esto, tanto el elevado costo de las licencias, como el desarrollo de software, por lo que se limita la capacidad de acceso a la información para el público en general.

Se pretende que cualquier usuario, sea investigador, funcionario, padre de familia o alumno, que esté interesado en trabajar

con los datos de uno o varios países, tenga un fácil acceso tanto a los datos, como a su posterior manejo y análisis, lo cual implica, que la obtención del conjunto de datos sea acorde con el formato que se ajuste a las capacidades de software y conocimiento de su manejo de parte del usuario.

Por otra parte, las bases de datos de PISA que la OCDE mantiene disponibles para descarga corresponden a las diferentes aplicaciones de la evaluación que ha realizado durante la existencia de este mismo programa, con la salvedad, de que dichas bases, no comparten una estructura uniforme o consistente respecto a las demás. Esto ocasiona que el análisis y consulta de información de estos datos sea mucho más complicado, al tener que implementar una solución diferente para cada cohorte o generación de pruebas, o bien, para leer la estructura de cada base de datos correspondiente a su propio año de aplicación. En este estudio se ha utilizado la base de datos de "PISA 2012 Dec 03", por ser la más reciente en el momento de inicio, por lo cual, se relatará un informe acerca de las implicaciones sobre su acceso, lectura y almacenamiento.

Es importante indicar que el término "base de datos" utilizado en esta sección no hace referencia a una base de datos relacional ni a un Sistema Gestor de Base de Datos (RDBMS por sus siglas en inglés), sino al término utilizado en la página web de la OCDE para referirse al conjunto de datos de cada evaluación PISA, integrado por distintos archivos y formatos, los cuales se explican a detalle en la sección de Metodología.

1.3. Objetivo

El objetivo de este estudio es documentar las incidencias del proceso de almacenamiento de la información de las bases de datos de PISA 2012 en un manejador de base de datos relacional mediante los lenguajes de programación AWK y PERL, a fin de desarrollar almacenes de datos homogéneos con acceso y recuperación más sencilla, así como los metadatos de grandes bases de datos abiertas provistos por fuentes originales tales como la OECD.

2. Metodología

El proyecto requería descubrir la manera más eficiente en el procesamiento y almacenamiento de la base de datos denominada "PISA 2012 Dec 03", cuyo peso de archivos oscila entre 14Mb y 1.14 Gb.

Los datos de PISA se componen de múltiples registros para los siguientes aspectos:

- Base de datos completa de PISA:
- Estudiantes (634 campos, 480,174 registros)
- Escuelas (291 campos, 18,139 registros)
- Padres (143 campos, 102,032 registros)
- Respuesta cognitiva (264 campos, 485,490 registros)
- Respuesta cognitiva calificada (216 campos, 485,490 registros)

- Base de datos de México
- Estudiantes (634 campos, 33,806 registros)
- Escuelas (291 campos, 1,471 registros)
- Padres (143 campos, 31,553 registros)
- Respuesta cognitiva (264 campos, 33,806 registros)
- Respuesta cognitiva calificada (216 campos, 33,806 registros)

Los resultados de PISA se encuentran en la página de la OECD, en la cual están disponibles para su descarga los archivos de texto en formato comprimido para el caso de los datos.

Los metadatos están tipificados como archivos de control para la herramienta estadística llamada Statistical Product and Service Solutions de IBM (SPSS), así como para la herramienta Statistical Analysis System (SAS), mismos que se pueden descargar o guardar como archivos de texto desde el mismo navegador.

Durante el desarrollo de este proyecto, también se trabajó con los archivos de texto de datos que fueron formateados acorde a los archivos de control de la herramienta que se había utilizado previamente denominada SPSS.

El archivo de datos contiene un renglón por cada registro, cada uno de los cuales contiene los datos individuales, separados por los caracteres definidos en el archivo de control.

El archivo de control tiene una estructura para su proceso en la herramienta SPSS, la cual se compone de las siguientes secciones:

- Identificador del campo
- Tamaño de caracteres del campo
- Tipo de datos del campo
- Descripción del campo
- Datos de tablas auxiliares
- Valores faltantes

Después de analizar los archivos se desarrollaron dos programas en los lenguajes de programación PERL y AWK para procesar la información y almacenarla en una base de datos para el manejador PostgreSQL.

Los programas se efectuaron siguiendo una metodología de desarrollo rápido, la cual consiste en un proceso iterativo de prototipos donde se realizan los cambios necesarios según se requiera, teniendo en cuenta el resultado de las pruebas.

¿Por qué utilizar PERL y AWK?

La decisión de utilizar ambos lenguajes de programación, se debió al soporte que existe para el procesamiento de archivos con ayuda de las expresiones regulares y que al no ser compilados permite la edición rápida del código fuente para realizar cualquier corrección o cambio, además de que ambos lenguajes pueden trabajarse sin problemas tanto en el sistema operativo Windows como en Linux sin tener que modificar el código.

El investigador que desea analizar datos de PISA, debe tener conocimiento sobre el acceso y almacenamiento de bases de datos, lo que implica tanto la descarga, como la integración de los metadatos con el conjunto de datos. Por otro lado, debido al tamaño de las bases de datos en ocasiones su manejo resulta intratable, puesto que como se comentó oscila desde unos cuantos megabytes por archivo, hasta poco más de un gigabyte.

La cantidad de registros, así como de variables es de tal magnitud que el costo computacional es muy alto, pues la velocidad de procesamiento de los equipos se ralentiza, no solamente por la lectura de la base de datos, sino también por las operaciones o consultas que se realizan sobre ellos.

Algunas de las desventajas en el software comercial, como es el caso de SPSS o SAS, es que en ocasiones la exportación de los datos a bases de datos no es posible debido a situaciones no previstas en la separación de metadatos o en la restauración de las bases de datos, tales como, falta de soporte para manejadores de bases de datos distintas al modelo relacional, conversión inadecuada en cuanto al formateo de variables derivado de la exportación, errores en la estructura del archivo, de la sintaxis, sea por el software o por la fuente. Tales errores pueden ocasionar fallas en los procedimientos de exportación o análisis de datos.

La mayoría de las versiones de Linux soportan ambos lenguajes por defecto, aunque cuando AWK no cuenta con librerías de bases de datos, se desarrollaron programas que se utilizaron para extraer desde los archivos de metadatos la información necesaria para interpretar el archivo de datos de PISA y poder almacenar la base de datos en PostgreSQL, lo que implica que no se necesitan librerías, ni módulos para acceder a una base de datos.

En el caso de PERL, se le dio preferencia en el desarrollo porque es considerado uno de los lenguajes más rápidos en el procesamiento de texto, ya que cuenta además con una conexión a base de datos más sencilla.

3. Resultados

Se desarrollaron múltiples prototipos en PERL para verificar que se realizara el almacenamiento de todos los registros de manera efectiva. Se logró la construcción de un programa que almacenaba dichos registros, pero se presentaron errores al procesar el archivo de metadatos (.SPSS) debido a que no había consistencia en algunos renglones. Se puede observar a manera de ejemplo, la línea 149 del archivo de metadatos (Ilustración 1), donde en los datos de Padres la variable NC tiene una longitud diferente a las demás variables, esta clase de inconsistencias tiene como consecuencia el surgimiento de errores de diverso tipos en el momento de procesar el archivo con las expresiones regulares.

Ilustración 1. Ejemplo de un segmento del archivo de metadatos de Padres

```
NC  "National Centre 6-digit Code"
CNT         "Country code 3-character"
OECD        "OECD country"
SUBNATIO    "Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)"
STRATUM     "Stratum ID 7-character (cnt + region ID + original stratum ID)"
```

Fuente: Metadatos del Cuestionario aplicado a Padres en la evaluación PISA de la OCDE.

En el caso de AWK se siguió un procedimiento similar al procedimiento utilizado con PERL. Se programaron dos scripts en AWK para procesar los datos de PISA. El primero de ellos, para leer el archivo de metadatos (.SPSS), así como para obtener la longitud, tipo y valores de cada variable en el archivo de datos, mientras que, el segundo se construyo para leer el archivo de datos y separar cada registro en campos, para posteriormente guardar toda esa información en un archivo CSV.

El archivo CSV, es un formato intermedio de almacenamiento de la información de los datos de PISA para su ulterior uso en la base de datos PostgreSQL. Esto se realiza, porque como ya se comentó anteriormente, el lenguaje AWK no cuenta con librerías oficiales para la conexión a base de datos como otros lenguajes de programación.

La falta de una librería de AWK para la conexión a PostgreSQL, derivó en la búsqueda de alternativas apropiadas, resultando como derivado de ello, la opción de elaborar varios archivos CSV que representaran las tablas de información para almacenarlas en la base de datos, siendo necesario generar un archivo SQL con las instrucciones para crear la nueva base de datos (con una nomenclatura correspondiente a los archivos originales de la prueba PISA), así como las tablas e instrucciones de copiado de los registros desde los archivos CSV hacia las tablas de la base de datos.

El proceso de transferencia de archivos CSV a PostgreSQL se realizó llamando al comando psql desde AWK, el cuál es el cliente en línea de comandos de PostgreSQL.

El resultado de este proceso fue la generación de una nueva base de datos con estructura de estrella en PostgreSQL, para cada archivo de datos PISA. Esto es, una tabla principal donde todas sus columnas son llaves foráneas hacia tablas auxiliares que contienen la descripción asociada a cada llave. Las tablas auxiliares son generadas con los metadatos extraídos del archivo SPSS, mientras que la tabla principal es generada con los registros del archivo de datos.

La creación de los archivos intermedios en formato CSV realizada para superar la limitante del lenguaje, tiene la ventaja adicional de que dichos archivos, pueden ser reutilizados ya sea para procesamientos posteriores, o, para la transferencia de información a diversas bases de datos. Además, la utilización del cliente psql, permitió obtener una tasa de transferencia mucho más rápida que la que se logra después de haber implementado la transferencia por medio de un cliente propio del lenguaje, con sentencias INSERT por lotes.
La implementación del script para la lectura de los metadatos (.SPSS) no fue trivial. La sintaxis fácilmente interpretada por el programa SPSS no podía ser interpretada por otros lenguajes de programación al no contar con los módulos necesarios para ello. Pero, la lectura del archivo SPSS, a manera de archivo de texto, facilitó un uso potente y flexible de las expresiones regulares en los lenguajes de programación utilizados para extraer la información necesaria.

La implementación del script en AWK, tuvo como resultado incidencias similares a lo sucedido con el script en Perl, entre ellas, algunos problemas con la estructura de los datos, de los cuales a continuación se mencionan los siguientes:

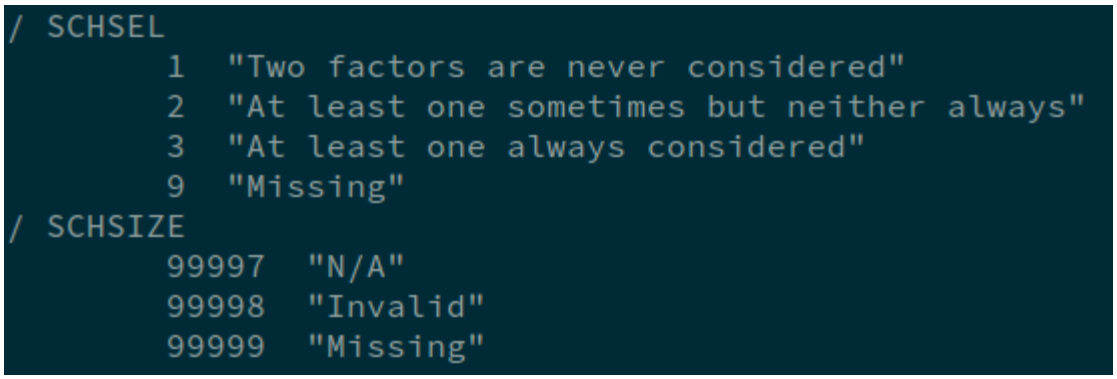
En primer lugar, uno de los principales obstáculos que se presentaron en la interpretación de los metadatos del archivo

SPSS, fue la identificación de variables declaradas que indicaban que podían tomar todos los valores posibles (categóricos) de su propia declaración, mientras que las variables numéricas, sólo indicaban los valores representativos correspondientes a los datos inválidos o perdidos acorde a su propia declaración (Ilustración 2).

Los resultados derivados de la implementación, facilitaron la identificación de problemas en el momento de transferir la información al modelo relacional, ya que, por ejemplo, mientras que la columna SCHSEL de la tabla principal mantenía la integridad referencial con la tabla auxiliar SCHSEL (al tener esta última todos los valores registrados posibles), la columna SCHSIZE de la tabla principal no mantenía dicha integridad, ya que esta columna podría tomar prácticamente cualquier valor pues no se especifica el diccionario de datos válidos, puesto que, los únicos datos que se registraban en la tabla auxiliar SCHSIZE eran los definidos previamente en el archivo SPSS (para datos inválidos o perdidos).

La solución encontrada para este tipo de casos fue la eliminación de la llave foránea de las columnas de la tabla principal para aquellas variables que declararan únicamente valores para datos inválidos o perdidos.

Ilustración 2. Archivo de metadatos de Escuelas



Fuente: Archivo de metadatos de Escuelas en la evaluación PISA de la OCDE.

Otro problema que se encontró en la estructura de los datos fue que, en el archivo de datos correspondiente a las distintas entidades evaluadas (Estudiantes, Padres, Escuelas, etc.), aparecían registros con valores de campos categóricos que no estaban declarados en las variables correspondientes en el archivo de metadatos. Por esto, se puede deducir que existían llaves registradas en algunas columnas de la tabla principal que no fueron registradas en sus propias tablas auxiliares correspondientes, lo que lleva a concluir la existencia de otros errores de integridad referencial.

La solución a este problema, consistió en la generación de los registros faltantes en las tablas auxiliares para evitar errores de integridad referencial al insertar los registros en la tabla principal.

Por último, se observó la existencia u ocurrencia de caracteres inválidos en algunas partes de los archivos de metadatos de cada entidad evaluada, lo que ocasionaba errores de diverso tipo cuando se intentaba insertar la información en PostgreSQL, ya que se utiliza por defecto, el conjunto de caracteres UTF-8.

La solución en esta ocasión, consistió en el procesamiento del archivo sql que había sido generado por AWK con la ayuda del comando SED, antes de ejecutarlo por medio del comando psql para transferir la información a PostgreSQL.

Una vez solucionado todos los problemas encontrados, se pudo transferir la información de los archivos de datos de PISA a las bases de datos en PostgreSQL satisfactoriamente.

4. Conclusiones

Los formatos de distribución de los resultados de PISA por la OCDE, limita a los usuarios de esta información de diversas maneras, entre ellas, el empleo y posiblemente la compra de licencia de herramientas estadísticas de uso comercial. Sirva de ejemplo, el almacenamiento de los metadatos en archivo de texto para los formatos SPSS y SAS, si bien, los datos están disponibles, estos mismos datos no pueden ser utilizados sin los metadatos que le dan sentido a la información, los cuales solamente están contenidos en los formatos propietarios.

Las restricciones que presenta la OCDE en la distribución de los resultados puede ser solucionada siguiendo las tendencias actuales de los datos abiertos, tal como se puede comprobar en portales de acceso abierto, como: http://datos.gob.mx o http://data.gov donde se pretende tener a la disponibilidad de cualquiera los datos públicos.

La OCDE, podría facilitar dicho acceso o consulta mediante la implementación del uso de servicios web para la obtención de resultados, o bien, podría incluir otro archivo de metadatos que no esté limitado a las herramientas estadísticas propietarias. Aunado a ello, sería recomendable crear o facilitar modelos y estructuras de datos más claros, debidamente estandarizados.

Si se utilizaran otros formatos de acceso o distribución de bases de datos, los resultados podrían ser utilizados por otros desarrolladores de sistemas para su almacenamiento. Por ejemplo, en este proyecto se procesaron los resultados para su almacenamiento en el manejador de base de datos de software libre "PostgreSQL".

Algunas de las ventajas de importar o exportar las bases de datos desde o hacia un manejador de base de datos, es una mayor eficiencia en el almacenamiento, ya que la información se encuentre separada en múltiples tablas. Otra ventaja, es la posibilidad de almacenar las bases de datos de otros cohortes de las evaluaciones de PISA, e inclusive, se podrían almacenar en el mismo rubro los datos y metadatos de otras pruebas similares.

Cuando los datos se encuentran ya almacenados en un manejador de bases de datos, los desarrolladores pueden construir clientes especiales para consultar sets de datos, o bien, construir adaptadores de software que permitan exportar sets de datos en una gran variedad de formatos que se acomoden a las necesidades particulares de los usuarios y de esta manera ampliar de una manera significativa el acceso a los datos.

Así un mayor número de investigadores podrían acceder a la información que requieren sin la necesidad de realizar grandes gastos en licencias de software, sin invertir enormes cantidades de tiempo o recursos humanos y materiales para conseguir a alguien con un perfil especializado que les permita emplear las plataformas estadísticas propietarias.

5. Referencias

1. Instituto Nacional para la Evaluación de la Educación. (2013). *Qué es PISA*. Recuperado el 2016 de Septiembre de 27, de Instituto Nacional para la Evaluación de la Educación: http://www.inee.edu.mx/index.php/proyectos/pisa/que-es-pisa
2. National Center for education Statistics. (2015). *Participation in PISA by Year*. (Institute of Education Sciences) Recuperado el 24 de Septiembre de 2016, de National Center for education Statistics: http://nces.ed.gov/surveys/pisa/countries.asp
3. Open Knowledge International. (s.f.). *¿Qué son los datos abiertos?* Recuperado el 26 de Abril de 2016, de Open Data Handbook: http://opendatahandbook.org/guide/es/what-is-open-data
4. Organización para la Cooperación y el Desarrollo Económicos. (2010). *OCDE*. Recuperado el 29 de Octubre de 2016, de Organización para la Cooperación y el Desarrollo Económicos: http://www.oecd.org/centrodemexico/laocde/
5. Organización para la Cooperación y el Desarrollo Económicos. (2015). *PISA Products-OECD*. Recuperado el 2016 de Agosto de 20 , de OECD: http://www.oecd.org/pisa/pisaproducts/