UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
CENTRO DE INVESTIGACIÓN EN SISTEMAS DE INFORMACIÓN

Desarrollo de una Solución de Inteligencia de Negocio para la obtención de indicadores para apoyar la Preservación Web

Trabajo Especial de Grado presentado ante la ilustre

Universidad Central de Venezuela por

Br. Adriana Lopez.

Br. Rosangela Sarno.

Para optar al título de Licenciado en Computación

Tutores:

Profa. Mercy Ospina

Profa. Concettina Di Vasta

Caracas, Octubre 2015

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN

ACTA

Quienes suscriben, miembros del jurado designado por el Consejo de la Escuela de Computación, para examinar el Trabajo Especial de Grado titulado **"Desarrollo de una Solución de Inteligencia de Negocio para la obtención de indicadores para apoyar la Preservación Web"** y presentado por las bachilleres: Br. Adriana Victoria Lopez Rangel C.I.: 25068696 y Br. Rosangela Sarno Marín C.I.: 20491553, a los fines de optar al título de Licenciado en Computación, dejamos constancia de lo siguiente:

Leído como fue dicho trabajo, por cada uno de los miembros del jurado, se fijó el día __ de _____ de _____, a las _____ horas, para que los autores lo defendieran en forma pública, lo que estos hicieron en la Sala ___ de la Escuela de Computación, mediante una presentación oral de su contenido, luego de lo cual respondieron a las preguntas formuladas. Finalizada la defensa pública del Trabajo Especial de Grado, el jurado decidió aprobar con la nota de ____ puntos.

En fe de lo cual se levanta la presente Acta, en Caracas el día ___ de _____ de ____.

_____	_____
Profa. Mercy Ospina (Tutora)	Profa. Concettina Di Vasta (Tutora)
_____	_____
Profa. Brenda López (Jurado Principal)	Profa. Fernando Crema (Jurado Principal)

Dedicatorias

A la persona que más amo. Mi motor, mi ejemplo y mi inspiración desde el día que vine al mundo. Gracias por tanto, mami. Eres mi modelo a seguir, te debo todo, Isbeth, ayudaste a poner mis sueños en el camino correcto para que se hagan realidad. Por esto y más, esto es para ti.

A quien siempre me ha impulsado a saltar al vacío y me sostuvo cuando más lo necesite, abuela ésta y muchas otras aventuras son gracias a lo positivo que ha sido tu influencia en mi vida.

Adriana Lopez

Isa, este trabajo es para ti. Para que te sirva de ejemplo, que todo en la vida se puede, y de que la dedicación y la perseverancia, a pesar de los problemas del camino, si dan sus frutos. No todo el camino es perfecto, pero el final lo es y nunca habrá una manera mejor de lograrlo. Para ti, hermanita, te amo.

Rosangela Sarno

Agradecimientos

A mi hermana, mi morocha, mi mejor amiga desde siempre, Alejandra. Gracias por ser mi compañera de aventuras, por defenderme, apoyarme, impulsarme y siempre estar para mí en este y todos los caminos que hemos recorrido juntas. Love u sister.

A mis tíos, Jhonny, Jenny, Lizabeth, Vivian y Moises, cada uno de ustedes que de un modo u otro me han ayudado a concretar este logro. En especial a mi tío Jhonny, fuiste un padre para mí, te extraño y a mi tía Vivian que ha sido como una suerte de hermana mayor.

A mi tía Xiomara, por ayudarnos en el momento que más lo necesitamos, sin ti se hubiesen complicado nuestros estudios, gracias infinitas.

A Lizabeth "Pichi" Rojas, quien sin entender nada estuvo conmigo en el desarrollo de este Trabajo Especial de Grado, además de ser una de mis compañeras de aventuras.

A mi novio, mi mejor amigo y confidente, Javier Flores, quien me recuerda siempre mi norte y me acompaña hacia este, quien a veces cree en mi más que yo misma y ha recorrido junto a mí este hermoso camino y con él me queda mucho más por recorrer.

A Piñero, quien arranco esta carrera conmigo como una compañera de clases y que hoy por hoy podemos decir que llegamos a la meta juntas como amigas.

A Humberto y David, mis partners in crime, que más que amigos, hoy puedo decir que son familia.

A Gabriela y Oriel, mis locas muchas gracias por todo el apoyo, las risas y los llantos, me encanto compartir con ustedes esta travesía.

Adriana Lopez

A mi hermanita, a mi confidente, a mi mejor amiga, a ti Isabella, por siempre estar a mi lado cuando más te he necesitado, por apoyarme y demostrarme que todo se puede. Te amo osa menor.

A mis padres, por ser grandes ejemplos de perseverancia, dedicación y trabajo duro. Gracias por ayudarme a alcanzar esta meta y por estar siempre a mi lado, apoyándome.

A mis hermanas de la vida, Anita, Alejandra y Rocío, por ser las mejores amigas que alguien podría desear, por ser mi pañito de lágrimas y mis motivadoras personales ¡Las amo mis niñas!

A mi tía María, por estar pendiente de mí, de que no necesitara nada.

A mi Nonna, gracias por prepararme todas esas comidas que me llenaron de energía para seguir adelante.

A toda la dinastía Marín, por ser grandes ejemplos de superación. Todos ustedes que me han demostrado que si se puede, que cualquier meta que uno se proponga es posible. Gracias por ser una familia tan loca e incondicional ¡Los amo a todos!

Rosangela Sarno

A Dios, por bendecirnos y permitirnos cumplir esta meta tan anhelada.

A Mercy Ospina, por su tutoría a lo largo de este Trabajo Especial de Grado, por haber confiado en nosotras para realizar este proyecto, por todos sus consejos y guía, por todo su apoyo incondicional y haber estado ahí para nosotras en todo momento.

A nuestra tutora Concettina Di Vasta, por sus consejos y guía, por siempre estar dispuesta y darnos ánimos en el momento que más lo necesitábamos.

Y a todos que de una forma u otra forma ayudaron a la realización de este Trabajo Especial de Grado.

Adriana Lopez y Rosangela Sarno

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
CENTRO DE INVESTIGACIÓN EN SISTEMAS DE INFORMACIÓN

Desarrollo de una Solución de Inteligencia de Negocio para la obtención de indicadores para apoyar la Preservación Web

Autores: Adriana Victoria Lopez Rangel.
Rosangela Sarno Marín.

Tutoras: Profa. Mercy Ospina.
Profa. Concettina Di Vasta.

Fecha: 05 de Octubre de 2015.

RESUMEN

Los Archivos Web son sistemas de información que surgen para archivar, de manera histórica, documentos que están publicados en la Web, considerados parte del patrimonio digital de las naciones, y tienen como objetivo preservar conjuntos seleccionados de páginas, o sitios web, y sus documentos mediante su replicación y/o migración de su formato original a otra representación. El formato de archivo utilizado para esta tarea es el formato WARC, el cual es un contenedor de documentos Web estandarizado y desarrollado con la finalidad de soportar la preservación. Actualmente, se está desarrollando un prototipo de Archivo Web en Venezuela, en la Facultad de Ciencias de la Universidad Central de Venezuela, como una iniciativa de Preservación Web para sitios web en Venezuela. En esta iniciativa de preservación web, se generan constantemente una gran cantidad de metadatos a partir de los rastreos realizados al conjunto de sitios web venezolanos seleccionados, sin embargo, estos metadatos están embebidos en los archivos WARC o en los *logs* de rastreo, por lo que no son accesibles de manera directa. Los metadatos son definidos como información estructurada que describe, explica, localiza o, de cierta forma, facilita la obtención, uso o manejo de algún recurso, y aunque pueden ser categorizados como descriptivos, administrativos y estructurales, son los metadatos administrativos quienes poseen la mayoría de los metadatos sobre los rastreos realizados dentro del Prototipo de Archivo Web. El objetivo de este Trabajo Especial de Grado, el cual forma parte de dicho prototipo, es definir los datos estadísticos e indicadores útiles para los diferentes usuarios, para apoyar y facilitar las tareas de administración del sistema. Este trabajo forma parte del Prototipo de Archivo Web de Venezuela, y fue realizado bajo una adaptación del método ciclo de vida de Ralph Kimball, para la extracción de los metadatos se utilizó la librería WAT, para el almacenamiento de dichos metadatos, se utilizaron archivos en formato JSON y MongoDB, y para mostrar los indicadores y datos estadísticos se utilizó la herramienta Pentaho.

Palabras Claves: Archivo Web, preservación Web, Formato WARC, Indicadores, Datos Estadísticos.

Índice de Contenido

Índice de Figuras

Índice de Tablas

Introducción

El patrimonio cultural es conocido como la herencia cultural que forma parte de la historia y las tradiciones de los diferentes pueblos que existen y han existido en el mundo, y que será transmitido a las generaciones futuras. Dicho patrimonio puede ser tanto tangible como intangible, aunque es dentro de éste último que se encuentra el patrimonio digital, el cual abarca libros digitales, sitios web, material multimedia, entre otros.

La problemática del patrimonio digital, específicamente el patrimonio web, es la facilidad con la que puede perderse, sea por modificaciones o eliminación del mismo, ya que no es autopreservable. Es por esto, que se han desarrollado mecanismos para conservar los recursos digitales que están en la web. A las actividades asociadas a este proceso de conservación, se les ha dado el nombre de preservación web y, como principal objetivo, busca almacenar la información web relevante para cualquier rama del saber en un lugar alterno, a su localidad actual, de manera segura.

Cada vez son más los países que se unen y comienzan a preservar su patrimonio digital, a veces incluyendo el de países aledaños. Es por esto que un conjunto de iniciativas se ha agrupado, formando el "Consorcio Internacional para la Preservación del Internet" (IIPC, por sus siglas en inglés), encargado de definir estándares y buenas prácticas para la creación y administración de los Archivos Web, con el objetivo de crear en un futuro, un repositorio universal.

Actualmente en Venezuela, se está desarrollando un prototipo de Archivo Web como una iniciativa que busca preservar los sitios web del país y que se encuentra funcionando en un servidor para poder rastrear y preservar dichos sitios web. Estos rastreos realizados han generado una gran cantidad de información relacionada al contenido preservado, que no es aprovechada en su totalidad, y para lograr esto se ha hecho necesario conocer cómo acceder a dicha información, su estructura y su semántica, de manera que pueda ser utilizada para facilitar la gestión de todo el Prototipo.

Es por lo antes mencionado que este Trabajo Especial de Grado propone estudiar esta la información descriptiva, denominada metadatos, que está almacenada en el Archivo, para desarrollar una solución de inteligencia de negocio que apoye a los usuarios en el cumplimiento de sus funciones y les facilite la toma de decisiones relacionadas al Prototipo de Archivo Web de Venezuela.

En el Capítulo 1, se plantea formalmente el problema a abarcar, dándole un contexto completo y justificándolo, así como se define el alcance y los objetivos de este trabajo. En el Capítulo 2, se definen todos los conceptos necesarios para poder desarrollar la solución al problema presentado en el capítulo anterior y las técnicas necesarias que ayudarán a alcanzar los objetivos. En el Capítulo 3, se explica la metodología a utilizar para el desarrollo del trabajo. En el Capítulo 4, se explica formalmente el desarrollo de la aplicación, siguiendo las actividades planteadas en la metodología, y se expone el producto final. Para finalizar, se ofrecen las Conclusiones alcanzadas y algunas propuestas para posibles trabajos futuros.

Capítulo 1 Problema de Investigación

En este capítulo se plantea el problema a solventar y se explica el contexto actual. También, se especifican los objetivos, la justificación y el alcance definidos para esta investigación.

1.1. Planteamiento del Problema

Actualmente en la Universidad Central de Venezuela, se está desarrollando un prototipo de Archivo Web como una iniciativa de Preservación Web para sitios web de Venezuela, el cual se ha implementado de manera modular e incremental. Actualmente consta de los siguientes módulos: un módulo de adquisición (productor), un módulo de indexación y almacenamiento (archivo), y un módulo de acceso al archivo web (consumidor), los cuales se puede apreciar en la Figura 1.

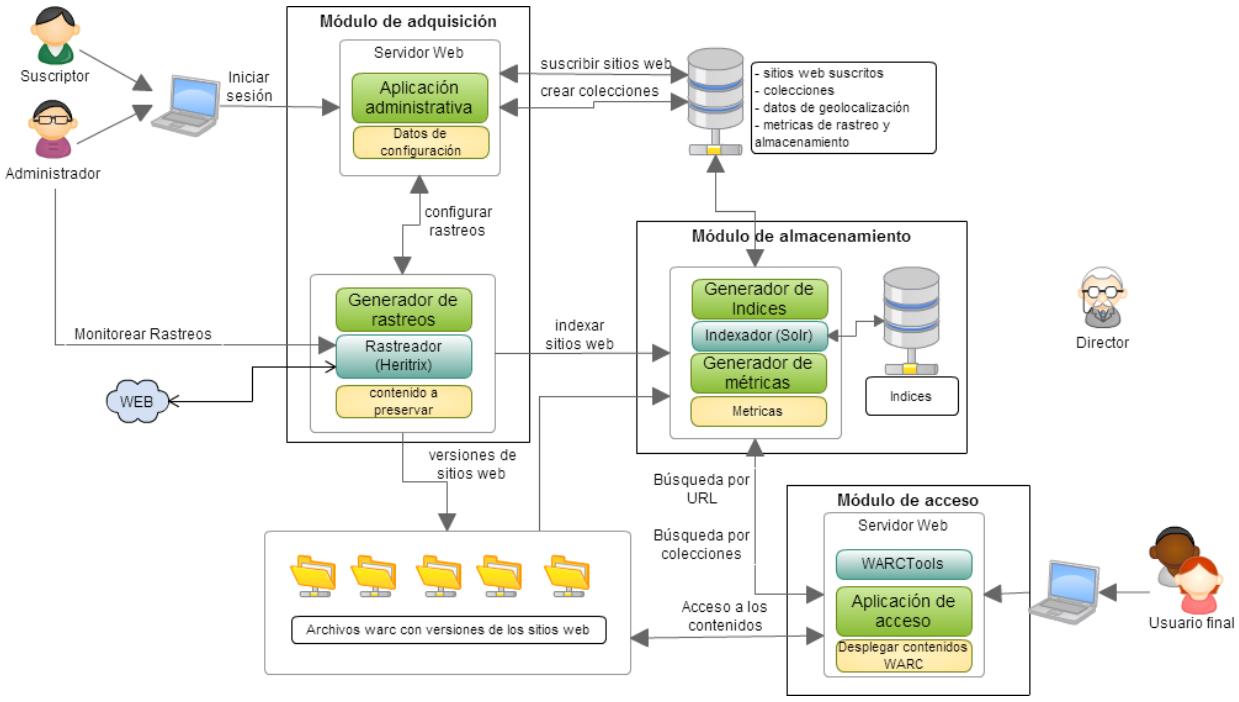


Figura 1 - Arquitectura de un Archivo Web
Fuente: Un Marco de Referencia para la Implementación de Archivos Web (Ospina Torres, 2014)

En el módulo de adquisición se encuentran las herramientas de rastreo, las cuales se encargan de inspeccionar la *World Wide Web* de forma metódica y automatizada, con la finalidad de crear una copia de todas las páginas Web visitadas para su posterior procesamiento, a este proceso se le denomina rastreo o cosecha. (Rivero & García, 2013)

El módulo de indexación y almacenamiento es donde se almacenan las páginas web cosechadas, para luego ser indexadas, y así facilitar futuros accesos a dichas páginas. (Rivero & García, 2013)

El módulo de acceso permite al usuario ingresar y ver el contenido de los archivos almacenados, de acuerdo a la búsqueda que hayan realizado. (Kabchi & Martínez, 2013)

Este sistema está dirigido a diferentes usuarios o actores, los cuales son aquellos entes que interactúan con el Archivo Web. A continuación, en la Tabla 1, se definen los usuarios para el Prototipo de Archivo Web.

Tabla 1 - Usuarios de un Archivo Web

Nombre	**Descripción**
Suscriptor	Rol desempeñado por las personas o los sistemas cliente, que proporcionan la información a ser conservada. Toman decisiones para incluir o excluir elementos (semillas –puntos de entrada de los sitios web a preservar–) o grupos de elementos (colecciones –conjuntos de semillas clasificados por tema–) en cada etapa del flujo, desde la adquisición hasta el almacenamiento. Tienen la responsabilidad de cumplir la política de selección.
Usuario Final	Rol desempeñado por las personas o los sistemas cliente, que interactúan con los servicios del Archivo para encontrar y adquirir información conservada de interés y estadísticas acerca de las métricas recolectadas.
Director	Rol responsable del manejo de los componentes funcionales, análisis de riesgos y costos, y definición de las políticas del Archivo a un nivel superior, así como de la coordinación entre administradores y suscriptores.
Administrador	Rol desempeñado por técnicos u operadores de rastreo, que controlan el flujo de trabajo y su operación diaria. Su tarea es desarrollar, construir, mantener y controlar el flujo de trabajo del Archivo.

Fuente: Un Marco de Referencia para la Implementación de Archivos Web (Ospina Torres, 2014)

En todo proceso de preservación, y en especial en la preservación web se genera una gran cantidad de información descriptiva, o metadatos (Sección 2.2). En esta iniciativa, se está generando una gran cantidad de metadatos de manera constante, a partir de los rastreos realizados a un conjunto de sitios web venezolanos seleccionados, los cuales se describen en la sección 2.2, con los que se pueden generar métricas o indicadores a ser utilizados por los actores antes descritos para apoyar la toma de decisiones. Sin embargo, estos metadatos están embebidos en los archivos WARC, que es un formato de archivo para la preservación descrito en la sección 2.1.3.2, o en los *logs* de rastreo, por lo que no son accesibles, de manera directa, a los usuarios quienes no pueden aprovecharlos de manera efectiva, para así poder conocer el desempeño del sistema.

En este prototipo, los metadatos no están siendo utilizados para la generación de indicadores y datos estadísticos, esto se debe a que son almacenados en un formato complejo que no permite su aprovechamiento, con la utilización de cualquier herramienta de Inteligencia de Negocio, por parte de los usuarios definidos para el Archivo Web. Adicionalmente, se tiene un conjunto de datos posiblemente

valioso para estos usuarios, pero no son utilizados por la falta de conocimiento y entendimiento de los mismos. Estos datos pueden acarrear información que apoye la toma de decisiones, correspondientes al rol de cada tipo de usuario, y permitan mejorar el rendimiento y la utilización del Prototipo de Archivo Web.

1.2. Objetivo

1.2.1. Objetivo General

- Desarrollar una solución de Inteligencia de negocio, utilizando métricas definidas a partir de los metadatos y la información almacenada en la base de datos de preservación del Prototipo de Archivo Web, para apoyar y facilitar la gestión del mismo.

1.2.2. Objetivos Específicos

- Analizar los metadatos existentes en el repositorio WARC y en la base de datos de Preservación, para conocer detalladamente la información almacenada.
- Definir los indicadores, relacionados con los metadatos almacenados, requeridos por los diferentes roles de usuarios del Archivo Web.
- Desarrollar el almacén de datos para los datos y metadatos obtenidos de los diferentes rastreos web.
- Construir un portal web de indicadores sobre los metadatos del Archivo Web.

1.3. Justificación

Actualmente, en el prototipo de Archivo Web se da cierta información descriptiva, como son los metadatos, que ha sido previamente rastreada y almacenada, y se puede consultar en una sección del módulo de adquisición del Prototipo de Archivo Web.

Sin embargo, esta información y los restantes metadatos, que se almacenan en el archivo, no están procesados en forma de indicadores que muestren el crecimiento de los contenidos almacenados en un lapso de tiempo determinado, el comportamiento de las páginas web a través del tiempo, los tipos de archivos almacenados (imágenes, videos, PDFs, entre otros) u otra información de interés sobre el comportamiento del Archivo Web. Por lo tanto, se requiere agregar una funcionalidad, a través del desarrollo de una solución de Inteligencia de Negocio integrada al módulo de adquisición, para que así los usuarios del Archivo Web puedan visualizar los indicadores provenientes de estos metadatos, y de esta manera tomar decisiones de manera rápida y oportuna.

1.4. Alcance

Se implementa una solución de Inteligencia de Negocio como parte del módulo de adquisición del Archivo Web de Venezuela, donde se puedan mostrar indicadores y métricas relacionados al proceso de preservación de las páginas web.

Se provee de un conjunto de métricas e indicadores para cada usuario previamente definido, ofreciéndole mayor información a éstos para la toma de decisiones de acuerdo a su rol dentro del sistema.

Se especifican los indicadores, así como también se implementan los procesos *ETL* y el almacén de datos. Adicionalmente a esto, se desarrolla la solución de Inteligencia de Negocio, utilizando Pentaho como aplicación de análisis de los datos, y se integra al módulo de adquisición del Prototipo de Archivo Web de Venezuela.

La principal limitante es la documentación escasa del formato WARC, debido a que es un formato nuevo y la información generada sobre éste es privativa. Es por esto que el Trabajo Especial de Grado se orienta a especificar los indicadores que puedan ser correctamente concebidos, basándose en la documentación encontrada de trabajos externos y/o generados por el Prototipo de Archivo Web.

Capítulo 2 Marco Conceptual

En este capítulo se explicarán los conceptos básicos para entender el contexto relacionado a este Trabajo Especial de Grado. Primero, se explicará lo que es la preservación web, incluyendo la definición de Archivo Web, los formatos existentes para su almacenado y la herramienta existente para la extracción de metadatos de esos formatos. Seguidamente, en la segunda parte, se explicarán los metadatos, incluyendo sus tipos. Tercero, se definirán los indicadores, con sus partes y tipos, y los indicadores definidos en el borrador de la ISO/DTR 14873:2013, creado en 2012. Como cuarto punto, se define Sistemas de Información y sus tipos existentes. Luego, se definirá Inteligencia de Negocio, sus características y funciones, y se explicará la arquitectura de una Solución de Inteligencia de Negocio. Por sexto y último punto, se explican las bases de datos *NoSQL*, sus características y tipos, así como el teorema de CAP y los pasos necesarios para transformar una base de datos a relacional a una base de datos *NoSQL* relacional.

2.1. Preservación Web

2.1.1. Patrimonio Digital

El patrimonio cultural se define como el conjunto de manifestaciones y objetos nacidos de la producción humana, que una sociedad ha recibido como herencia histórica, y que constituyen elementos significativos de su identidad como pueblo. (Llull P., 2005)

La UNESCO (2003) ha clasificado al patrimonio cultural como tangible, siendo éste los objetos arqueológicos, históricos, artísticos, etnográficos, tecnológicos, religiosos y aquellos de origen artesanal o folklórico, que constituyen colecciones importantes para las ciencias, la historia del arte y la conservación de la diversidad cultural del país, como los monumentos, conjuntos de construcciones y sitios con valor histórico, estético arqueológico, científico, etnológico y antropológico, o como patrimonio intangible, es decir, las expresiones y prácticas culturales, constituido por la poesía, los ritos, los modos de vida, la medicina tradicional, la religiosidad popular, las tecnologías tradicionales, entre otros elementos, de cada país.

Cuando se crea la *World Wide Web* (WWW), se ha conseguido una manera más sencilla de compartir parte de este patrimonio e infinidad de información actualizada, generada por libros o investigaciones, que se han sido desarrolladas por universidades, organizaciones y/o empresas. El problema de este, como se mencionó anteriormente, es la velocidad de cambio o desaparición de esta información, que podría ser útil para generaciones futuras.

Es por esto que nace la definición de patrimonio de origen digital, siendo éste los materiales informáticos, de valor perdurable, dignos de ser conservados por generaciones futuras, como puede ser publicaciones electrónicas, bases de datos, páginas web, entre otros. (UNESCO, 2003)

2.1.2. Definición de Preservación Web

Conforme con la UNESCO (2003), gran parte de la enorme cantidad de información que se produce en el mundo es de origen digital y existe en una gran variedad de formatos: textos, bases de datos, imágenes, audios, videos, programas informáticos, páginas web, entre otros. Para las instituciones culturales que tienen a su cargo la recolección y preservación del patrimonio cultural, definir qué elementos deben conservarse para las generaciones futuras y cómo proceder en su selección y conservación, se está volviendo un problema apremiante. La información digital producida hoy en día, en prácticamente todas las áreas de las actividades humanas, y concebida para ser consultada utilizando computadoras, podría perderse si no se elaboran técnicas y políticas específicas para su conservación.

Además de que los métodos tradicionales de preservación no pueden aplicarse tal cual al material digital, porque las "publicaciones" de la red comúnmente utilizan datos almacenados en varios servidores, ubicados en diferentes partes del mundo, existe un tipo de patrimonio digital que presenta características propias, como son las páginas web, que conlleva a idear nuevos métodos de preservación, denominados preservación web (Masanès, 2006). El patrimonio web comparte ciertas características con el resto del patrimonio digital, pero presenta algunas propias el alto volumen de datos, pues se estima que en Internet existen mil millones de páginas, su cambio constante, pues se ha estimado que su vida media es muy corta (Cho & Garcia-Molina, 2000), calculada entre cuarenta y cuatro (44) días a dos (2) años, su diversidad de formatos, entre otros. Es precisamente por estos problemas que se hace imperiosa la necesidad de preservar la información digital en formato web.

Cabe destacar que el patrimonio web posee dos (2) características fundamentales: una fuente única (servidor web) y un identificador único, el cual suele ser un Identificador Uniforme de Recursos (*URI*, por sus siglas en inglés). (Ospina, Modelo de Archivo de la Web para Venezuela, 2011)

Desde el punto de vista de la conservación, un recurso web tiene dos (2) características importantes:

- Depende de su única fuente para existir.
- Los servidores web pueden adaptar el contenido para cada instancia de recurso, haciéndolo diferente por cada solitud para la misma *URI*.

La Web, desde este punto de vista, no es un contenedor de archivos fijos, sino una caja negra con recursos, de los cuales el usuario sólo recibe instancias. (Ospina, 2011)

Por la creciente necesidad de preservar los recursos de la web, se crea el Consorcio Internacional de Preservación de Internet (IIPC) con el objetivo de adquirir, preservar y permitir la accesibilidad al conocimiento y la información, que se encuentra en el Internet, a futuras generaciones en cualquier parte del mundo. Conformado por las Bibliotecas, universitarias o nacionales, de 45 países, la IIPC se dedica a mejorar las herramientas, estándares y mejores prácticas para el archivado web, además de promover la colaboración internacional entre sus miembros, y el acceso y uso de los Archivos Web para investigaciones y patrimonio cultural (IIPC, 2012).

El Consorcio Internacional de Preservación de Internet (IIPC, 2012) afirma:

> *El internet ha dado lugar a una era sin precedente, donde se comparte el conocimiento, la creatividad, la innovación y la conexión. Como consecuencia, las instituciones dedicadas a la preservación y documentación del conocimiento y la cultura contemporánea se enfrentan a nuevos retos". Muchas cosas que las instituciones dedicadas a la preservación digital recolectan, son actualmente accesibles desde la web, como las publicaciones académicas, los materiales de campañas, las obras de arte, los documentos gubernamentales, correspondencia y noticias. Las páginas son cada vez más dinámicas, su contenido cambiando constantemente, por lo que es importante capturar esta información en tiempo real y asegurar su preservación para las próximas generaciones*.

Para determinar qué debe cumplir un sistema que asegure la preservación web, la UNESCO ha determinado que, la preservación digital (por tanto la preservación web) debe tener las siguientes características:

- Ser accesible para cualquier persona.
- Garantizar la protección de información delicada o de carácter privado.
- Disponer de un marco jurídico y técnico que proteja su autenticidad.
- No debe estar sujeto a límites temporales, geográficos, culturales o de formato. Se debe propiciar, con el tiempo, una representación de todos los pueblos, naciones, culturas e idiomas.

Además, la IIPC ha definido cuatro (4) actividades para la preservación web, las cuales se pueden visualizar en la Figura 2.

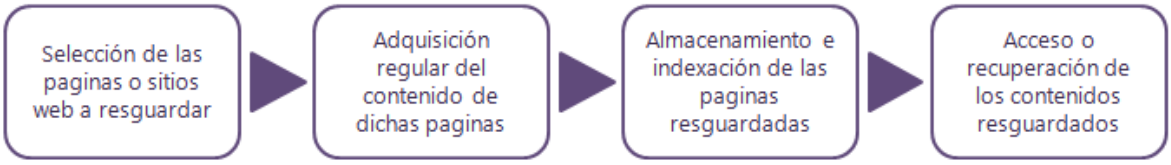


Figura 2 - Actividades para la Preservación Web
Fuente: Desarrollo de una Aplicación para Acceder a Contenidos de un Archivo Web en Formato WARC (Ospina, Martinez, Kabchi, & León, 2014)

A continuación, se presenta una breve descripción de estas actividades:

- La selección permite limitar el ámbito del archivo, pudiendo preservar contenidos locales o de un tipo en particular, como por ejemplo, contenidos de un país o educativos solamente. (Ospina, Martinez, Kabchi, & León, 2014)
- La adquisición logra que se puedan almacenar los cambios generados sobre los contenidos que se preservan a través del tiempo. (Ospina, Martinez, Kabchi, & León, 2014)
- El almacenamiento requiere estrategias que permitan preservar grandes volúmenes de información (del orden de los *Terabytes*), millones de archivos y diferentes formatos. Para este fin, se han desarrollado formatos de archivos contenedores específicos, cuyo objetivo principal es

superar la limitación de los sistemas de archivos propios de los sistemas operativos donde se alojan los Archivos Web. (Ospina, Martinez, Kabchi, & León, 2014)

- El acceso o recuperación de los contenidos está estrechamente ligado a la forma en que se encuentran almacenados, pero debido a la naturaleza hipertextual y multimedia de la web, se espera que el usuario final pueda acceder a este contenido de manera similar a cuando lo hace en los servidores originales. (Ospina, Martinez, Kabchi, & León, 2014)

2.1.3. Archivo Web

Los Archivos Web son sistemas de información (ver la sección 2.4 donde se define Sistemas de Información) que surgen para archivar, de manera histórica, documentos que están publicados en la Web, considerados parte del patrimonio digital de las naciones. Se puede definir un documento web como un documento basado en el lenguaje de marcas *HTML*[1] (*HyperText Markup Lenguage*), también llamado página web, y los demás archivos asociados en su composición como imágenes, videos, hojas de estilo, *scripts*, entre otros, que puede ser localizado a través de un *URL* y que, normalmente, forma parte de un sitio web. (Ospina Torres, 2014)

Los Archivos Web tienen como objetivo preservar conjuntos seleccionados de páginas o sitios web, y sus documentos mediante su replicación y/o migración de su formato original a otra representación. Los sitios replicados son mantenidos completos, es decir, acompañados de los archivos, imágenes, gráficas y aspecto visual, y son almacenados en servidores de preservación en un ambiente seguro. (Masanès, 2006)

Archivo Web recolecta fragmentos de la *World Wide Web* (WWW), preservando estas colecciones en un formato de archivo y ofreciendo estos archivos, para que puedan ser accedidos y utilizados por generaciones futuras. (IIPC, 2012)

Una forma de almacenar los contenidos dentro de un Archivo Web es el Formato WARC, el cual será descrito a continuación.

2.1.3.1. Formato ARC

Según Burner & Kahle (1996), ARC (*The ARChive*, en inglés) es un formato de archivado creado por *System Enhancement Associates* (SEA), el cual recoge la data en largos archivos agregados, actualmente de 100MB, para facilitar el almacenamiento en un sistema de archivos convencional.

Cuando el formato ARC fue creado, fue diseñado para cumplir los siguientes requisitos:

- Ser un archivo auto contenido, es decir, debe permitir que los objetos agregados sean identificados y desempaquetados sin la necesidad de utilizar un archivo de índices adicional.
- El formato debe ser extensible para alojar los archivos recuperados utilizando diversos protocolos de red, incluyendo http y ftp.

[1] Lenguaje de marcado estándar utilizado para crear páginas web.

- El archivo debe permitir concatenar múltiples archivos en un flujo de datos.
- Una vez escrito, el registro debe ser viable, es decir, la integridad del archivo no debe depender en la creación subsecuente de un índice de contenidos.

Figura 3 - Formato Archivo ARC
Fuente: (Burner & Kahle, 1996)

En el caso de la preservación web, el formato ARC ha sido, tradicionalmente, utilizado para almacenar los rastreos web en secuencias de bloques de contenido recogidos de la web. Cada captura en un archivo ARC es precedida por una cabecera de una línea que, muy brevemente, describe el contenido recogido y su longitud, y es seguido directamente por los mensajes de respuesta del protocolo de recuperación y su contenido, como se puede apreciar en la Figura 3. El formato de archivo ARC original ha sido utilizado por el *Internet Archive*, desde 1996, para la gestión de billones de objetos, y por varias bibliotecas nacionales.

La motivación para entender el formato ARC, surgió por la discusión y las experiencias del Consorcio Internacional de la Preservación del Internet (IIPC), entre cuyos miembros se incluyen las bibliotecas nacionales de Australia, Canadá, Dinamarca, Finlandia, Francia, Islandia, Italia, Noruega, Suecia, La Biblioteca Británica (Reino Unido), La biblioteca del Congreso (Estados Unidos de América) y el *Internet Archive*. La biblioteca Digital de California y el Laboratorio Nacional de Los Álamos también aportaron en la extensión y generalización del formato.

2.1.3.2. Formato WARC

Según la ISO 28500 (2009), el formato WARC (*Web ARChive*) es un contenedor de archivos, extensión del formato de archivo ARC, que permite concatenar múltiples registros de recursos (objetos de datos), cada uno compuesto de un set de cabeceras de texto simple y un bloque de datos arbitrario en un archivo largo.

En la Figura 4, se muestra el formato de un registro WARC perteneciente a un archivo WARC. Existen diversos Tipos de registros WARC, que son explicados más adelante, pero cabe destacar que todos tienen el mismo formato y, en la cabecera, poseen los mismos campos obligatorios.

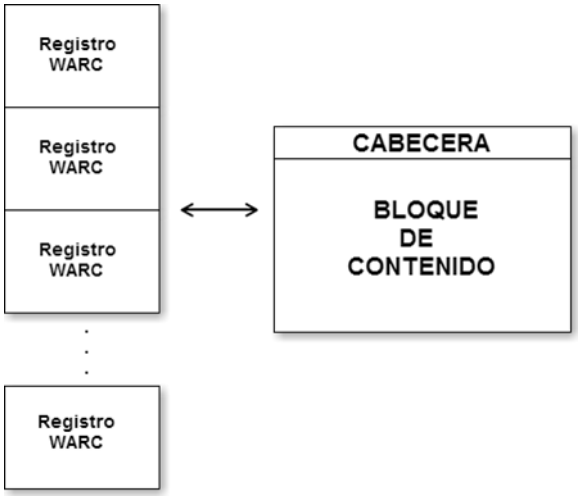


Figura 4 - Formato Registro WARC
Fuente: (ISO 28500, 2009)

El formato WARC es un estándar de estructura, que permite gestionar y almacenar billones de recursos recolectados de la Web, o de cualquier otra fuente. Éste es utilizado para construir aplicaciones que obtienen, gestionan y permiten el acceso e intercambio de contenido, tales como el rastreador web, de código abierto, *Heritrix*. La manera en que los archivos WARC son creados, y los recursos almacenados y prestados, depende de la implementación del software y las aplicaciones.

Además del contenido primario registrado en los ARCs, el formato extendido WARC aloja contenido secundario relacionado, tal como metadatos asignados, duplicaciones abreviadas de detección de eventos, transformaciones posteriores, y segmentación de recursos extensos. La extensión, también, puede ser útil para aplicaciones más generales que el archivado web. Para ayudar el desarrollo de herramientas que sean compatibles con versiones anteriores, el contenido WARC es claramente distinguible del contenido ARC.

El formato de archivo WARC está hecho lo suficientemente diferente del formato de archivo ARC, así las herramientas de software pueden detectar inequívocamente y procesar correctamente ambos tipos de registros, los ARCs y los WARCs. Dada la gran cantidad de datos de archivo existente en el formato ARC, es importante que el acceso y uso de éstos no sea interrumpido durante la transición al formato WARC.

El formato WARC especifica un método para la combinación de múltiples recursos digitales en un archivo de archivos añadido, junto con información relacionada. Los recursos son anticuados, identificados por los *URIs* y precedidos de cabeceras de texto simple. Por convención, los archivos con este formato se nombran con extensión ".warc" y tienen la extensión *MIME* definida como aplicación/warc.

Características

Entre las principales características de los WARCs, se tienen las mencionadas a continuación (Ospina Torres, 2014):

- Guarda contenido junto a la información de los protocolos de cosecha.
- Guarda metadatos enlazados a otros datos.
- Permite compresión de datos y preservar integridad de los registros.
- Permite manejos de registros de gran tamaño.
- Detecta duplicados y transformaciones posteriores de un archivo.

Objetivos

Según la ISO 28500 (2009), el formato WARC tiene los siguientes objetivos:

- Capacidad para almacenar tanto el contenido de la carga útil, como la información de control de los principales protocolos de la capa de aplicaciones de *Internet*, incluyendo HTTP, FTP, NNTP y SMTP.
- Capacidad para almacenar metadatos ligados a otros datos almacenados, como por ejemplo el clasificador de tema, idioma descubierto, codificación.
- Soporte para la compresión y el mantenimiento de la integridad de datos.
- Capacidad para almacenar toda la información de control del protocolo de adquisición (por ejemplo, solicitar cabeceras), y no sólo respuesta de información.
- Capacidad para almacenar los resultados de las transformaciones de datos vinculados a otros datos almacenados.
- Capacidad para almacenar un evento de detección de duplicados vinculado a otros datos almacenados, para reducir el almacenamiento en la presencia de recursos idénticos o sustancialmente similares.
- Ser suficientemente diferente del formato ARC, para que las herramientas de software pueden detectar y procesar los registros, tanto WARC y ARC, correctamente y sin ambigüedades. Dada la gran cantidad de datos de los archivos existentes en el formato ARC anterior, es importante que el acceso y uso de este legado no se interrumpa durante la transición al formato WARC.
- Capacidad para almacenar identificadores de registro únicos globales.
- Soporte para el manejo determinista de registros largos, por ejemplo, el truncamiento o segmentación.

Tipos de registros WARC

Hay ocho (8) tipos de registro WARC actualmente definidos, *'warcinfo'*, *'response'*, *'resourse'*, *'request'*, *'metadata'*, *'revisit'*, *'conversion'*, y *'continuation'*. El propósito y uso de cada tipo se describen a continuación.

- *warcinfo*: describe los registros que le preceden hasta el fin de archivo, fin de entrada de datos o hasta otro registro del mismo tipo. Normalmente, aparece una sola vez y al principio del archivo WARC. Para un Archivo Web, usualmente contiene la descripción de un rastreo web, es decir,

puede contener información como la duración, la profundidad y/o el propósito. El formato de la descripción puede contener datos como el tamaño máximo del archivo o la tasa de rastreo.

- *response*: contiene la respuesta completa de un protocolo, tal como una respuesta completa HTTP, incluyendo cabeceras y contenido-cuerpo, de una recuperación de Internet. Usualmente, la carga útil de estas respuestas reflejan el objetivo principal de la colección del servicio de archivado, cuya responsabilidad es la de distinguir la carga útil de las cabeceras de protocolo durante el procesamiento subsecuente. Un registro response, normalmente, incluye los parámetros llamados "IP-Address" (Dirección-IP) y "Related-Record-ID" (ID-Registro-Relacionado).
- *resource*: contiene un recurso sin toda la información de respuesta del protocolo. Por ejemplo: un archivo directamente recuperado de un repositorio de acceso local, o el resultado de una recuperación en red, donde la información del protocolo ha sido descartada. Un registro "resource" suele incluir el parámetro llamado "Related-Record-ID" (ID-Registro-Relacionado).
- *request*: contiene la manera en la cual el contenido de un registro primario fue solicitado, por ejemplo, en el contexto de rastreo en la web, contendrá la solicitud HTTP. Un registro de este tipo incluye el parámetro llamado "Related-Record-ID" (ID-Registro-Relacionado).
- *metadata*: posee el contenido creado para alguna futura descripción, explicación o acompañar algún contenido recogido, que no haya sido cubierto por otro tipo de registro de ninguna manera. Un registro "metadata" casi siempre se referirá a un registro de otro tipo, teniendo este otro registro el contenido original, o transformado, que había sido recolectado. Sin embargo, está permitido que un registro "metadata" esté referenciando a cualquier otro tipo de registro, incluyendo otro de su mismo tipo, o de no referenciar a otro registro en general. Pueden ser creados cualquier cantidad de registros "metadata" que referencien a otro registro en específico. Los formatos potenciales de este tipo de registro son el [ANVL], [RDF] o cualquier otro formato basado en XML. Un registro "metadata" suele incluir el parámetro "Related-Record-ID" (ID-Registro-Relacionado).
- *revisit*: describe la "revisitación" de un contenido ya archivado, e incluye solo un bloque de contenido abreviado, el cual debe ser interpretado en relación a un registro anterior. Típicamente, un registro "revisit" suele ser utilizado en sustitución a uno "response" o "resource", para indicar que el contenido visitado era un duplicado, parcial o completo, de algún material previamente archivado. Un registro "revisit" debería ser utilizado únicamente cuando, al interpretar un registro, se requiere consultar un registro previo; otros tipos de registros deberían ser preferidos si el registro actual es entendible por sí solo. No se requiere que algún "revisit" de un URI previamente visto utilice un "revisit", solo aquellos que se refieran a otros registros. El formato de un bloque de contenido, de este tipo de registro, puede variar para alcanzar diferentes objetivos, tal como almacenar la aparente magnitud de la diferencia de la visita anterior, o codificar el contenido visitado como un "diff" del contenido previamente almacenado. El propósito de este tipo de registro es el de reducir la redundancia al almacenar, cuando se recupera repetidas veces

el mismo contenido o similar, mientras aún se registra que ha ocurrido una revisitación, sumando detalles sobre el estado actual de un contenido visitado relacionado a la versión archivada. Un registro "revisit" requiere el parámetro llamado "Related-Record-ID" (ID-Registro-Relacionado).

- *conversion*: contiene una versión alternativa del contenido de otro registro que ha sido creado como resultado de un proceso de archivado. Normalmente, este es utilizado para que contengan las transformaciones del contenido que mantienen la viabilidad del contenido después de que desaparecen las herramientas disponibles para el formato original en el que fue almacenado dicho contenido. Según sea necesario, el contenido original puede ser migrado (transformado) a un formato más viable para mantener la información disponible con herramientas actuales, mientas que se minimiza la pérdida de información. Cualquier cantidad de registros de transformación podrán ser creados, referenciando a un registro de origen que puede contener, también, contenido transformado. Cada transformación debe resultar en un registro independiente y completo, sin dependencias al registro original. Los registros de tipo "metadata" pueden ser utilizados para descripciones adicionales de los registros de transformaciones. Un registro de conversión requiere el parámetro llamado "Related-Record-ID" (ID-Registro-Relacionado).
- *continuation*: necesita ser lógicamente anexado a un registro anterior, como otro archivo WARC, para crear lógicamente el registro completo. Este tipo de registro es utilizado cuando el archivo WARC va a exceder el límite deseado, éste es dividido en segmentos para mantener el tamaño del archivo. Un registro de continuación requiere los parámetros "Segment-Origin-ID" (ID-Segmento-Origen) y "Segment-Number" (Número-Segmento), y normalmente incluye el parámetro "Related-Record-ID" (ID-Registro-Relacionado).

2.1.3.3. Transformación del Archivo Web (*Web Archive Transformation*, WAT)

La especificación para la Transformación del Archivo Web (WAT) describe una forma estructurada para almacenar los metadatos generados por los rastreos web, y simplifica el análisis de grandes conjuntos de datos producidos por dichos rastreos. La estructura de los metadatos extraídos es optimizada para el análisis de los datos. La data de los WAT puede ser utilizada para crear reportes del análisis de los datos, eficientemente, basados en extensos conjuntos de datos. (Stern, 2011)

Las utilidades WAT (*WAT Utilities*, en inglés) son utilizadas para extraer metadatos de los archivos WARC. Dichas utilidades extraen metadatos de los archivos WARC y estructuran los metadatos utilizando un formato optimizado, que puede ser analizado en un ambiente de procesamiento distribuido. WAT estructura los datos utilizando *JavaScript Object Notation* (JSON).

Los programas de utilidades WAT están disponibles en la librería WAT, la cual está desarrollada en *JAVA*[2] por *Internet Archive*. Ellos producen metadatos, estructurados y optimizados, para el análisis de los datos.

[2] Lenguaje de programación concurrente, orientado a objetos y específicamente diseñado para tener la menor cantidad de dependencias posible.

Las utilidades WAT generan datos en formato *JSON*, utilizando *STDOUT*, a partir de un archivo ARC o WARC, el cual puede o no estar comprimido (GZIP). El archivo ARC o WARC puede ser un archivo local, un archivo accesible por HTTP (http://) o un archivo accesible (hdfs://) del Sistema de Archivo *Hadoop* (HDFS). (Stern, 2011)

2.1.4. Modelo de Información

2.1.4.1. Modelo OAIS

El modelo de referencia para un Sistema de Información de Archivo Abierto (OAIS) es un modelo desarrollado por el Comité de Sistemas de Datos del Espacio (CCSDS, por sus siglas en inglés), establecido en 1984, como un foro de las agencias espaciales de E.E.U.U. para el desarrollo cooperativo de estándares que apoyen el manejo de datos de las investigaciones espaciales. Surge como parte de una iniciativa para desarrollar normas que apoyaran la conservación, a largo plazo, de los datos obtenidos de los satélites y otros tipos de misiones espaciales.

OAIS se desarrolla como un modelo general, aplicable a otros contextos de preservación digital (Consultative Committee for Space Data Systems, 2012), y tiene como objetivo servir de guía para ayudar a entender los desafíos de los Archivos que se relacionan a objetos de información digital, proporcionando un lenguaje común de alto nivel que pueda facilitar la discusión a través de las comunidades interesadas en la preservación digital, por lo que la UNESCO recomienda su uso para el diseño de Archivos Digitales (UNESCO, 2003).

En la conceptualización de este tipo de archivos, el modelo OAIS define los elementos listados a continuación:

Ambiente OAIS

En la Figura 5, se muestra la interacción del Archivo con agentes externos. El modelo de referencia OAIS identifica y describe estas entidades externas, y caracteriza las interfaces entre dichas entidades y el Archivo (Ospina Torres, 2014). A continuación, se describen los elementos que forman parte del ambiente OAIS:

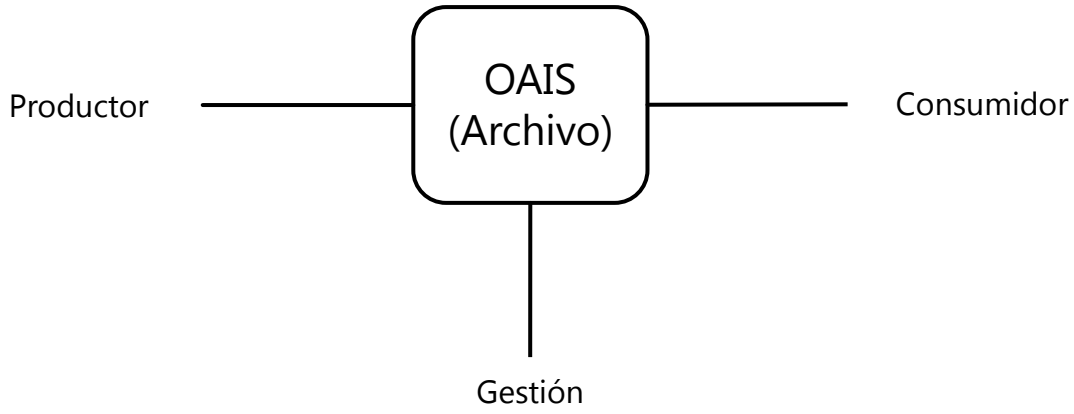


Figura 5 - Ambiente OAIS
Fuente: (Lavoie, 2004)

- Productor: rol desempeñado por las personas o los sistemas cliente, que proporcionan la información a ser conservada.
- Consumidor: rol desempeñado por las personas o los sistemas cliente, que interactúan con los servidores de OAIS para encontrar y adquirir información conservada de interés.
- Gestión: responsable del manejo de los componentes funcionales y las políticas del Archivo, así como del día a día de las operaciones del archivo. (Ospina Torres, 2014)

Modelo funcional

Describe la gama de actividades que deben llevarse a cabo por un Archivo. Define seis capas de servicio de alto nivel o componentes funcionales, como se puede ver en la Figura 6, que en conjunto cumplen el rol doble del OAIS: preservar y proveer acceso a la información custodiada. (Ospina Torres, 2014)

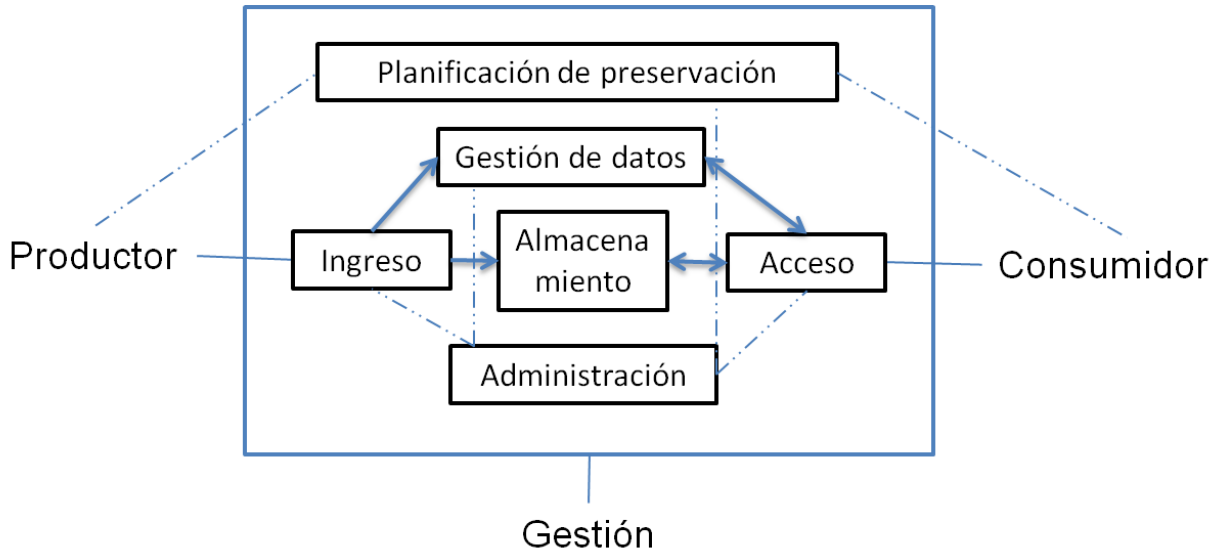


Figura 6 - Modelo funcional OAIS
Fuente: (Lavoie, 2004)

- Ingreso: acepta presentaciones de los productores y los prepara para el almacenamiento y la gestión dentro del archivo.
- Almacenamiento: para el almacenamiento, mantenimiento y recuperación de contenido del Archivo.
- Gestión de datos: maneja repositorios de los metadatos, que describen la información almacenada en el Archivo, realiza la gestión de información sobre el Archivo y sus propiedades.
- Administración: maneja las operaciones cotidianas del Archivo y coordina las actividades de los demás componentes funcionales del OAIS.
- Planificación de preservación: vigilancia del medio ambiente de la OAIS para garantizar la conservación a largo plazo de los contenidos del Archivo.
- Acceso: el apoyo a los consumidores (usuarios) en la búsqueda y recuperación de contenidos del Archivo. (Ospina Torres, 2014)

Modelo de información

Provee una descripción de alto nivel de los objeto de información manejados por el archivo. El modelo de información OAIS está construido alrededor del concepto de paquete de información (IP por sus siglas en inglés, *Information Package*) y de la manera en que esta estructura se transforma y mueve a través del Archivo. (Ospina Torres, 2014)

A medida que los IP son adquiridos y procesados por el Archivo pueden sufrir transformaciones a medida que fluye entre los componentes funcionales, por lo que el modelo define tres tipos de IP:

- El SIP (*Submit Information Package*): paquete de información recibido por el producto.
- El AIP (*Archive Information Package*): paquete de información almacenado en el Archivo.
- El DIP (*Delivery Information Package*): paquete de información entregado al cliente.

En la Figura 7, se muestra como los diferentes tipos de Paquetes de Información interactúan con los componentes de la arquitectura.

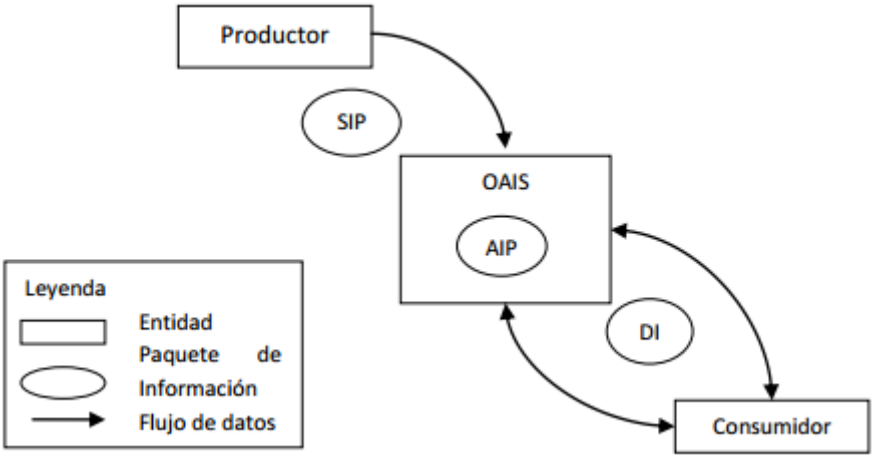


Figura 7 - Flujo de Datos en un *OAIS*
Fuente: (CCSDS, 2002)

2.1.4.2. Adaptación del modelo OAIS a Archivos Web

En 2006 (Masanès), se hace una adaptación del modelo funcional *OAIS*, donde define las tareas necesarias para que los Archivos Web puedan cumplir su objetivo de preservación. Dichas tareas se listan a continuación:

- Selección de las páginas web a resguardar.
- Adquisición regular del contenido de dichas páginas.
- Almacenamiento e indexación de las páginas resguardadas.
- Recuperación o consultas sobre la información resguardada.

En la Figura 8, se muestra la arquitectura funcional propuesta por el IIPC para Archivos Web y basada en el modelo OAIS. Aquí se agrupan una serie de herramientas que dan soporte a las tareas previamente definidas, de acuerdo al rol que tienen dentro del Archivo Web.

- Las herramientas de ingreso de datos (*Data ingest tools*) se encargan de la adquisición regular de los sitios web seleccionados y sus documentos web asociados. (Masanès, 2006)
- El proceso de selección, aunque forma parte de las políticas y lineamientos de cada Archivo, es soportado por las herramientas de ingreso a través de un repositorio de los sitios a preservar y la frecuencia de cambio para la preservación de nuevas versiones. (Masanès, 2006)
- Para la indexación de los documentos se hace uso de la tecnología de motores de búsqueda e indexación (*Index & Search engine*), desarrolladas por los buscadores de internet como Google, Altavista y *Yahoo*. (Masanès, 2006)
- El almacenamiento es llevado a cabo por las herramientas de almacenamiento (*Storage tools*) y los manejadores de contenido (*content management*). (Masanès, 2006)

- El acceso a los datos se hace a través de las herramientas de acceso (*Access tools*). (Masanès, 2006)

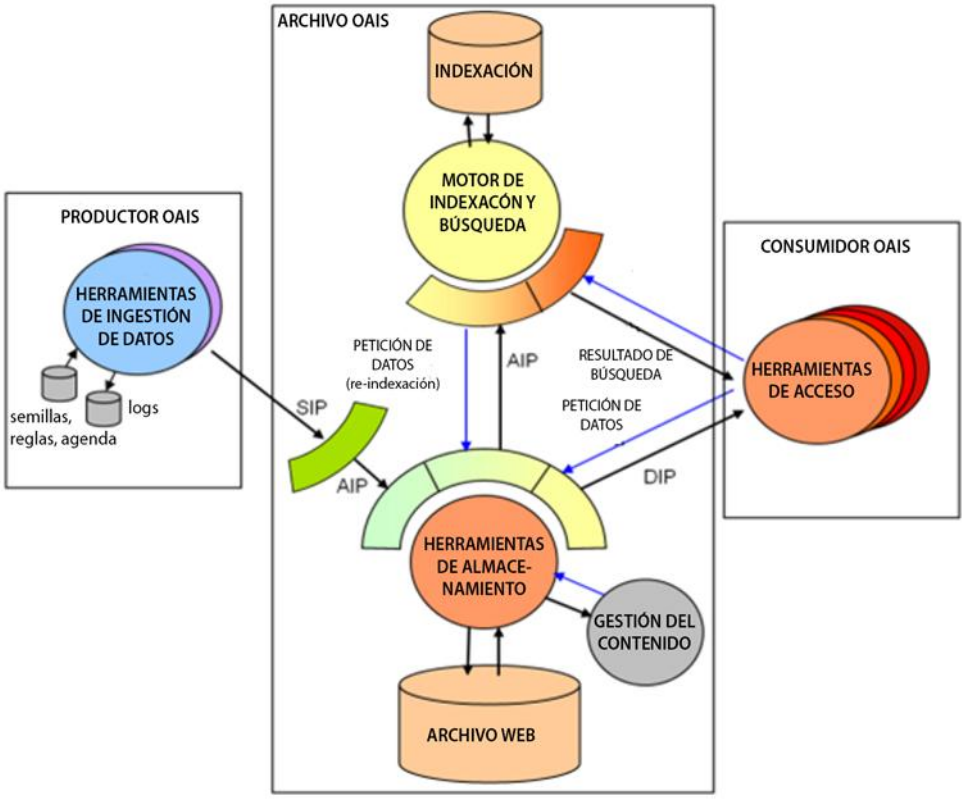


Figura 8 - Arquitectura Funcional IIPC, basada en modelo *OAIS* (Masanès, 2006)
Fuente: (Ospina Torres, 2014)

Modelo de Información

A continuación, se presentan los elementos de información en un Archivo Web, junto con la relación entre éstos y el modelo de información *OAIS* (sección 0), y el proceso de transformación por las actividades del Archivo Web. Dichos elementos de información se detallan en la Tabla 2.

Tabla 2 - Elementos de Información de un Archivo Web

Elementos de Información	**Descripción**	**Modelo de Información – Modelo OAIS**
Semillas, frecuencia de rastreo	Información que permite adquirir de manera regular el contenido a almacenar	SIP
Archivo de configuración	Archivo usado por el rastreador para realizar la adquisición	
Datos de localización geográfica del URL	Datos que permiten conocer la ubicación geográfica y el IP del sitio web rastreado	
Archivos WARC	Paquetes de información que almacenan los	AIP

Tabla 2 - Elementos de Información de un Archivo Web

Elementos de Información	**Descripción**	**Modelo de Información – Modelo OAIS**
	contenidos preservados por sitio web	
Colecciones	Agrupaciones de URL de sitios web en categorías o temas que dependen del contenido web del sitio web	Grupos de AIP
Log de Heritrix, Registros WARC descriptivos	Información descriptiva de la preservación	Metadata
Documentos web dentro de WARCS, que conforman los sitios web: • HTLM • CSS • Imágenes • Scripts Otros	Información a ser desplegada al usuario final	DIP

Fuente: (Ospina Torres, 2014)

2.2. Metadatos

2.2.1. Definición

Los metadatos son definidos, por el *National Information Standards Organization* (2004), como información estructurada que describe, explica, localiza o, de cierta manera, facilita la obtención, uso o manejo de algún recurso. Los metadatos pueden describir recursos en cualquier nivel de agregación. Pueden describir una colección, un solo recurso o un componente, parte de un recurso más grande.

El término metadatos es utilizado diferentemente entre las comunidades. Algunas lo utilizan para referirse a información entendible para las máquinas, mientras que otros lo usan solo para archivos que describen recursos electrónicos. En el contexto de las bibliotecas, los metadatos son comúnmente utilizados para cualquier esquema formal de descripción de recursos, aplicando a cualquier tipo de objeto, digital o no digital.

La motivación para crear metadatos descriptivos es para facilitar el descubrimiento de información relevante. También, los metadatos pueden ayudar a organizar recursos electrónicos, facilitar la interoperabilidad y la integración de recursos heredados (*legacy*, en inglés), proveen identificación digital, y soportan el archivado y preservación de dichos recursos.

2.2.2. Tipos de Metadatos

Aunque existen diversos tipos de metadatos (National Information Standards Organization, 2004), pero los tres (3) tipos principales son los siguientes:

- Metadatos descriptivos: con propósitos de descubrimiento e identificación, este tipo de metadatos describe un recurso, pudiendo incluir un resumen, autor y palabras claves.
- Metadatos estructurales: indica como los objetos compuestos están ensamblados, como el orden de las páginas dentro de un libro.
- Metadatos administrativos: provee información sobre el recurso para facilitar el manejo del mismo, tal como cuándo y cómo fue creado, tipo de archivo y otros detalles técnicos, y quien puede accederlo.

El tipo de metadatos utilizado en el prototipo son los metadatos administrativos. Adicionalmente, este tipo de metadatos se divide en dos (2) grandes subtipos:

- Metadatos para el manejo de derechos, el cual maneja los derechos de propiedad intelectual.
- Metadatos para la preservación, el cual contiene información requerida para archivar y preservar el recurso. Cabe destacar que éstos son los metadatos almacenados dentro del Prototipo de Archivo Web

2.2.3. Formato de la Transformación del Archivo Web (WAT)

Las Utilidades WAT, explicadas en la sección anterior, poseen un formato específico en el cual extraen los metadatos almacenados dentro de los WARCs. A continuación, haciendo referencia a la especificación de los archivos metadatos de los Archivos Web (Goel, 2011), se explicará el formato propuesto por ellos.

2.2.3.1. Cabecera del registro *Metadata*

La cabecera del registro consiste en la primera línea declarando el registro presente en el formato WARC con un número de versión dado, y seguido de un determinado número de líneas correspondiente a los campos explicados en la Tabla 3, y finalizando en un salto de línea.

Tabla 3 - Campos de la Cabecera del registro Metadata

Campo de la Cabecera	**Descripción**
WARC-Type	El tipo de registro WARC. Ajustado a "*metadata*"
WARC-Target-URI	El *URI* original correspondiente al contenido primario.
WARC-Date	Una marca de tiempo de 14 dígitos, que representa el instante de captura de los datos del contenido primario.
WARC-Record-ID	Un identificador asignado al registro actual, que es globalmente único para su período de uso previsto

Tabla 3 - Campos de la Cabecera del registro Metadata

Campo de la Cabecera	Descripción
WARC-Refers-To	Es el *WARC-Record-ID* del registro WARC primario siendo descrito.
Content-Type	Es el Tipo *MIME* de la información contenida en el bloque del registro *Metadata*. Ajustado a "*application/json*"
Content-Length	El número de octetos en el bloque del registro *metadata*.

Fuente: (Goel, 2011)

2.2.3.2. Bloque de contenido del registro Metadata

El bloque del registro *metadata* utiliza el formato "*application/json*" (*JSON* anidado) para describir los metadatos del registro WARC primario. Los metadatos son organizados en dos bloques, *Container* y *Envelope*, y todos sus campos son opcionales (ver Figura 9).

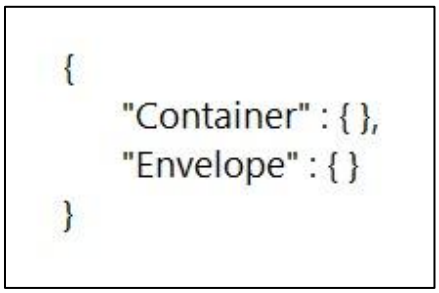


Figura 9 - Estructura de bloque de contenido
Fuente: (Goel, 2011)

Container

La primera parte del bloque de contenido viene dada en una estructura anidada de información, como se ve en la Figura 10, y contiene los campos descritos a continuación en la Tabla 4.

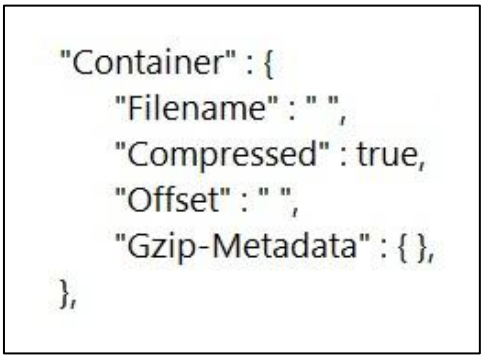


Figura 10 - Estructura del *Container* del Bloque de Contenido
Fuente: (Goel, 2011)

Tabla 4 - Campos iniciales del bloque de datos del registro Metadata

Campo Metadatos	Descripción

Tabla 4 - Campos iniciales del bloque de datos del registro Metadata

Campo Metadatos	**Descripción**
Filename	Nombre del archivo WARC donde se encuentra almacenado el registro
Compressed	Indica si el archivo está comprimido
Offset	Desplazamiento del archivo correspondiente al registro en el archivo
Digest	Parámetro que indica el nombre del algoritmo y el valor calculado de un resumen aplicado al archivo
Gzip-Metadata	El bloque Gzip de metadatos

Fuente: (Goel, 2011)

Adicionalmente, en la Tabla 5, se describen los campos pertenecientes a la estructura más interna del bloque de contenido, es decir, el *Gzip*.

Tabla 5 - Campos de la estructura interna, Gzip-Metadata, del *Container*

Campo Metadatos	**Descripción**
Header-Length	Indica el largo del octeto de la cabecera del Gzip
Footer-Length	Indica el tamaño del octeto del pie de página del Gzip. Ajustado a 8.
Deflate-Length	Indica el tamaño del octeto de todo el miembro Gzip desinflado, incluyendo los datos de a cabecera y el pie de página.
Inflated-CRC	Indica el CRC inflado
Inflated-Length	Indica el tamaño inflado
F-Extra	Campos y valores adicionales

Fuente: (Goel, 2011)

Envelope

La segunda y última parte del bloque de contenido, también se encuentran almacenados los metadatos en una estructura anidada, presentada en la Figura 11 y explicada en la Tabla 6.

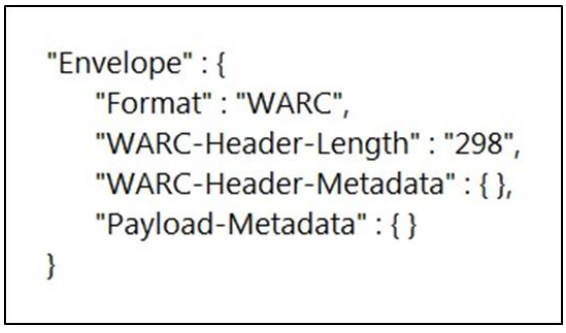


Figura 11 - Estructura del *Envelope* del Bloque de Contenido
Fuente: (Goel, 2011)

Tabla 6 - Campos iniciales del Envelope del bloque de datos del registro Metadata

Campo Metadatos	Descripción
Format	Indica si el registro es un registro ARC o WARC.
WARC-Header-Length	Número de octetos en la cabecera del registro
WARC-Header-Metadata	El bloque de metadatos de la cabecera
Payload-Metadata	La bloque de metadatos de carga útil

Fuente: (Goel, 2011)

Dentro de esta estructura, se encuentran anidadas las estructuras *WARC-Header-Metadata* y *Payload-Metadata* que se pueden observar en las Figura 12 y Figura 13, y se explican en las Tabla 7 y Tabla 8 respectivamente.

```
"WARC-Header-Metadata" : {
    "WARC-Type" : " ",
    "WARC-Record-ID" : " ",
    "WARC-Date" : " ",
    "Content-Length" : " ",
    "Content-Type" : " ",
    "WARC-Concurrent-To" : " ",
    "WARC-Block-Digest" : " ",
    "WARC-Payload-Digest" : " ",
    "WARC-IP-Address" : " ",
    "WARC-Refers-To" : " ",
    "WARC-Target-URI" : " ",
    "WARC-Truncated" : " ",
    "WARC-Warcinfo-ID" : " ",
    "WARC-Filename" : " ",
    "WARC-Profile" : " ",
    "WARC-Identified-Payload-Type" : " ",
    "WARC-Segment-Origin-ID" : " ",
    "WARC-Segment-Number" : " ",
    "WARC-Segment-Total-Length" : " "
}
```

Figura 12 - Campos de la estructura WARC-Header-Metadata del *Envelope*
Fuente: (Goel, 2011)

Tabla 7 - Descripción de los campos de la estructura WARC-Header-Metadata del Envelope

Campo Metadatos	Descripción
WARC-Type	Tipo del registro WARC
WARC-Record-ID	Identificador asignado al registro WARC
WARC-Date	Marca de tiempo (14 dígitos) que representa el instante en que fueron capturados los datos
Content-Length	Número de octetos en el bloque del registro (Declarado)
Content-Type	Tipo MIME de la información contenida en el bloque del registro

Tabla 7 - Descripción de los campos de la estructura WARC-Header-Metadata del Envelope

Campo Metadatos	**Descripción**
	(Decarado)
WARC-Concurrent-To	WARC-Record-ID de cualquiera de los registros creados como parte del mismo evento de captura del registro
WARC-Block-Digest	Parámetro que indica el nombre del algoritmo y valor calculado de un compendio aplicado a todo el bloque del registro
WARC-Payload-Digest	Parámetro que indica el nombre del algoritmo y el valor calculado de un compendio aplicado a la carga útil referida o contenida por el registro, la cual no es necesariamente equivalente al bloque del registro. La carga útil de un bloque *application/http* es su '*entity-body*'. En contraste con el campo anterior, éste también puede ser utilizado para los datos que no están necesariamente presente en el actual bloque del registro o cuando el registro se encuentra segmentado.
WARC-IP-Address	La dirección numérica del Internet contactada para obtener el contenido
WARC-Refers-To	WARC-Record-ID de un único registro para el cual el registro guarda contenido adicional
WARC-Target-URI	URI original cuya captura dio lugar a la información contenida en el registro
WARC-Truncated	Indica la "truncación" de un bloque de contenido con su motivo
WARC-Warcinfo-ID	WARC-Record-ID del registro '*warcinfo'* asociado
WARC-Filename	Nombre del archivo que contiene el registro '*warcinfo*'. Aplicable solo para registros '*warcinfo'*.
WARC-Profile	*URI* que representa el tipo de análisis y manejo aplicado en el registro '*revisit'*. Aplicable solo para registros '*revisit*'.
WARC-Identified-Paylos-Type	Tipo de contenido de la carga útil del registro según lo determinado por una verificación independiente
WARC-Segment-Origin-ID	Identifica el registro inicial en una serie de registros segmentados cuyos bloques de contenido son reensamblados para obtener un bloque de contenido lógicamente completo. Aplicable solo para registros '*continuation'*.
WARC-Segment-Number	Reporta de ordenamiento relativo de registros en una secuencia de registros segmentados.
WARC-Segment-Total-Length	Reporta el tamaño total de todos los bloques de contenido segmentados cuando han sido concatenados juntos. Aplicable solo para registros '*continuation'*.

Fuente: (Goel, 2011)

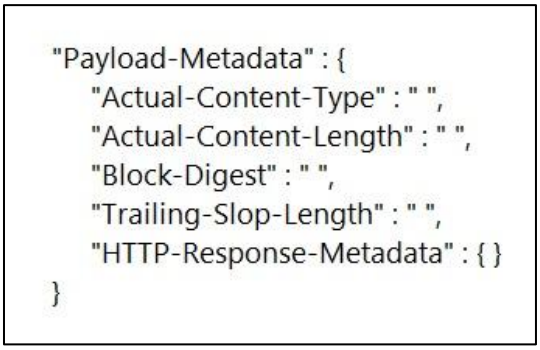


Figura 13 - Campos de la estructura Payload-Metadata del *Envelope*
Fuente: (Goel, 2011)

Tabla 8 - Descripción de los campos de la estructura Payload-Metadata del *Envelope*

Campo Metadatos	**Descripción**
Actual-Content-Type	Tipo MIME de la información contenida en el bloque del registro según determina una verificación independiente (Actual/Detectado)
Actual-Content-Length	Número de octetos en el bloque del registro (Actual)
Block-Digest	Parámetro que indica el nombre del algoritmo y el valor calculado de un compendio aplicado a todo el bloque del registro.
Trailing-Slop-Length	Número de bytes finales de decantación
Metadata	El bloque de metadatos específicos al tipo del registro (HTTP-Response-Metadata, DNS-Response-Metadata, etc.)

Fuente: (Goel, 2011)

2.3. Indicadores

Cuando se habla de indicadores, normalmente suele confundirse con métricas, las cuales son definidas como medidas numéricas que representan un pedazo de la data del negocio.

Sin embargo, según la Real Academia Española (2001), un indicador es algo que indica o sirve para indicar. Cuando introducimos esta definición en contexto, se puede definir que un indicador entonces es una métrica atada a un objetivo.

Formalmente, se pueden definir los indicadores clave de rendimiento (KPI, *Key Performance Indicator*) como datos que ayudan a medir, objetivamente, la evolución de un sistema, es decir, ayudan a evaluar hasta qué punto se están logrando los objetivos estratégicos. (AEC, 2013)

En las organizaciones, sin una retroalimentación regular y objetiva sobre el progreso en el alcance de los resultados que se quieren, éstas toman decisiones subjetivamente, suponiendo que dicha decisión es la adecuada sin un soporte real y objetivo que las fundamente. En cambio, con buenos indicadores de rendimiento, se pueden tomar las decisiones correctas en el momento indicado para alcanzar los objetivos y no desperdiciar tiempo, dinero o esfuerzo en el proceso.

Según la Asociación Española de Calidad (2013), los indicadores tienen como objetivo aportar a la empresa un camino correcto para que logre cumplir con las metas establecidas, teniendo que cumplir los siguientes objetivos al ser un sistema de medición de objetivos para la empresa:

- Comunicar la estrategia.
- Comunicar las metas.
- Identificar problemas y oportunidades.
- Diagnosticar problemas.
- Entender procesos.
- Definir responsabilidades.
- Mejorar el control de la empresa.
- Identificar iniciativas y acciones necesarias.
- Medir comportamientos.
- Facilitar la delegación en las personas.
- Integrar la compensación con la actuación.

A continuación, se describen las partes necesarias para definir un indicador correctamente:

2.3.1. Partes de un Indicador

A la hora de definir los indicadores, hay que fijar una serie de parámetros para cada uno de ellos. Según Bernal (2013) y Mondragón (2014) las partes esenciales que deben definirse junto al indicador son las siguientes:

- Definición: la identificación del indicador es primordial, debe ser concreto y definir claramente la función que cumple.
- Forma de cálculo / ratio: cuando se trata de indicadores cuantitativos, se debe tener muy claro la fórmula matemática para el cálculo de su valor, lo cual implica la identificación exacta de los factores o variables que lo conforman y la manera en la que ellos se relacionan.
- Unidades: la manera en la que se expresa el valor de determinado indicador está dada por las unidades, las cuales son seleccionadas en función a las necesidades de la organización.
- Periodicidad: se debe fijar cada cuánto tiempo se va a medir, mensualmente, trimestralmente, anualmente, semanalmente, diariamente, cada hora, instantáneamente, entre otras. Si el indicador es clave para el buen funcionamiento se deberá medir y controlar más frecuentemente que si es un indicador secundario menos importante.
- Proceso: la actividad o proceso que está asociado al indicador.
- Responsable: el departamento o persona que es responsable del proceso o la actividad que se está midiendo.

Ahora bien, sobre los resultados del indicador, debemos compararlos con un valor preestablecido: un objetivo, una expectativa y/o un límite.

- Objetivo: valor que queremos alcanzar. Este debe ser ambicioso, alcanzable, estar cuantificado y acotado en el tiempo.
- Expectativa: es el valor ideal del indicador, aunque no siempre es alcanzable.
- Límites legales: es el límite que nos impone la ley, y que no podemos propasar. Es diferente a los objetivos, porque el objetivo marca un propósito voluntario fijado por nosotros, y el límite legal es un valor que estamos obligados a cumplir.
- Límite de aceptabilidad: aparte de lo anterior, también se puede fijar un valor límite para considerar que el proceso funciona. Conociendo cuál es el funcionamiento normal del proceso, fijamos un valor, por debajo del cual asumiremos el proceso está funcionando mal y deberemos tomar acciones.

A continuación, otros puntos importantes a tomar en cuenta:

- Propósito del indicador: ¿Por qué medimos este dato? ¿Para qué sirve esta medición? Todos los indicadores deben tener un propósito lo suficientemente argumentado como para que lo que ganamos obteniendo ese dato sea más valioso que el tiempo que perdemos en medirlo.
- Grupos de interés: ¿A quién beneficia que estemos controlando el aspecto medido por el indicador? Pueden ser grupos de interés los clientes, los proveedores, los empleados, la dirección, los accionistas, el entorno, etc.
- Destinatarios: ¿Quién va a recibir y revisar los datos del indicador? Por lo general, los destinatarios suelen ser los responsables del proceso, los jefes de sección, y la dirección.
- Soporte: ¿En qué formato se va a almacenar? ¿Quién va a recopilar los datos? ¿Cómo se va a distribuir? Lo más común es almacenarlos en *Excel* o *PDF* y enviarlos a sus destinatarios por email, impresos o en una carpeta compartida.

Es por esto que se definen distintos tipos de indicadores, los cuales serán descritos en la siguiente sección.

2.3.2. Tipos de Indicadores

En teoría, se pueden establecer indicadores para cualquier aspecto medible, pero en el contexto de orientación a los procesos, podemos encontrar los siguientes tipos de indicadores (Beltrán, 2006):

- Indicadores de Procesos: están relacionados con el conjunto de actividades que forman parte del proceso.
- Indicadores de Resultados: se refiere al comportamiento del proceso como un todo.
- Indicadores de Eficacia: representan el cociente entre producción real y la esperada, independiente de los recursos utilizados para lograrlo.
- Indicadores de Eficiencia: el concepto de eficiencia se refiere al grado de cumplimiento de los objetivos planteados, sin garantizarlo, es decir, en qué medida la organización, está cumpliendo con sus objetivos tomando en cuenta los recursos con los que cuenta.

- Indicadores de Gestión: son medidas utilizadas para determinar el nivel de cumplimiento de los objetivos de una actividad perteneciente a un proyecto o del proyecto en sí. Los indicadores de gestión están relacionados con la administración de un proceso o actividad.

Adicionalmente, a pesar de la distinción existente entre los tipos de indicadores, éstos son generados por un motivo, el cual es explicado a continuación:

2.3.3. Importancia de los Indicadores

Mondragón (2014) señala cual es la importancia de los indicadores en los siguientes puntos:

- Elemento de planificación: durante los procesos de planificación se utilizan con frecuencia los indicadores para establecer la meta u horizonte a donde se quiere llegar.
- Estándar de seguimiento y control: el indicador ayuda a entender o muestra el estado del problema, ayuda a determinar la brecha entre lo planificado o esperado y el punto actual en el que se hace la valoración o medición. Un indicador es una señal de alerta que induce a reconocer que es necesario resolver un problema.
- Herramienta para la toma de decisiones: permite establecer métricas a través de las cuales se demuestre el cumplimiento de un objetivo o una meta en determinado proceso, proporcionando la información de apoyo para la toma de decisiones, así como también el planteamiento de políticas y estrategias para solucionar el problema.

2.3.4. Antecedentes de uso de indicadores en la preservación web

A medida que se van involucrando más instituciones en el archivado web, una necesidad mundial surgió por directrices en la administración y evaluación de los productos y actividades del Archivo Web. Es por esto que, en 2009, el Comité Técnico 46 de la *ISO* (la división de información y documentación) decidió crear un grupo que trabajara en "Estadísticas e Indicadores de Calidad para Archivos Web".

En Octubre de 2012, este comité entrega un borrador del informe técnico (ISO Working Group, 2012) para que sea evaluado y sometido a votación por los diferentes cuerpos pertenecientes a la ISO. Adicionalmente, hicieron público dicho borrador para obtener realimentación de una comunidad mayor, publicándolo en la página oficial del Consorcio Internacional de la Preservación del Internet (IIPC, por sus siglas en inglés).

Actualmente, en 2013, fue publicado el documento *ISO* oficial, bajo el nombre ISO/TR 14873:2013, haciendo énfasis en que este reporte técnico no avala ni recomienda el uso de alguna aplicación en específico, aunque el uso de algunos puede ocasionar variaciones en los resultados. Adicionalmente, también resaltan que dentro de dicho reporte técnico se enfocan en principios y métodos para el Archivado Web, pero no abarca opciones alternativas para recolectar recursos del Internet.

2.3.5. Datos Estadísticos e Indicadores de Preservación

En las tablas presentadas a continuación, se muestran todos los datos estadísticos e indicadores definidos para los Archivos Web, obtenidos del borrador del ISO/TR 14873:2013 (ISO Working Group, 2012), y que son clasificados de la siguiente manera:

- Datos estadísticos para el desarrollo de una colección.
- Datos estadísticos principales para la categorización de la colección.
- Datos estadísticos básicos sobre el uso del Archivo Web.
- Datos estadísticos principales para el uso de la colección.
- Datos estadísticos para la caracterización avanzada del uso del Archivo Web.
- Preservación del Archivo Web.
- Costo del Archivo Web.
- Indicadores de calidad.

2.3.5.1. Datos estadísticos principales para el desarrollo de una colección

Los datos estadísticos, mostrados en la Tabla 9, son los considerados como principales y que todo Archivo Web debe poder obtener, ya que describen el estado actual de éste.

Tabla 9 - Datos estadísticos principales para el desarrollo de una colección

Dato Estadístico	Propósito	Ejemplo
Cantidad/Número de objetivos[3]	Objetivos de colección / datos cuantitativos	8.000 objetivos
Cantidad/Número de capturas de objetivo[4]	Objetivos de colección / datos cuantitativos	14.000 rastreos de un objetivo
Tiempo de selección de objetivo[5]	Objetivos de colección / datos cuantitativos	2 horas
Número/Cantidad de URLs	Datos cuantitativos	14 miles de millones de URLs
Distribución de URLs por códigos de estatus[6]	Tipo / el número de recursos	2 millones de recursos rastreados con éxito
Cantidad de recursos recolectados	Número de recursos	5 millones de recursos rastreados
Número/Cantidad de dominios o hosts	Datos cuantitativos	3 millones de nombres de dominio
Tamaño en bytes (comprimido y descomprimido)	Datos cuantitativos	200 terabytes sin comprimir; 160 terabytes comprimidos

[3] Un objetivo es un conjunto de recursos a ser colectado y su alcance puede variar de recursos interrelacionados en el mismo dominio, presentado como una página web, a un único recurso. Cada rastreo es la captura de un objetivo. Para AWV son las semillas

[4] Rastreos realizados

[5] Tiempo utilizado por el suscriptor para determinar un objetivo

[6] Respuesta del servidor al solicitar un recurso

Tabla 9 - Datos estadísticos principales para el desarrollo de una colección

Dato Estadístico	Propósito	Ejemplo
Cantidad/Número de archivos WARC	Datos cuantitativos	18.000 archivos WARC

Fuente: (ISO Working Group, 2012)

2.3.5.2. Datos estadísticos principales para caracterización de la colección

Por otro lado, los datos estadísticos mostrados en la Tabla 10, son aquellos que nos permiten observar cómo se distribuyen los datos dentro del Archivo Web, incluyendo las características de los mismos.

Tabla 10 - Datos estadísticos principales para caracterización de la colección

Dato Estadístico	Propósito	Ejemplo	Comentario
Distribución por dominios de nivel superior (TLD) o segundo nivel[7]	Distribución geográfica	70% del Archivo alojado utilizando .fr (TLD)	
Distribución por volumen de recursos por dominio[8]	Análisis de dominio	3% de los dominios del Archivo contiene el 30% del total de URLs	
Distribución de URLs por dominio	Análisis de dominio	15% de los URLs son contenidos por www.nytimes.com	
Distribución por tipos de formatos[9]	Caracterización de formatos	60% de los recursos del Archivo son html/text	
Cobertura cronológica	Análisis temporal	El Archivo contiene recursos recolectados desde 1996 hasta hoy	
Almacenamiento cronológico	Análisis temporal	En 2012, se almacenaron 200 MB en rastreos.	
Idioma[10]	Caracterización de idioma	El 70% de los objetivos están en inglés.	Muchos recursos no poseen metadatos relacionados al idioma.
Cantidad/Número de permisos solicitados	Productividad	20 permisos han sido solicitados	Único de Archivos Web selectivos
Cantidad/Número de permisos concedidos[11]	Productividad	60% de los permisos solicitados han sido concedidos	Único de Archivos Web selectivos

[7] Los dominios de nivel superior indican la distribución geográfica de los recursos en un Archivo Web. Los dominios de segundo nivel indican la amplia naturaleza de los recursos en un Archivo.

[8] Útil para revelar las características de los recursos alojados en ciertos tipos de dominios.

[9] Tipos MIME.

[10] Ayuda a analizar una variedad de cuestiones sociales y culturales, y su reflejo en la Web.

[11] Parte legal del Archivo Web

Tabla 10 - Datos estadísticos principales para caracterización de la colección

Dato Estadístico	Propósito	Ejemplo	Comentario
Número/Cantidad de nominaciones[12]	Productividad	30% del Archivo es seleccionado manualmente	Único de Archivos Web selectivos
Temas	Caracterización de temas	75% de los objetivos del Archivo Web son académicos	Únicos de Archivos Web selectivos.

Fuente: (ISO Working Group, 2012)

2.3.5.3. Datos estadísticos básicos sobre el uso del Archivo Web

Adicionalmente, los datos estadísticos listados en la Tabla 11, son especialmente definidos para instituciones, como Bibliotecas Nacionales, que desean conocer a grosso modo el uso del Archivo Web por parte de los usuarios.

Tabla 11 - Datos estadísticos básicos sobre el uso del Archivo Web

Dato Estadístico	Tipo	Cálculo	Importancia
Vistas página	Cantidad	El número de veces que la página ha sido vista	Alta
Visitas (Sesiones)	Cantidad	Una visita es la interacción de un individuo con un sitio web, consistiendo en uno o más solicitudes a la página. Si un individuo no ha realizado otra acción (normalmente vistas a otras páginas) en el sitio en un periodo específico de tiempo, la visita será terminada por el exceso de tiempo inactivo.	Alta
Visitantes únicos	Cantidad	Número de individuos inferidos dentro de un plazo de informes designado, con actividades que consisten en uno o más visitas a un sitio. Cada individuo es contado solo una vez en la medida de visitante único para el periodo de reportes.	Media
Evento	Dimensión y/o Cantidad	Cualquier acción registrada que posea una fecha específica y un tiempo, asignado a éste por el servidor o el navegador.	Baja

Fuente: (ISO Working Group, 2012)

2.3.5.4. Datos estadísticos principales para el uso de la colección

En la Tabla 12, se muestran datos estadísticos que complementan los mostrados en la sección anterior.

Tabla 12 - Datos estadísticos principales para el uso de la colección

Dato Estadístico	Propósito	Ejemplo
Número de páginas vistas	Grado de uso	48.318 páginas en el Archivo Web de UK fueron vistas en Junio del 2012

[12] Cuando las semillas tienen que ser aprobadas por algún ente, luego de ser nominadas.

Tabla 12 - Datos estadísticos principales para el uso de la colección

Dato Estadístico	Propósito	Ejemplo
Número de visitas	Grado de uso	Hubo 11.415 visitas al Archivo Web de UK en Junio del 2012
Número de visitantes no-duplicados	Grado de uso	Hubo 9.434 visitantes únicos al Archivo Web de UK en Junio del 2012
Duración de visita	Utilidad/relevancia del archivo	En promedio, cada visita al Archivo Web de UK en Junio del 2012, duró 3 minutos y 25 segundos
Página vista por visita	Utilidad/relevancia del archivo	4,23 páginas fueron vistas por visita al Archivo Web de UK en Junio del 2012
Términos de búsqueda utilizados en un archivo web	Comportamiento del usuario	La palabra clave más utilizada para buscar el archivo web de UK, en Junio de 2012, fue "goji berry"

Fuente: (ISO Working Group, 2012)

2.3.5.5.Datos estadísticos para la caracterización avanzada del uso de un Archivo Web

Los datos estadísticos de la Tabla 13, son una extensión para aquellas instituciones que deseen evaluar más a fondo el uso de su Archivo Web.

Tabla 13 - Datos estadísticos para la caracterización avanzada del uso de un Archivo Web

Dato Estadístico	Tipo	Cálculo	Importancia
Página de entrada	Dimensión	Primera página de una visita.	Media
Página destino	Dimensión	Vista de una página con la intención de identificar el inicio de la experiencia del usuario, como resultado de un esfuerzo de comercialización definido.	Baja
Página de salida	Dimensión	Última página de un sitio, accedida en una visita, significando el final de una visita/sesión.	Baja
Duración visita	Cantidad	Tiempo de una sesión. El cálculo suele ser la marca de tiempo de la última actividad en la sesión menos la marca de tiempo de la primera actividad de la sesión	Alta – Indicador de cuán útil es el archivo web para el usuario
Referente	Dimensión	Referente es un término genérico que describe la fuente del tráfico a una página o visita.	Media
Página referente	Dimensión	Página referente describe la fuente de tráfico de una página.	Media
Visitante nuevo	Cantidad	Número de visitantes únicos con actividades que incluyen la primera visita a un sitio en un tiempo determinado. Nótese que "primera visita" es	Baja

Tabla 13 - Datos estadísticos para la caracterización avanzada del uso de un Archivo Web

Dato Estadístico	Tipo	Cálculo	Importancia
		respecto a cuándo los datos empezaron a ser recolectados apropiadamente, utilizando la actual herramienta.	
Visitante recurrente	Cantidad	Número de visitantes únicos con actividades que consisten en una visita a un sitio durante un periodo de tiempo y donde el visitante también visitó el sitio antes de dicho período.	Media
Visitante repite	Cantidad	Número de visitantes únicos con actividad que consiste de una o más visitas a un sitio en un periodo de tiempo	Media
Visita de una página	Dimensión o cantidad	Una visita que consiste en la visita a una sola página	Media
Porcentaje de abandono	Proporción	Visitas a una sola página dividida por páginas de entrada	Media
Páginas vistas por visita	Proporción	Número de páginas vistas en un reporte periódico, dividido por el número de visitas en el mismo periodo de tiempo	Alta – Indicador de la utilidad del archivo para el usuario
Visitantes (Localización geográfica)	Cantidad	Reporte sobre Geo-IP de la dirección IP del solicitante	Baja
Términos de búsqueda para encontrar el Archivo Web	Cantidad	Términos utilizados en los motores de búsqueda para encontrar el sitio web de la herramienta de acceso del Archivo Web	Media
Términos de búsqueda utilizados en el Archivo Web	Cantidad	Términos de búsqueda utilizados en la herramienta de acceso para encontrar capturas archivadas	Alta – Indicador de "temas actuales" al buscar rastreos

Fuente: (ISO Working Group, 2012)

2.3.5.6. Preservación del Archivo Web

Para planificar y gestionar la preservación de los datos del archivo Web, los siguientes datos estadísticos son de ayuda para conocer el estado actual del archivo Web. Éstos están clasificados en el flujo de bits (*bit-stream*, Tabla 14), preservación de los metadatos (Tabla 15) y la preservación lógica del Archivo Web (Tabla 16).

Tabla 14 - Datos estadísticos para la preservación del *bit-stream*

Dato Estadístico	Propósito	Ejemplo	Comentario

Tabla 14 - Datos estadísticos para la preservación del *bit-stream*

Dato Estadístico	Propósito	Ejemplo	Comentario
Cantidad de datos perdidos o deteriorados	Seguridad y resistencia	Se han perdido 25 MB de datos. Se han deteriorado 50 *URLs*.	Obtenido por la comparación de *checksums* regulares.
Volumen de recursos replicados	Seguridad y resistencia	150 *terabytes* del Archivo Web están replicados	

Fuente: (ISO Working Group, 2012)

Tabla 15 - Datos estadísticos relacionados a la preservación de los metadatos

Tipo de Metadatos	Descripción	Estándar utilizado (si aplica)	Porcentaje de recursos que contienen los metadatos	Comentario
Descriptiva	Metadatos DCMI	*Dublin Core Metadata* (DCMI) LCSH	30%	Término del material asignado manualmente por curadores y almacenada en una herramienta de conservación web
Procedencia	Archivos de configuración		90%	Archivos de configuración de los rastreos del 2004 fueron descartados
Técnica	Formatos de archivos (Tipo *MIME*)	*Multipurpose Internet Mail Extension (MIME) Part Two: Media Types*	100%	Todos los archivos cosechados tienen información tipo *MIME*, pero esto puede no ser confiable
Derechos	Permisos para archivar y proveer acceso en línea		100%	Requerido solamente por objetivos con acceso abierto

Fuente: (ISO Working Group, 2012)

Tabla 16 - Datos estadísticos para la preservación lógica del Archivo Web

Dato Estadístico	Propósito	Ejemplo
Distribución por formatos de archivos identificados	Capacidad de preservación	60% del archivo se encuentra en formato HTML
Número de formatos por los cuales una estrategia de preservación fue definida	Capacidad de preservación y compromiso	Cinco (5) formatos han definido la estrategia de preservación: HTML, JPEG, GIF, PNG y PDF.

Fuente: (ISO Working Group, 2012)

2.3.5.7. Costo del Archivo Web

Al medir el costo asociado al Archivo Web, es necesario establecer seis (6) categorías por las cuales se pueden clasificar los gastos generados por las actividades de archivado web. Estas categorías se explican en la Tabla 17, presentada a continuación:

Tabla 17 - Datos estadísticos asociados al Costo del Archivo Web

Categoría	Descripción	Ejemplo
Hardware	Gastos asociados a la adquisición y mantenimiento de la infraestructura necesaria para realizar las actividades de archivado web.	Se han invertido $1500 en el servidor del Archivo Web.
Procesamiento	Costos generados en recursos para que el Archivo Web pueda funcionar, como lo son el servicio de electricidad o el proveedor de Internet.	El servidor consume 1050 Watts por semana o genera gastos de $150 por semana.
Software	Depende de la opción de software tomada, en el caso de que no sean software libre. También se incluye el costo de desarrollo adicional u operaciones técnicas requeridas.	La licencia del software de rastreo genera un gasto anual de $600.
Personal	Cantidad de horas/hombre empleadas en las actividades del Archivo Web.	Para el monitoreo de los rastreos se emplean 12 horas diarias.
Legal	Tomado en consideración solo cuando se requiere asesoramiento legal.	La demanda realizada a la institución generó $1.5M en compensación a la compañía demandante.
Cooperación Internacional	Costo de membresías para organizaciones como la IIPC y/o gastos generados por viajes relacionados al Archivo Web.	El viaje para la charla divulgativa del Archivo Web de Venezuela generó gastos de $3000.

Fuente: (ISO Working Group, 2012)

2.3.5.8. Indicadores de calidad

A continuación, en la Tabla 18 se listan los indicadores de calidad definidos en el informe técnico de la ISO (ISO Working Group, 2012). Éstos se encuentran clasificados en: administrativos, calidad del proceso de recolección, accesibilidad y uso, y preservación.

Tabla 18 - Indicadores de Calidad definidos en documento ISO

Categoría	Nombre	Objetivo	Comentarios

Tabla 18 - Indicadores de Calidad definidos en documento ISO

Categoría	Nombre	Objetivo	Comentarios
Administrativo	Costo por URL recolectado	Evaluar la eficiencia de los procesos relacionados al archivado Web	Un costo bajo por URL recolectado demuestra una alta eficiencia en el proceso. Un costo alto puede también indicar un alto nivel de conservación
	Porcentaje del personal involucrado en el Archivo Web	Indicar el compromiso de la institución con el Archivo Web	Especialmente para bibliotecas e instituciones que tienen personal para diversas actividades, no sólo el Archivo Web
Calidad del proceso de recolección	Porcentaje de recursos desaparecidos de la Web viva durante un periodo de tiempo	Evaluar el valor del Archivo Web	Es considerado un sitio web desaparecido aquél que presenta una respuesta del DNS o cuando se genera una respuesta 404. Si es posible, el método más confiable es la revisión manual
	Porcentaje conseguido del alcance establecido	Evaluar si los resultados del archivado Web corresponden con los establecidos	El alcance obligatorio o la cobertura del Archivo Web puede ser establecido como los sitios web nacionales
	Porcentaje de solicitudes otorgadas por dueños	Evaluar la efectividad de las solicitudes de permisos	Un nivel alto indica la efectividad en la actividad de solicitud de permisos. Es recomendable almacenar los rechazos explícitos y el número de solicitudes sin respuesta
Accesibilidad y uso	Porcentaje de recursos accesibles por usuarios finales	Evaluar la disponibilidad del Archivo Web	La unidad en la que se calcula el indicador debería ser reportado
	Porcentaje de recursos indexados por textos completos	Evaluar la capacidad de búsqueda del Archivo Web	La búsqueda por textos completos aumenta considerablemente la accesibilidad y la usabilidad del Archivo Web
	Porcentaje de recursos catalogados	Evaluar la capacidad de búsqueda y el nivel de conservación del	Catalogar los recursos del Archivo Web aumenta la accesibilidad y la usabilidad,

Tabla 18 - Indicadores de Calidad definidos en documento ISO

Categoría	Nombre	Objetivo	Comentarios
		Archivo Web	así como también ayuda a integrar los recursos del Archivo Web con otros recursos retenidos por la biblioteca o institución
	Porcentaje anual de recursos accedidos	Evaluar la amplitud del uso actual del Archivo Web	Utilizar los dominios como unidad de medida muestra la amplitud del uso y entendimiento del Archivo Web por parte de los usuarios.
	Porcentaje de usuarios de la biblioteca utilizando el Archivo Web	Evaluar el uso del Archivo Web por parte de los usuarios de los servicios de la institución	Si es posible, el uso debería ser reportado en dos categorías de usuarios: los que realizan consultas dentro de la institución y los que realizan consultas desde afuera
Preservación	Porcentaje de los recursos con al menos una replicación	Evaluar la capacidad de preservación del flujo de datos	La unidad de medida debería ser reportada
	Porcentaje de los recursos deteriorados o perdidos	Evaluar la seguridad en el almacenamiento del Archivo Web	Un porcentaje bajo indica una seguridad elevada
	Porcentaje de recursos con un formato de archivo identificado	Evaluar el conocimiento institucional del Archivo Web y la capacidad de preservación	La unidad de medida debería ser reportada
	Porcentaje de recursos cuyo formato ha definido una estrategia de preservación	Evaluar el compromiso de la institución con la preservación lógica del Archivo Web	La unidad de medida debería ser reportada
	Porcentaje de recursos revisados por virus	Evaluar la utilización segura, del Archivo Web, por parte de otras colecciones y usuarios finales	En la detección de virus, las instituciones deberían definir políticas de no remoción de los virus del Archivo Web, sino bloquear el acceso a los recursos infectados

Fuente: (ISO Working Group, 2012)

2.4. Inteligencia de Negocio

Una rama donde se estudian y utilizan ampliamente los indicadores es la Inteligencia de Negocio, y los sistemas de Inteligencia de Negocio, por lo que en esta sección se van a definir los conceptos necesarios para entender este tipo de sistemas.

2.4.1. Sistemas de Información (SI)

Actualmente, las empresas buscan satisfacer las necesidades del cliente de la manera más rápida y efectiva, por lo que se ven en la necesidad de utilizar la tecnología actual y los sistemas de información, para así apoyar con estos la regularización de la información dentro de las empresas. Es por este hecho que, Laudon y Laudon (2012), definen a los sistemas de información como "componentes interrelacionados que trabajan en conjunto para recolectar, procesar, almacenar y diseminar información para soportar la toma de decisiones, la coordinación, el control, el análisis y la visualización en una organización".

2.4.1.1. Tipos de Sistemas de Información

Según Laudon y Laudon (2012), una organización de negocios tiene sistemas para dar soporte a los distintos grupos de niveles de administración. Estos sistemas incluyen sistemas de procesamiento de transacciones (TPS), sistemas de información gerencial (MIS), sistemas de soporte de decisiones (DSS) y sistemas de apoyo a ejecutivos (ESS), su jerarquía se puede visualizar en la Figura 14.

Figura 14 - Tipos de Sistemas de Información
Fuente: (Laudon & Laudon, 2012)

Laudon y Laudon (2012) clasifica los tipos de sistemas de información en los siguientes:

Sistemas de procesamiento de transacciones

Los gerentes operacionales necesitan sistemas que lleven el registro de las actividades y transacciones elementales de la organización, como ventas, recibos, depósitos en efectivo, nóminas, decisiones de créditos y el flujo de materiales en una fábrica. Los sistemas de procesamiento de transacciones (TPS) proveen este tipo de información.

Un sistema de procesamiento de transacciones es un sistema computarizado que efectúa y registra las transacciones diarias de rutina necesarias para realizar negocios, como introducir pedidos de ventas, reservaciones de hoteles, nómina, registro de empleados y envíos. El principal propósito de los sistemas en este nivel es responder a las preguntas de rutina y rastrear el flujo de transacciones por toda la organización. Los datos en el sistema se combinan en distintas maneras para crear informes de interés para la gerencia y las agencias gubernamentales. Los gerentes necesitan el TPS para supervisar el estado de las operaciones internas y las relaciones de la empresa con el entorno externo. Los TPS también son importantes productores de información para los otros sistemas y funciones de negocios.

Sistemas de información general para el soporte de decisiones

La gerencia de nivel medio necesita sistemas para ayudar con las actividades de monitoreo, control, toma de decisiones y administrativas, es decir, necesita los sistemas de información gerenciales (MIS). Estos sistemas proveen reportes sobre el desempeño actual de la organización, los cuales se utiliza para supervisar y controlar la empresa, además de predecir su desempeño en el futuro, sintetizando e informando sobre las operaciones básicas de la compañía, mediante el uso de datos suministrados por los sistemas de procesamiento de transacciones. Los MIS no son flexibles y tienen poca capacidad analítica, ya que utilizan rutinas simples, a diferencia de los sofisticados modelos matemáticos o las técnicas estadísticas.

En contraste, los sistemas de soporte de decisiones (DSS) brindan apoyo a la toma de decisiones que no es rutinaria, es decir, se enfocan en problemas que son únicos y cambian con rapidez. Tratan de responder interrogantes sobre casos hipotéticos aplicables dentro del negocio. Aunque los DSS usan información interna de los TPS y MIS, a menudo obtienen datos de fuentes externas. Estos sistemas usan una variedad de modelos para analizar los datos, y están diseñados para que los usuarios puedan trabajar con ellos directamente.

Por otro lado, los sistemas de apoyo a ejecutivos (EIS) ayudan a la gerencia de nivel superior a tomar estas resoluciones, es decir, se encargan de las decisiones no rutinarias que requieren de juicio, evaluación y perspectiva, por lo que presentan gráficos y datos de diversas fuentes a través de una interfaz usable para los gerentes de nivel superior. Estos sistemas están diseñados para incorporar datos sobre eventos externos, como leyes fiscales o competidores nuevos, pero también obtienen información sintetizada proveniente de sistemas MIS y DSS. Los EIS incluyen, cada vez en mayor grado, los análisis de inteligencia

de negocio para analizar tendencias, realizar pronósticos y "desglosar" los datos para obtener mayores niveles de detalle.

Los sistemas MIS, DSS y EIS, en conjunto, son los que integran una solución de Inteligencia de Negocio actualmente, tema que será abarcado en la siguiente sección, donde se define la Inteligencia de Negocio, su arquitectura, componentes, entre otros conceptos.

2.4.2. Definición de Inteligencia de Negocio

Según Loshin (2012), la Inteligencia de Negocio (BI, por sus siglas en inglés) son procesos, tecnologías y herramientas necesarias para transformar datos en información, información en conocimiento y conocimiento en planes de negocios rentables.

El término inteligencia de negocios, se utiliza para indicar el conjunto de conceptos y métodos que brinda apoyo en la toma de decisiones de las organizaciones, utilizando sistemas de apoyo basados en hechos. El objetivo básico es apoyar de forma sostenible y continua a las organizaciones, facilitando la información necesaria para la toma de decisiones.

"BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un almacén de datos), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones.

El proceso de Inteligencia de Negocio incluye la comunicación de los descubrimientos y efectuar los cambios.

Las áreas incluyen clientes, proveedores, productos, servicios y competidores." (Cano, 2007)

2.4.3. Características de una Solución de Inteligencia de Negocio

Según Cano (2007), toda solución de Inteligencia de Negocio debe cumplir con las siguientes características:

- Visión unificada de los datos: Todos los datos deben estar localizados en un único repositorio de datos, sin importar el tipo de datos o la fuente de donde provenga, para así dar la sensación de que los datos están centralizados.
- Creación personalizada de informes y consultas: permite el desarrollo de consultas y reportes a la medida sobre información contenida en los Almacenes de Datos.
- Vistas gráficas e interactivas para la presentación de información analítica: A través de cuadros de mandos integrales y estratégicos se facilita la visualización de los indicadores de negocio.
- Capacidad de procesamiento de grandes volúmenes de datos: las soluciones de BI permiten realizar consultas comparando los datos actuales con los históricos.

2.4.4. Funciones de una Solución de Inteligencia de Negocio

- Permiten reunir, estandarizar y centralizar toda la información de la empresa, mediante un almacén de datos, permitiendo así su explotación sin esfuerzo.
- Posibilita la extracción de información de los datos y el conocimiento de la información, con la utilización del software adecuado.
- Permiten el perfeccionamiento de las consultas de alto nivel, realizando las transformaciones oportunas a cada sistema, y liberando los servidores operacionales.

2.4.5. Arquitectura de una Solución de Inteligencia de Negocio

Una arquitectura básica para una solución de inteligencia de negocio está formada por los siguientes elementos (ver Figura 15):

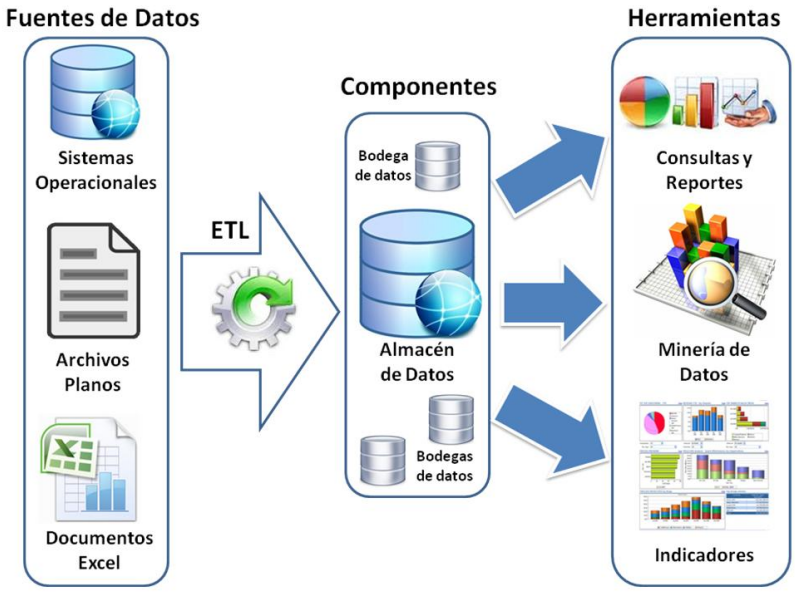


Figura 15 - Arquitectura Básica de una solución de Inteligencia de Negocio
Fuente: (Cano, 2007)

Según Cano (2007), los componentes que conforman la arquitectura de una solución de inteligencia de negocio son los siguientes:

2.4.5.1. Fuentes de datos

Básicamente, son los datos que alimentan de información el almacén de datos, los cuales salen de los sistemas operacionales o transaccionales, incluyendo aplicaciones desarrolladas a la medida, ERP, CRM, SCM, entre otros sistemas, con información relativa a la actividad rutinaria de la empresa y de los sistemas de información departamentales, incluyendo previsiones, presupuestos y hojas de cálculo, que se requieren para el análisis del negocio.

2.4.5.2. Extracción, Transformación y Carga

Comúnmente este proceso es conocido como ETL por sus siglas en inglés (Extract, Transformation, and Load) y se estima que a menudo consume un 70% del tiempo y esfuerzo para la construcción de un almacén de datos y se utiliza para migrar datos de un punto a otro. Su nombre claramente indica las 3 fases que conforman el proceso.

- Extracción: en esta fase se obtienen los datos provenientes de las diferentes fuentes externas e internas como sistemas transaccionales, archivos planos, hojas de cálculo, entre otros. Además, esta fase también incluye un filtrado de los datos, de tal manera que se eliminen datos redundantes o de poco interés.
- Transformación: en la fase de transformación se aplican una serie de reglas a los datos que han sido extraídos y que serán cargados en el almacén de datos. Algunos de estos datos pueden no necesitar modificación alguna, pero otros podrían necesitar por ejemplo: un nuevo formato, consolidarse datos de diferentes fuentes, rechazar datos no requeridos, crear datos derivados de otros, entre otros.
- Carga: es la última fase en el proceso ETL, y en ella se cargan los datos transformados al almacén de datos. Dado a que en esta fase existe una interacción directa con la base de datos, al momento de insertar los datos, se activan las restricciones existentes, lo cual ayuda a la limpieza de datos y complementa todo el proceso de ETL.

2.4.5.3. Área Intermedia

Normalmente, la información que se tiene en los sistemas transaccionales no está preparada para la toma de decisiones, por lo tanto se busca almacenar los datos de una forma que maximice su flexibilidad, facilidad de acceso y administración. (Cano, 2007)

El área intermedia es un espacio temporal y de carácter volátil, sobre el cual se van a ejecutar los procesos de ETL. Se usa para hacer una primera extracción rápida de las fuentes de datos y almacenarlos temporalmente mientras se analizan, limpian, mejoran y, posteriormente, se carga al almacén de datos.

2.4.5.4. Almacén de datos

Como se definió previamente, el almacén de datos es la base para los Sistemas de Soporte de Decisión (DSS) y su principal papel es el de integrar datos de diversas fuentes en una única base de datos centralizada y accesible por las aplicaciones que dan soporte para el proceso de toma de decisiones gerenciales. Con esto se simplifica el problema de acceso a la información y en consecuencia, acelera el proceso de análisis, consultas y el menor tiempo de uso de la información. De manera que un directivo o analista pueda realizar evaluaciones, sin la mediación del personal informático de la empresa, por lo que dedica tiempo al análisis y extracción de valor añadido de la información sin pérdida de tiempo. (Inmon, 1996)

2.4.5.5. Cubo

Es una estructura multidimensional generada por la intersección de las dimensiones enfocadas en el hecho a medir en una plataforma de base de datos multidimensional o procesamiento analítico en línea (OLAP). El nombre de cubo se refiere a un caso específico de tres dimensiones. Los cubos pueden tener un número indefinido de dimensiones, por este motivo también son llamados hipercubos. (Ponniah, 2001)

Los cubos son subconjuntos de datos de un almacén de datos, organizado y sumarizado dentro de una estructura multidimensional. Los datos se sumarizan de acuerdo a factores de negocio seleccionados, proveyendo el mecanismo para la rápida y uniforme tiempo de respuesta de las complejas consultas (ver Figura 16).

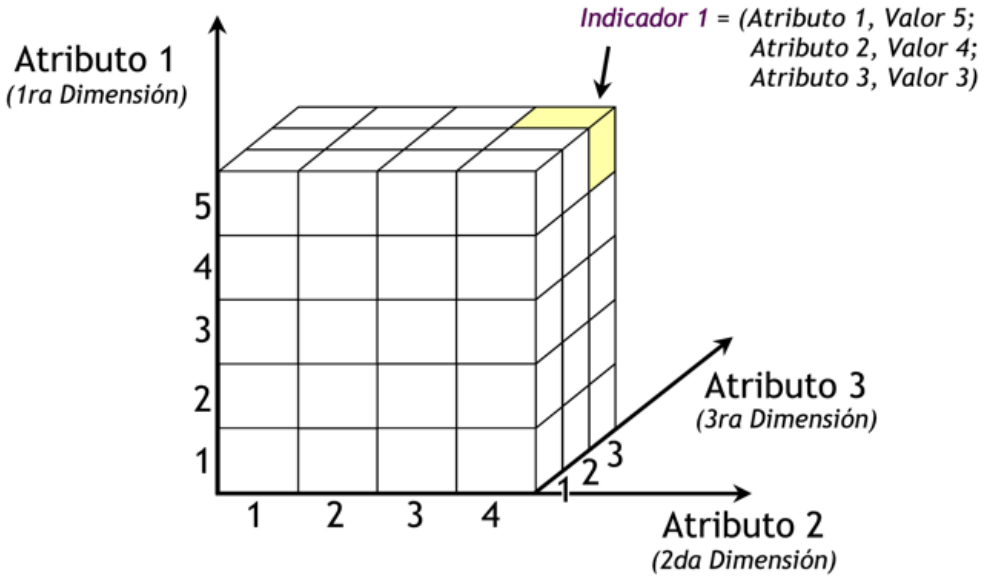


Figura 16 – Cubo
Fuente: (Ponniah, 2001)

- Generadores de informes: Utilizadas por desarrolladores profesionales para crear informes estándar para grupos, departamentos o la organización.
- Herramientas de usuario final de consultas e informes: empleadas por usuarios finales para crear informes para ellos mismos o para otros; no requieren programación.
- Herramientas OLAP: permiten a los usuarios finales tratar la información de forma multidimensional para explorarla desde distintas perspectivas y periodos de tiempo.
- Herramientas de cuadros de mando: permiten a los usuarios finales ver información crítica para el rendimiento con un simple vistazo utilizando iconos gráficos y con la posibilidad de ver más detalle para analizar información detallada e informes, si lo desean.
- Herramientas de planificación, modelización y consolidación: permite a los analistas y a los usuarios finales crear planes de negocio y simulaciones con la información de Inteligencia de Negocio. Pueden ser para elaborar la planificación, los presupuestos, las previsiones. Estas herramientas proveen a los cuadros de mando con los objetivos y los umbrales de las métricas.

- Herramientas de minería de datos: permiten a estadísticos o analistas de negocio crear modelos estadísticos de las actividades de los negocios. Minería de datos es el proceso para descubrir e interpretar patrones desconocidos en la información mediante los cuales resolver problemas de negocio. Los usos más habituales de la minería de datos son: segmentación, venta cruzada, sendas de consumo, clasificación, previsiones, optimizaciones, etc.

2.5. Bases de Datos NoSQL

Todo sistema de Inteligencia de Negocio requiere un componente de almacenamiento de datos, donde usualmente está contenido el almacén de datos. Sin embargo, en la actualidad han surgido nuevos tipos de bases de datos que cumplen ciertas características necesarias para el almacenamiento de grandes volúmenes de datos, como son las Bases de Datos NoSQL.

El término *NoSQL* fue utilizado por primera vez en 1998, por Carlo Strozzi, para referirse a una base de datos de código abierto, que omitía el uso de *SQL*, pero sí seguía el modelo relacional. Actualmente, significa "*Not Only SQL*" (No solo SQL), implicando que cuando se diseña una solución o un producto, hay más de un mecanismo de almacenamiento que puede ser utilizado basado en las necesidades, no únicamente las bases de datos relacionales.

2.5.1. Definición *NoSQL*

NoSQL engloba una amplia variedad de diferentes tecnologías de bases de datos, que fueron desarrolladas en respuesta a un aumento en el volumen de los datos almacenados, la frecuencia en la cual estos datos son accedidos y las necesidades de rendimiento y procesamiento. Por otro lado, las bases de datos relacionales no fueron diseñadas para hacer frente a los desafíos de escalabilidad y rapidez que enfrentan las aplicaciones modernas, ni fueron construidas para aprovechar el bajo costo del almacenamiento y el poder de procesamiento disponible actualmente. (MongoDB, Inc., 2015)

Formalmente, no existe una definición de *NoSQL*, pero se pueden establecer un conjunto de observaciones que se encuentra en cualquier base de datos *NoSQL* (Sadalage, 2014):

- No se utiliza el modelo relacional.
- Buen rendimiento en clústeres de servidores.
- Mayormente código abierto.
- Construido para los estados web del siglo XXI.
- Sin esquema.

2.5.2.Teorema de *CAP*

En el año 2000, Erick Brewer presentó su discurso de apertura en el *ACM Symposium on the Principles of Distributed Cumputing* (Simposio ACM sobre los Principios de la Computación Distribuida), donde planteó que existen tres requerimientos de sistema, esenciales y necesarios, para la realización exitosa del diseño,

la implementación y el lanzamiento de aplicaciones en sistemas distribuidos. Esos requerimientos esenciales son: Consistencia (*Consistency*), Disponibilidad (*Availability*) y Tolerancia a Particiones (*Partition toleration*), es decir, CAP por sus siglas en inglés. (Roe, 2012)

- Consistencia: se refiere a cuando un sistema opera plenamente o no, es decir, si un sistema se encuentra en un estado consistente después de la ejecución de una operación y cómo.
- Disponibilidad: significa que un sistema es capaz de continuar con las operaciones a pesar de presentar alguna falla en algún componente de hardware o software.
- Tolerancia a Particiones: representa la capacidad del sistema para continuar operando en presencia de particiones de red. También es descrita como la capacidad del sistema a tolerar la adición y/o eliminación dinámica de nodos.

El teorema *CAP* establece que, en cualquier sistema distribuido, se puede favorecer solo a dos (2) características entre las mencionadas anteriormente (ver Figura 17), por lo que Brewer indica que se debe utilizar como criterio de selección los requerimientos más críticos para el negocio, optando entre las propiedades *ACID* y *BASE*. Actualmente, muchas bases de datos *NoSQL* intentan proveer opciones para el desarrollador que le permitan elegir donde puede modificar la base de datos de acuerdo a sus necesidades.

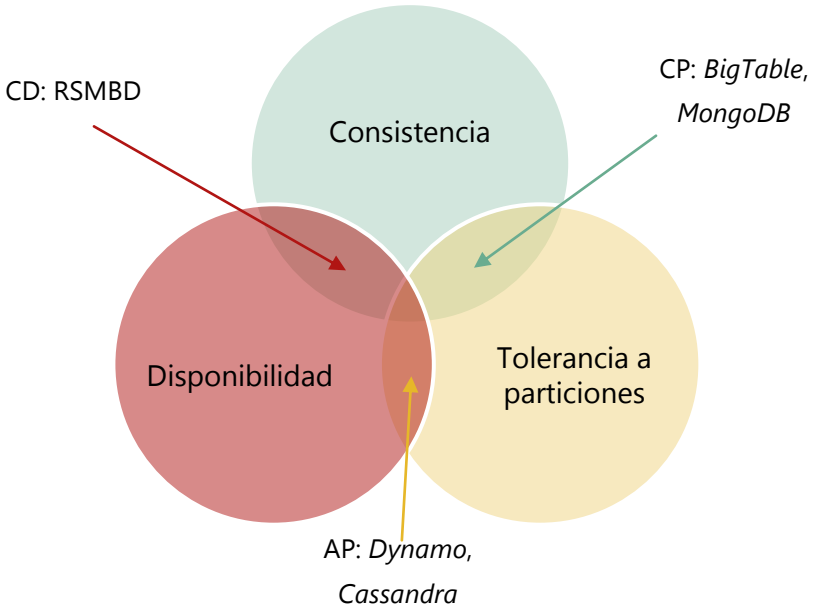


Figura 17 - Diagrama de *CAP*
Fuente: (Roe, 2012)

2.5.3.Propiedades *BASE*

En las bases de datos relacionales se cumplen las propiedades *ACID*, que se caracterizan por proveer consistencia sobre disponibilidad, por otro lado las bases de datos *NoSQL* cumplen las propiedades *BASE* las cuales dan mayor prioridad a la disponibilidad sobre la consistencia. La palabra *BASE* viene dada por

Basically Available (básicamente disponible), *Soft state* (estado flexible) y *Eventually consistent* (eventualmente consistente), y son las propiedades que permiten asegurar la disponibilidad de los datos. A continuación, se explican brevemente las propiedades: (Roe, 2012)

- Básicamente disponible: esta propiedad asegura que siempre habrá una respuesta a cualquier solicitud, aunque la respuesta no siempre sea correcta.
- Estado flexible: el estado del sistema puede cambiar con el paso del tiempo. Así que, a pesar de que no haya una entrada de datos, pueden estar realizándose cambios en el sistema dada la siguiente propiedad.
- Eventualmente consistente: con el tiempo el sistema será consistente, ya que los datos se propagarán a todos los nodos correspondientes en un momento futuro.

2.5.4. Tipos de BD *NoSQL*

Las bases de datos *NoSQL* pueden ser categorizadas en cuatro (4) tipos:

2.5.4.1. Almacenamiento Clave-Valor

La base de datos clave-valor es la más simple de utilizar, ya que cada objeto dentro de la base de datos es almacenado como clave con su valor correspondiente. El usuario puede obtener un valor por su clave, asignarle un valor a una clave o eliminar la clave de la base de datos, gracias a un mapa o diccionario (DHT). Este tipo de base de datos favorece la escalabilidad sobre la consistencia, pero limita considerablemente las funcionalidades de análisis y consultas complejas *ad-hoc*.

Algunos ejemplos del almacenamiento clave-valor son *Riak* y *Voldemort*. Algunos de éstos, como *Redis*, ofrecen la posibilidad de asignarle un tipo de dato a cada valor, aumentando su funcionalidad. (Sadalage, 2014)

2.5.4.2. Bases de datos orientadas a Grafos

Las bases de datos orientadas a grafos se inspiraron en la teoría de grafos de Euler, utilizando nodos, aristas y propiedades para representar y almacenar datos. La organización del grafo permite que los datos sean almacenados una vez y que puedan ser interpretados de diferentes maneras, basándose en las relaciones.

En estas bases de datos, la realización de *joins* o atravesar relaciones es bastante rápido. Las relaciones entre los nodos no son calculadas en el momento en que se ejecuta la consulta, sino que es persistente como una relación.

Como la mayoría del atractivo de las bases de datos orientadas a grafos viene de las relaciones y sus propiedades, se requiere mucho trabajo pensando y diseñando el modelo de las relaciones en el dominio en que se está trabajando. (Sadalage, 2014)

Algunas de las bases de datos de este tipo que se encuentran disponibles son *Neo4J, Infinite Graph, OrientDB, HyperGraphDB o FlickDB*.

2.5.4.3. Bases de datos orientadas a Columna

Las bases de datos orientadas a columna almacenan los datos en familias de columnas como las filas que tienen muchas columnas, asociadas a una clave de fila. Las familias de columnas son grupos de datos relacionados que, frecuentemente, son accedidas juntas.

Cada familia de columnas puede ser comparada con un contenedor de filas en una base de datos tradicional, donde la llave identifica una fila y ésta consiste en múltiples columnas. La diferencia radica en que en varias filas no necesitan tener las mismas columnas, y las columnas pueden ser añadidas a cualquier fila en cualquier momento sin tener que añadirlo a otras filas. (Sadalage, 2014)

Cuando una columna está compuesta por un mapa de columnas, entonces se tiene una "super columna". Éstas consisten en un nombre y un valor, que es un mapa de columnas.

Algunas de las bases de datos orientadas a columna son *Cassandra, HBase, Hypertable y DynamoDB*.

2.5.4.4. Bases de datos Documentales

Las bases de datos documentales almacenan y recuperan documentos, que puede ser XML, JSON, BSON, entre otros. Estos documentos son autodescriptivos y los datos se encuentran estructurados en un árbol jerárquico, dentro del cual se encuentran los pares clave-valor, clave-areglos o hasta documentos anidados. Los documentos que se almacenan son similares entre ellos, pero no tienen que ser exactamente iguales. Algunos de las bases de datos más populares son: *MongoDB, CouchDB, Terrastore, OrientDB y RavenDB*.

Adicionalmente, el modelo de datos de las bases de datos documentales posee una estructura jerárquica, como se puede ver en la Figura 18, y está compuesto por los siguientes componentes:

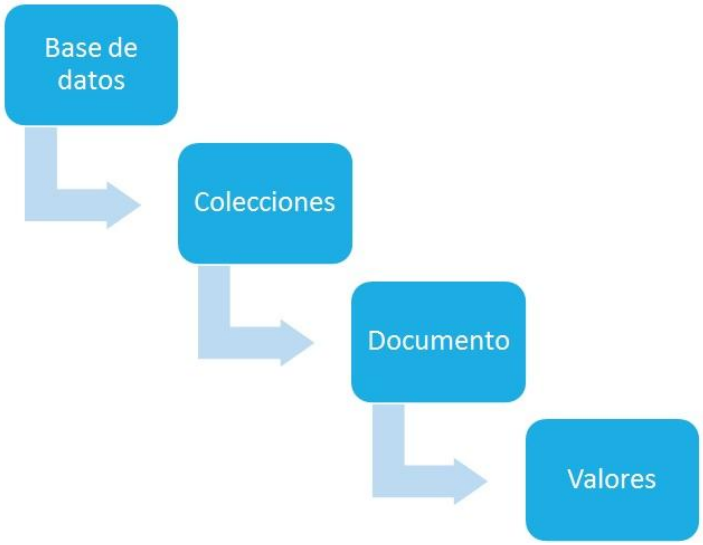


Figura 18 - Estructura de Modelo de Datos de BD Documentales
Fuente: Basado en (MongoDB, Inc., 2015)

- Base de Datos: contenedor físico de colecciones o documentos.
- Colección: agrupaciones de documentos. Cabe destacar que no necesariamente está presente en todas las bases de datos documentales.
- Documento: unidad básica de datos. Equivalente a un registro en una base de datos tradicional.
- Valores: datos atómicos, que se pueden encontrar almacenados en listas o arreglos.

2.5.5. Transformación de Base de Datos Relacional a *NoSQL* Documental

Las bases de datos tradicionales vienen dadas por tablas y sus relaciones explicitas entre ellas, pero en las bases de datos documentales no es posible definir estas relaciones tan explícitamente, ya que las estructuras que se manejan son documentos (ver Figura 19).

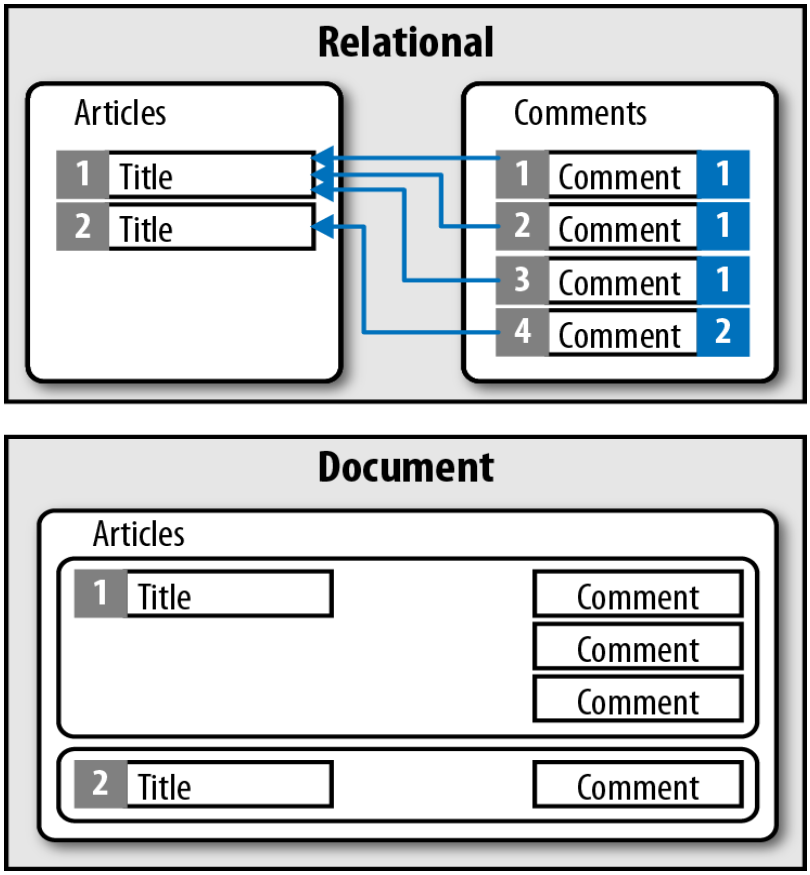


Figura 19 - Diferencia entre Bases de Datos Relacionales y Documentales
Fuente: (Francia, 2012)

Si se requiere llevar un modelo dimensional tradicional a un formato documental, es importante conocer y manejar las técnicas que permiten realizar dicha transformación del modelo dimensional relacional a un modelo documental.

Existen dos (2) patrones de modelado que permiten establecer la estructura de los datos que tendrán los documentos dentro de la base de datos documental, que en una base de datos estarían almacenados en diferentes tablas. Estos patrones son denominados por MongoDB (MongoDB, Inc., 2015) como modelos de datos y los define de la siguiente manera:

2.5.5.1. Modelo de datos embebido

Se basa en incrustar un documento dentro de otro, como se ve en la Figura 20, con la finalidad de hacerlo parte del mismo registro de la base de datos. Este patrón da como resultado un modelo de datos desnormalizado, que permite a las aplicaciones ejecutar menos consultas y actualizaciones para completar operaciones comunes. Este patrón se utiliza cuando se tienen relaciones uno-a-uno o uno-a-muchos entre entidades. (MongoDB, Inc., 2015)

Figura 20 - Ejemplo modelo de datos embebido
Fuente: http://docs.mongodb.org/master/core/data-model-design/

En general, embeber provee un mejor rendimiento para operaciones de lectura, así como la habilidad de solicitar y recuperar datos en una sola operación a la base de datos. También permite actualizar datos relacionados en una sola operación de escritura.

De igual manera, embeber datos relacionados en documentos puede ocasionar que los documentos crezcan, después de ser creados, hasta el punto de afectar el rendimiento de las operaciones de escritura y conllevar a la fragmentación de los datos.

2.5.5.2. Modelo de datos referencial

Busca imitar el comportamiento de las claves foráneas para relacionar datos que se encuentran en colecciones diferentes, es decir, describe las relaciones utilizando referencias entre documentos, ocasionando que sea un modelo de datos normalizado, como se aprecia en la Figura 21. Este patrón se utiliza cuando embeber puede resultar en datos duplicados, sin beneficiar el rendimiento de lectura; para representar relaciones muchos-a-muchos, y para modelar largos conjuntos de datos jerárquicos. (MongoDB, Inc., 2015)

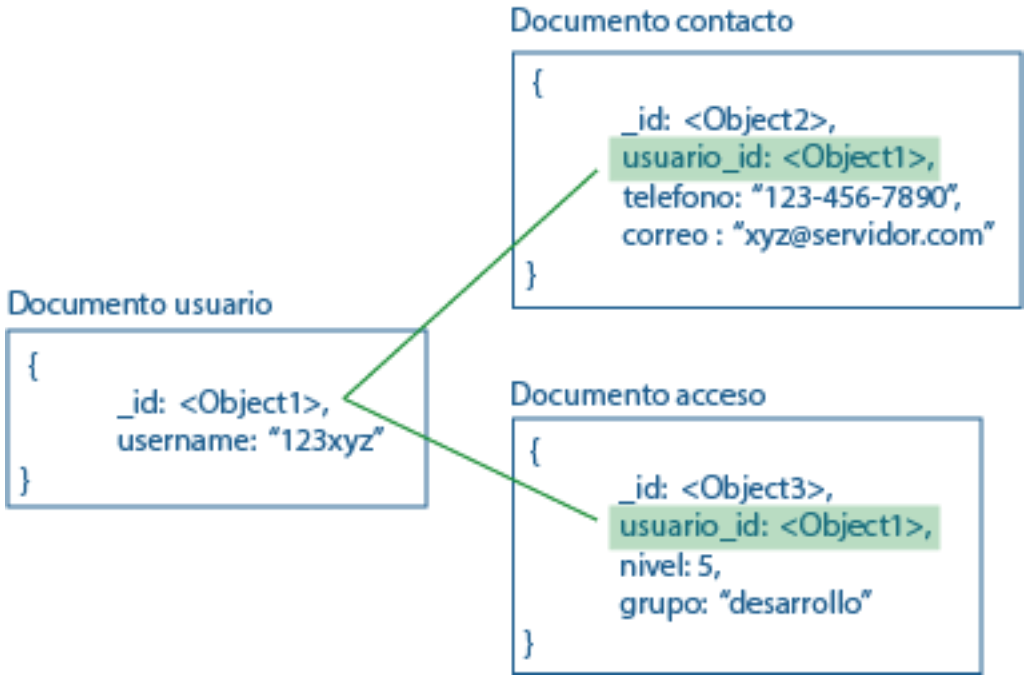


Figura 21 - Ejemplo modelo de datos referencial
Fuente: http://docs.mongodb.org/master/core/data-model-design/

Aunque este modelo provee más flexibilidad, las aplicaciones del lado del cliente deben realizar el seguimiento de las consultas para resolver las referencias, es decir, la normalización puede requerir más solicitudes al servidor.

Capítulo 3 Marco Metodológico

En esta sección, se propone una adaptación de la metodología propuesta por Kimball, explicando a grosso modo las actividades dentro de ésta y los cambios realizados en la implementación del diseño físico.

3.1. Metodología de Desarrollo

La metodología ascendente (*bottom-up*) propuesta por Kimball (1998), denominada ciclo de vida dimensional del negocio (*Business Dimensional Lifecycle*), se utiliza a la hora de diseñar y crear almacenes de datos. La construcción de un almacén de datos es sumamente complicado, y Kimball nos propone una metodología que nos ayuda a simplificar esa complejidad. Las tareas de esta metodología se muestran en la Figura 22.

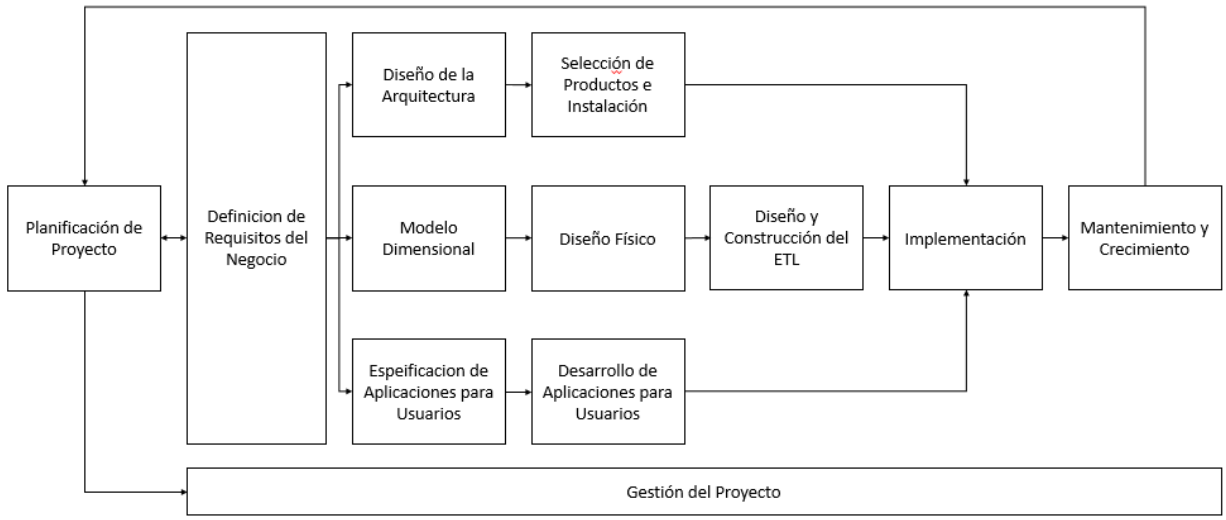


Figura 22 - Ciclo de Vida de Kimball
Fuente: (Kimball, 1998)

Para este Trabajo Especial de Grado, se realiza una adaptación a la metodología mostrada, modificando algunas actividades y complementado otras con metodologías de desarrollo de software. En la Figura 23, se muestra el Ciclo de Vida adaptado y se explican a continuación las actividades que fueron modificadas:

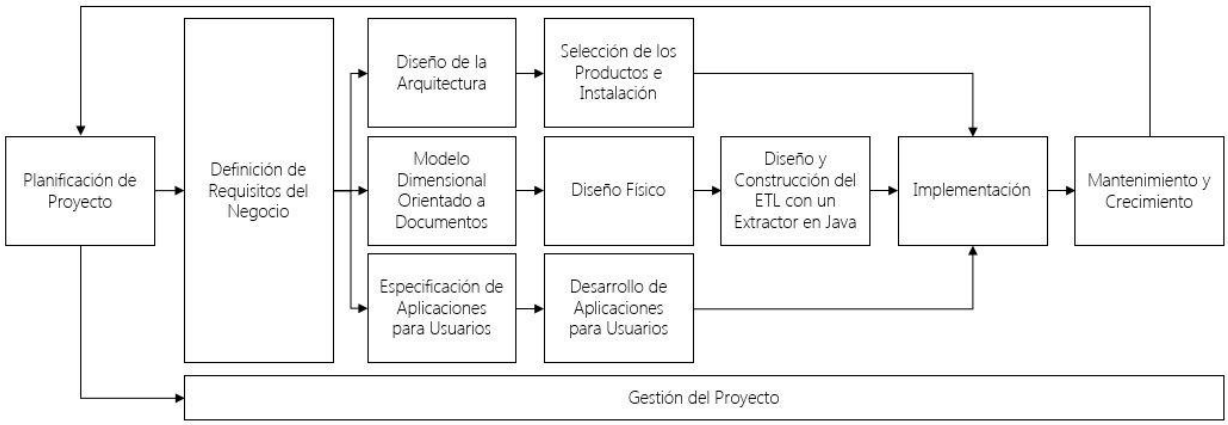


Figura 23 - Adaptación Ciclo de Vida de Kimball
Fuente: Elaboración propia basada en (Kimball, 1998)

- Planificación y Gestión del Proyecto: este primer paso da inicio al desarrollo del proyecto. Aquí, se debe definir el proyecto, elaborar un plan de trabajo y gestionar la puesta en marcha del mismo, definiendo y respetando el alcance establecido.
- Definición de Requerimientos: aunque la definición de requerimientos está en gran parte influenciada por entrevistas con personal de negocio y técnico, siempre es conveniente tener el más amplio conocimiento posible sobre el negocio y el ambiente donde éste se desarrolla. En esta fase, y a partir del análisis, se puede construir una herramienta de la metodología denominada matriz de procesos/dimensiones, o la matriz de Bus.
- Diseño Arquitectura Técnica: cubre los procesos y herramientas que se aplican a los datos. Es importante conocer que existen principios de diseño que deben presentar todas las herramientas, procesos y componentes de las capas de servicio.
- Selección de las herramientas: en esta fase, se decide qué herramientas serán usadas para el desarrollo del proyecto, tales como el motor de la BD, herramientas de ETL, entre otros. Luego de la instalación, se debe probar el correcto funcionamiento de las herramientas en todos los ambientes disponibles.
- Modelo Dimensional orientado a Documentos: para el diseño del modelo dimensional, se utiliza como guía el patrón de diseño planteado en la sección Modelo de datos embebido, utilizado para transformar un modelo relacional a un modelo documental.
- Diseño y Construcción del ETL con un extractor en Java: para el proceso ETL se plantea hacer una aplicación en Java, por lo que el desarrollo de ésta se guiará por los principios del Manifiesto ágil (Beck, y otros, 2001) listados a continuación:
    - Se acepta que los requisitos cambien, incluso en etapas tardías del desarrollo.
    - Los responsables del negocio y los desarrolladores trabajan en conjunto, de forma cotidiana, durante todo el proyecto.
    - Desarrollo sostenible.
    - La atención continua a la excelencia técnica y al buen diseño mejora la agilidad.

  - A intervalos regulares, el equipo reflexiona sobre cómo ser más efectivo para, a continuación, ajustar y perfeccionar su comportamiento en consecuencia.
- Especificación de aplicaciones para Usuarios: en esta fase, se definen las herramientas a ser utilizadas para el desarrollo de la plataforma BI, en donde el usuario podrá visualizar los indicadores. Una vez instaladas las herramientas, se debe verificar su correcto funcionamiento.
- Desarrollo de aplicaciones para Usuarios: en esta actividad, se utilizaron como guías las buenas prácticas definidas en el método AgilUs, un método de desarrollo ágil para la creación de software usable, creado en el Centro de Ingeniería de Software y Sistemas (ISYS) de la Escuela de Computación. Las buenas prácticas se explican a continuación:
  - Diseño centrado en el usuario (DCU): éste es un enfoque de diseño y desarrollo que se centra en los deseos, limitaciones y necesidades de los usuarios finales de un software. En las técnicas de DCU, es relevante que los desarrolladores realicen pruebas constantes para verificar el curso que lleva el desarrollo del sistema y su interfaz de usuario. De este modo, el usuario guía indirecta pero influyentemente el proceso de desarrollo del sistema. (Acosta, 2005)
  - Diseño basado en prototipos: tras una inspección inicial, los desarrolladores producen un primer prototipo, los especialistas y usuarios lo evalúan, los analistas preguntan directamente al usuario sus opiniones sobre el desarrollo, y con esa retroalimentación, los desarrolladores se disponen a producir el siguiente prototipo. Este ciclo continúa hasta que se tiene un producto listo para la entrega. (Acosta, 2005)
  - Desarrollo ágil, iterativo e incremental: se recomienda desarrollar un sistema simple que satisfaga las necesidades actuales de los usuarios, preparándose para cambios futuros. El desarrollo por incrementos permite proveer resultados sin necesidad de esclarecer todos los requisitos de una vez, al inicio del desarrollo. La iteratividad permite regresar a etapas anteriores una vez recibida la retroalimentación, producto de las evaluaciones realizadas. (Acosta, 2005)
  - Usabilidad como atributo de la calidad: la usabilidad es considerada un atributo de la calidad interna y externa del software, por lo que se sigue la recomendación del estándar ISO/IEC 9126-1 (2001). (Acosta, 2005)
  - Interacción continua con el usuario: se propicia un intercambio cara a cara, ya que la presencia constante y participativa del usuario es fundamental. El equipo de desarrollo sólo puede tomar decisiones tras realizar evaluaciones de usabilidad, y ésta sólo puede ser determinada por el usuario a través de las pruebas de aceptación con dicho usuario. (Acosta, 2005)
- Implementación: culminación del proceso de desarrollo del proyecto, que representa el desarrollo de la convergencia de todas las tecnologías, datos y la aplicación de análisis, que sea accesible directamente por los usuarios. También, es necesario realizar jornadas de entrenamiento para los usuarios finales, y se ofrece el servicio de soporte técnico a la organización

Capítulo 4 Marco Aplicativo

En esta sección se explicará el desarrollo de la solución de inteligencia de negocio, siguiendo la metodología descrita en el capítulo anterior, así como también las tecnologías utilizadas en cada actividad del proyecto, en caso de aplicar.

4.1. Definición de Requerimientos

Una vez leído el borrador del documento de la ISO (ISO Working Group, 2012), y basándonos en la información disponible dentro de los WARCS y las bases de datos existentes del Prototipo de Archivo Web de Venezuela, como primera actividad se definieron los indicadores descritos en la Tabla 19 y especificados en la Tabla 20.

Tabla 19 - Descripción de los Indicadores Propuestos

N°	Nombre	Descripción	Actores	Utilidad
1	Cantidad de semillas	Cantidad de sitios web rastreados actualmente	Administrador, Director y Suscriptor	Definir y medir objetivos del Archivo Web, monitoreo técnico y análisis de costos de todos los niveles del flujo de trabajo.
2	Cantidad de rastreos	Cantidad de trabajos de rastreos realizados	Administrador, Director y Suscriptor	Planificación de la preservación, definir y empaquetar solicitudes de recolecciones a gran escala.
3	Cantidad de colecciones	Cantidad de colecciones actualmente	Administrador, Director y Suscriptor	Planificar y organizar proyectos específicos de cosechado, diseñar mayor nivel de granularidad para propósitos de gestión, descripción y acceso del Archivo Web.
4	Cantidad de URLs	Cantidad de URLs procesados en el archivo web	Administrador, Director y Suscriptor	Monitoreo técnico y análisis de costos de todos los niveles del flujo de trabajo.
5	Distribución de URLs por código de estatus	Porcentaje de distribución de los URLs por su código de status	Administrador, Director y Suscriptor	Planificación de preservación, monitoreo técnico y análisis de costos de todos los
6	Cantidad de	Cantidad de WARCs	Administrador,	

Tabla 19 - Descripción de los Indicadores Propuestos

N°	Nombre	Descripción	Actores	Utilidad
	WARCs	almacenados en el archivo web	Director y Suscriptor	niveles del flujo de trabajo.
7	Duración promedio rastreo	Tiempo promedio que toma hacer el rastreo completo de todos los WARCs de una versión de una semilla	Administrador, Director y Suscriptor	Evaluación costos de adquisición
8	Tamaño del archivo web	Total bytes rastreados descomprimidos	Administrador, Director y Suscriptor	Caracterización y monitoreo del Archivo Web.
9	Distribución de URLs	Porcentaje de distribución de los URLs por semilla o colección	Administrador, Director, Suscriptor y Usuario Final	Planificación de preservación
10	Distribución por tipos de formatos	Porcentaje de distribución de los tipo MIME	Administrador, Director y Suscriptor	Evaluar los riesgos y las prioridades, y definir estrategias de migración y emulación para los archivos a largo plazo
11	Cantidad de URL por formato	Cantidad de URL de cada tipo MIME	Administrador, Director, Suscriptor y Usuario Final	
12	Cobertura cronológica	Fecha desde la cual se empezó a rastrear un objetivo	Administrador y Director	Caracterización y monitoreo del Archivo Web
13	Costo por objetivo recolectado	Porcentaje ocupado por objetivo dentro del tamaño del rastreo	Administrador, Director y Suscriptor	Evaluación costos de adquisición
14	Porcentaje de recursos desaparecidos de la web viva durante un periodo de tiempo	Porcentaje de distribución de los recursos desaparecidos de la web en un periodo de tiempo	Administrador y Director	Evaluación de riesgos por no actuar y la necesidad de ejecutar acciones rápidamente

Fuente: Elaboración propia

Tabla 20 - Especificación de los Indicadores Propuestos

N°	Fórmula	Unidad	Criterio de clasificación	Representación

Tabla 20 - Especificación de los Indicadores Propuestos

N°	Fórmula	Unidad	Criterio de clasificación	Representación
1	Conteo de semillas	#	Por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)
2	Conteo de rastreos	#	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico
3	Conteo de colecciones	#	Por fecha (mes y año)	Gráfico histórico
4	Conteo de URL	#	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico
5	Conteo URL por código estatus/ Conteo URL * 100	%	Por semilla, por colección, por fecha (año)	Gráficos de torta, gráfico de barras
6	Conteo de WARCs	#	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico
7	∑duración rastreos/cantidad rastreos	Minutos	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico
8	∑bytes rastreados	Bytes	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico
9	Conteo de URL del archivo web	%	Por semilla, por colección, por fecha (mes y año)	Gráfico de torta
10	Cantidad de bytes de formato/total URL formatos	%	Por semilla, por colección, por fecha (año)	Gráfico de torta, gráfico de línea
11	Conteo de recursos por formato	%	Por semilla, por colección, por fecha (año)	Gráfico de torta
12	Fecha inicio rastreo	Fecha	Por semilla, por colección, por fecha	Gráfico histórico
13	∑bytes rastreados/conteo URL	Bytes	Top 10 por semilla, por colección, por fecha (año)	Gráfico de torta, gráfico de barra
14	∑semillas desaparecidas/semillas	%	Por colección, por	Gráfico de torta

Tabla 20 - Especificación de los Indicadores Propuestos

N°	Fórmula	Unidad	Criterio de clasificación	Representación
	totales * 100		fecha	

Fuente: Elaboración propia

4.2. Diseño Técnico

Originalmente, la arquitectura de la solución de Inteligencia de Negocio fue inicialmente diseñada para utilizar las bases de datos "Preservación" y "Appacceso", así como también los *logs* generados por *Heritrix*[13], integrando todos los datos en una base de datos intermedia, desarrollada en *PostgreSQL*, incluyendo los extraídos de los WARCS. Luego, se construiría el almacén de datos, en el mismo sistema manejador, de acuerdo a los indicadores definidos en la sección anterior.

Cuando se realiza el análisis de los metadatos almacenados dentro de los WARCs, su estructura y la gran cantidad de datos que representan, se pudo observar que la arquitectura planteada originalmente no iba a ser suficiente para crear e implementar la solución deseada, mucho menos para solventar la necesidad de información de los diferentes usuarios definidos en el Prototipo. Es por esto que se plantea una nueva arquitectura, utilizando tecnologías más adecuadas a la cantidad de datos a ser manipulados dentro de la solución.

Como se puede ver en la Figura 24, las fuentes de datos se mantienen casi iguales, siendo éstas la base de datos "Preservación" y los metadatos extraídos de los archivos WARC, los cuales están en formato JSON. Para extraer la información de las fuentes de datos, se implementa un extractor en Java que, utilizando *WAT Utilies*, extrae cada registro "metadata" de los WARCs y genera un documento JSON por cada registro, es decir, un documento por recurso, manteniendo la estructura definida en el Formato de la Transformación del Archivo Web (WAT) (sección 2.2.3). Luego se extraen los datos de la base de datos "Preservacion" y se insertan al inicio de cada documento JSON. Al finalizar el procesamiento, se insertan los documentos JSON en el almacén implementado en *MongoDB*.

[13] Heritrix es un software que permite recorrer las páginas Web en profundidad. Es una herramienta de software libre, extensible, escalable y de calidad de archivado del "Internet Archive".

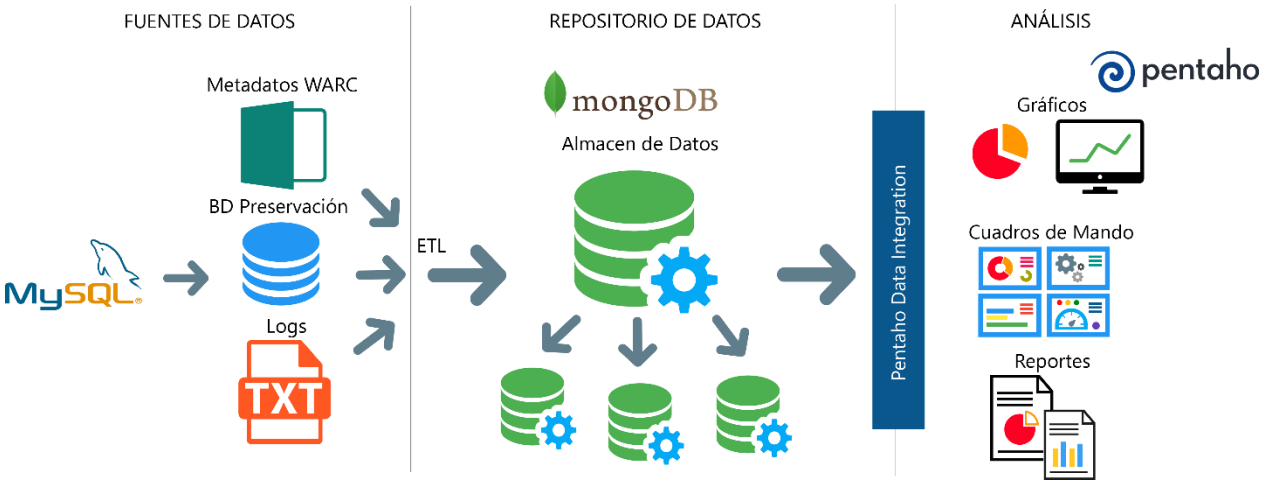


Figura 24 – Arquitectura lógica de la Solución
Fuente: Elaboración propia

Una vez que los documentos se encuentren almacenados en *MongoDB*, se utiliza *Pentaho Data Integration* para extraer y transformar los datos, y poder generar los gráficos creados utilizando *Sparkl* y *CTools* de *Pentaho*, conformando éstos la herramienta de análisis implementada para visualizar los indicadores.

4.3. Definición de Herramientas

4.3.1. WAT Utilities

Como se menciona anteriormente, las utilidades WAT (*WAT Utilities*, en inglés) son utilizadas para extraer metadatos de los archivos WARC, estructurando los metadatos utilizando el Formato de la Transformación del Archivo Web (WAT).

4.3.2. MongoDB

MongoDB es una base de datos documental, de código abierto, que provee alto rendimiento, disponibilidad y escalabilidad automática. Para esta base de datos *NoSQL*, un registro es representado por un documento, que es una estructura de datos compuesta de duplas campo-valor.

4.4. Modelo Dimensional Orientado a Documento y Diseño Físico

En la Figura 25, se muestra el diseño lógico del almacén de datos de la solución, es decir, el modelo dimensional. Para realizarlo, se necesitó conocer la información requerida por los indicadores y que necesitaba ser extraída de la base de datos y los *logs*, para así poder determinar las dimensiones que conformarían el modelo.

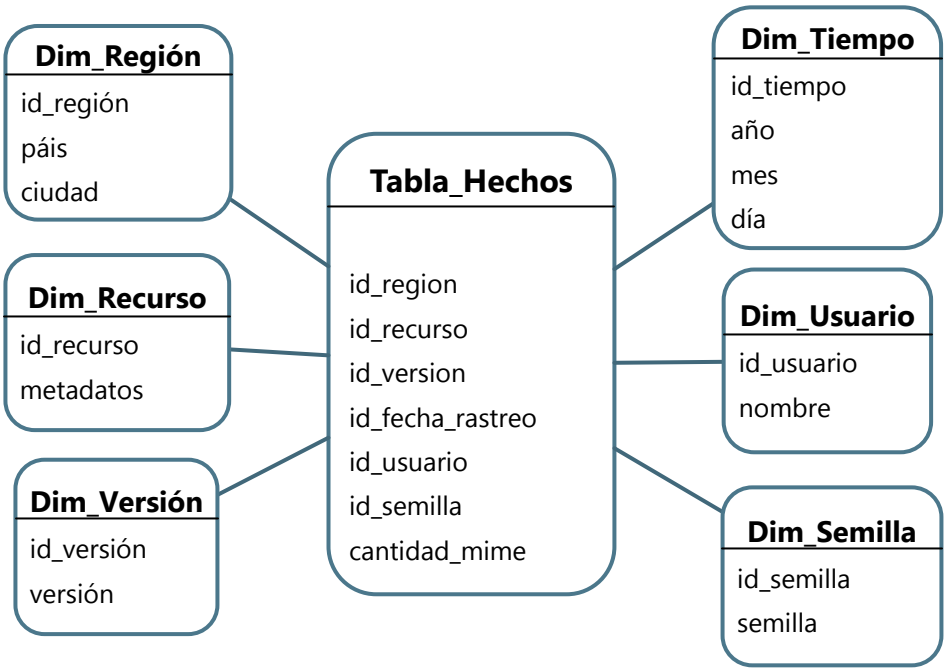


Figura 25 - Modelo Dimensional del Almacén de Datos
Fuente: Elaboración propia

Para implementar el diseño lógico en una base de datos documental, se utiliza el Patrón de modelado de datos embebido, explicado en la sección 2.5.5.1, es decir, al crear cada documento se embebieron los datos pertenecientes a las dimensiones dentro del mismo documento, al inicio de éste, y los metadatos extraídos con WAT Utilities se ubican después de ésta, manteniendo la estructura original en la que se extrajeron, como se puede apreciar en la Figura 26.

Como se mencionó anteriormente (sección 4.2), este proceso de extracción se realiza por cada recurso recolectado, definiendo el nivel de detalle (granularidad) en los recursos recolectados por los rastreos.

Figura 26 - Modelo Dimensional Documental
Fuente: Elaboración propia

En cuanto al diseño físico, se implementa la base de datos en el sistema de bases de datos *NoSQL MongoDB*. Se crea una base de datos llamada "metadatos" y, dentro de ésta, una colección llamada "documentos", en la cual se insertan un conjunto de documentos con datos referentes a los metadatos de los archivos WARC del Archivo Web.

4.5. Proceso ETL

Para generar el contenido del documento descrito en la sección anterior, y almacenar en el repositorio los metadatos de los archivos WARC junto con los datos descriptivos de los mismos, se crea un algoritmo en Java que realiza el proceso ETL, que sigue los pasos mostrados en la Figura 27.

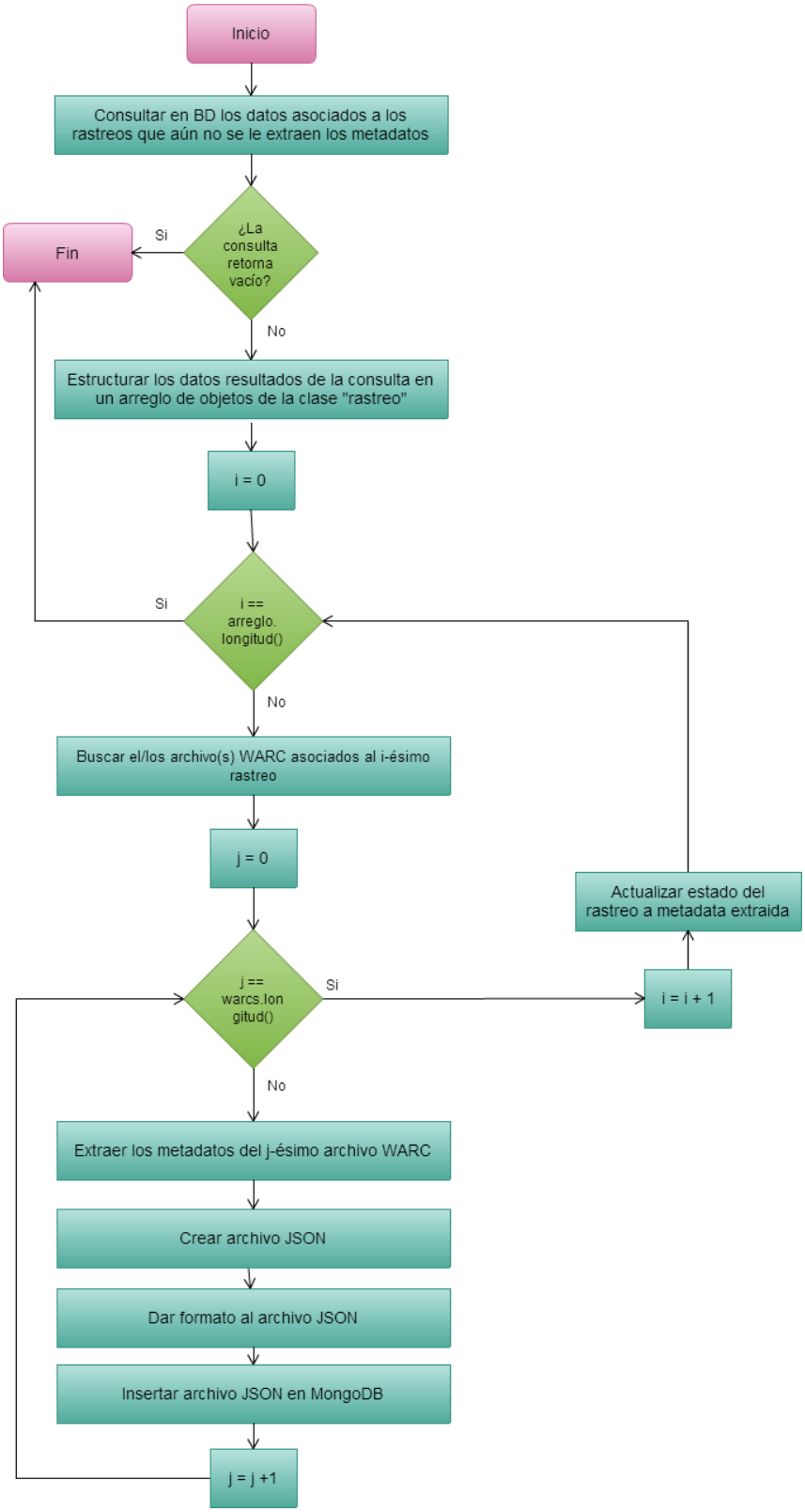


Figura 27 - Diagrama de flujo del Proceso ETL
Fuente: Elaboración propia

El primer paso es realizar la extracción de los datos - correspondientes a los archivos WARC rastreados previamente - de la base de datos de preservación, la cual se encuentra en *MySQL*. Se utiliza el conector *JDBC mysql-connector-java-5.1.35-bin.jar* para poder conectarse y acceder a la base de datos, con el objetivo de ejecutar la sentencia SQL que extrae los datos necesarios, ésta última se puede visualizar en la Figura 28. Algunos de estos datos se encuentran en los *logs*, que son generados por los rastreos de los Archivos WARC, los cuales se accederán y extraerán.

```
SELECT  v.nombre,       -- Nombre version
        t.nombre_job,   -- Nombre job
        t.ciudad,       -- Nombre ciudad
        c.nombre,       -- Nombre coleccion
        u.nombre,       -- Nombre usuario
        u.apellido,     -- Apellido usuario
        p.id,           -- ID
        p.fecha_fin,    -- Fecha rastreo fin
        t.url           -- Semilla
FROM app.app_predictions p
JOIN app.app_traces t
ON (p.traces_id = t.id)
JOIN app.app_versions v
ON (p.versiones_id = v.id)
JOIN app.app_coleccions c
ON (t.coleccion_id = c.id)
JOIN app.app_user u
ON (t.user_id = u.id)
WHERE p.estado = "Finalizado" AND p.extraccion_metadata = 0;
```

Figura 28 - Sentencia SQL que extrae los datos de los archivos WARCs
Fuente: Elaboración propia

Posteriormente, se procede a extraer los metadatos de los Archivos WARC utilizando la biblioteca *WAT Utilities* y el siguiente comando: "java -jar archive-metadata-extractor.jar mywarcfile.warc.gz". Una vez ejecutado, se procede a leer su salida y crear el archivo *JSON*, que próximamente será insertado en el repositorio de datos. La función que realiza la creación del archivo se puede ver en la Figura 29.

```
public void crearArchivo(String ruta, String comando, resultado resultados,int i) throws IOException{
    String json = new String();
    FileWriter file = new FileWriter(ruta);
    Process proceso = Runtime.getRuntime().exec(comando);
    BufferedReader stdInput = new BufferedReader(new InputStreamReader(proceso.getInputStream()));
    while ((json = stdInput.readLine()) != null){
        if(json.indexOf("\"Container\": {") != -1) {
            file.append("\"id\":\""+resultados.getJob()+resultados.getVersion()+"\","
                    + "\"job\":\""+resultados.getJob()+"\","
                    + "\"version\":\""+resultados.getVersion()+"\","
                    + "\"semilla\":\""+resultados.getSemilla()+"\","
                    + "\"tamaño\":\""+resultados.getTamaño()+"\","
                    + "\"ciudad\":\""+resultados.getCiudad()+"\","
                    + "\"coleccion\":\""+resultados.getColeccion()+"\","
                    + "\"usuario\":\""+resultados.getUsuario()+"\","
                    + "\"fechaRastreo\":\""+resultados.getFechaRastreo()+"\","
                    + "\"duracion\":\""+resultados.getDuracion()+"\","
                    + "\"añoD\":\""+resultados.getAñoD()+"\","
                    + "\"mesD\":\""+resultados.getMesD()+"\","
                    + "\"diaD\":\""+resultados.getDiaD()+"\","
                    + "\"horaD\":\""+resultados.getHoraD()+"\","
                    + "\"minutosD\":\""+resultados.getMinutoD()+"\","
                    + "\"segundoD\":\""+resultados.getSegundoD()+"\","
                    + "\"milisegundoD\":\""+resultados.getMilisegundoD()+"\","
                    + "\"añoF\":\""+resultados.getAñoF()+"\","
                    + "\"mesF\":\""+resultados.getMesF()+"\","
                    + "\"diaF\":\""+resultados.getDiaF()+"\","
                    + "\"horaF\":\""+resultados.getHoraF()+"\","
                    + "\"minutosF\":\""+resultados.getMinutoF()+"\","
                    + "\"segundoF\":\""+resultados.getSegundoF()+"\",");
            file.append("\"Container\": {");
        }else
            file.append(json);
    }
    file.close();
}
```

Figura 29 - Función que crear archivo *JSON*
Fuente: Elaboración propia

Finalmente, se le da el formato adecuado al archivo *JSON*, previamente creado, como se puede apreciar en la Figura 30, y se inserta en el repositorio de datos que es implementado en *MongoDB*, el cual fue accedido utilizando el conector *mongo-java-driver-3.0.1.jar*.

```
{
    "id":"CIENCIAS20130618030138",
    "job":"CIENCIAS",
    "version":"20130618030138",
    "semilla":"http://www.ciens.ucv.ve",
    "tamaño":"1584699295",
    "ciudad":"Caracas",
    "coleccion":"Educación",
    "usuario":"MercyOs",
    "fechaRastreo":"18/06/2013-03:01:38",
    "duracion":"7h30m50s966ms",
    "añoD":"0",
    "mesD":"0",
    "diaD":"0",
    "horaD":"7",
    "minutosD":"30",
    "segundoD":"50",
    "milisegundoD":"966",
    "añoF":"2013",
    "mesF":"6",
    "diaF":"18",
    "horaF":"3",
    "minutosF":"1",
    "segundoF":"38",
    "Container":{"Compressed":true,"Filename":"WEB-20130618030214850-00
    "Envelope":{"Actual-Content-Length":"379","Block-Digest":"sha1:YQVD
}
```


Figura 30 - Archivo *JSON*
Fuente: Elaboración propia

4.6. Especificación de Herramientas tecnológicas de Inteligencia de Negocio usadas en la solución

A continuación se explican las distintas herramientas seleccionadas para implementar la solución de Inteligencia de Negocio.

4.6.1. *Pentaho*

Es una plataforma de inteligencia de negocio, orientada a soluciones y centrada en procesos que incluye todos los principales componentes requeridos para implementar soluciones basadas en procesos, para la gestión y toma de decisiones empresariales. Es una plataforma compuesta de diferentes programas que satisfacen los requisitos fundamentales para una solución de inteligencia de negocio. Ofrece soluciones para la gestión y análisis de la información, incluyendo el análisis multidimensional *OLAP*, presentación de informes, minería de datos, creación de cuadros de mando para el usuario, entre otros.

Actualmente, existe una versión comercial y una versión de software libre desarrollada por la comunidad. La plataforma ha sido desarrollada bajo el lenguaje de programación Java y tiene un ambiente de implementación, también basado en *Java*, haciendo de *Pentaho* una solución flexible que cubre una gran cantidad de necesidades empresariales.

4.6.2. *Pentaho Data Integration*

Herramienta que extrae, limpia e integra la información disponible en aplicaciones y bases de datos separadas, y la coloca en manos del usuario, proveyendo consistencia. Esto se debe a que centraliza una versión de todos los recursos de información. Incluye cinco (5) herramientas, las cuales son:

- *Spoon*: herramienta gráfica para diseñar *ETLs* (ver Figura 31).
- *Pan*: ejecuta transformaciones diseñadas en el *Spoon*.
- *Chef*: herramienta para ejecutar trabajos complejos, que automatizan los procesos de actualización de la base datos.
- *Kitchen*: herramienta que ayuda a ejecutar los trabajos por lotes, permitiendo iniciar y controlar fácilmente el proceso *ETL*.
- *Carter*: servidor web que permite la supervisión remota del proceso *ETL*.

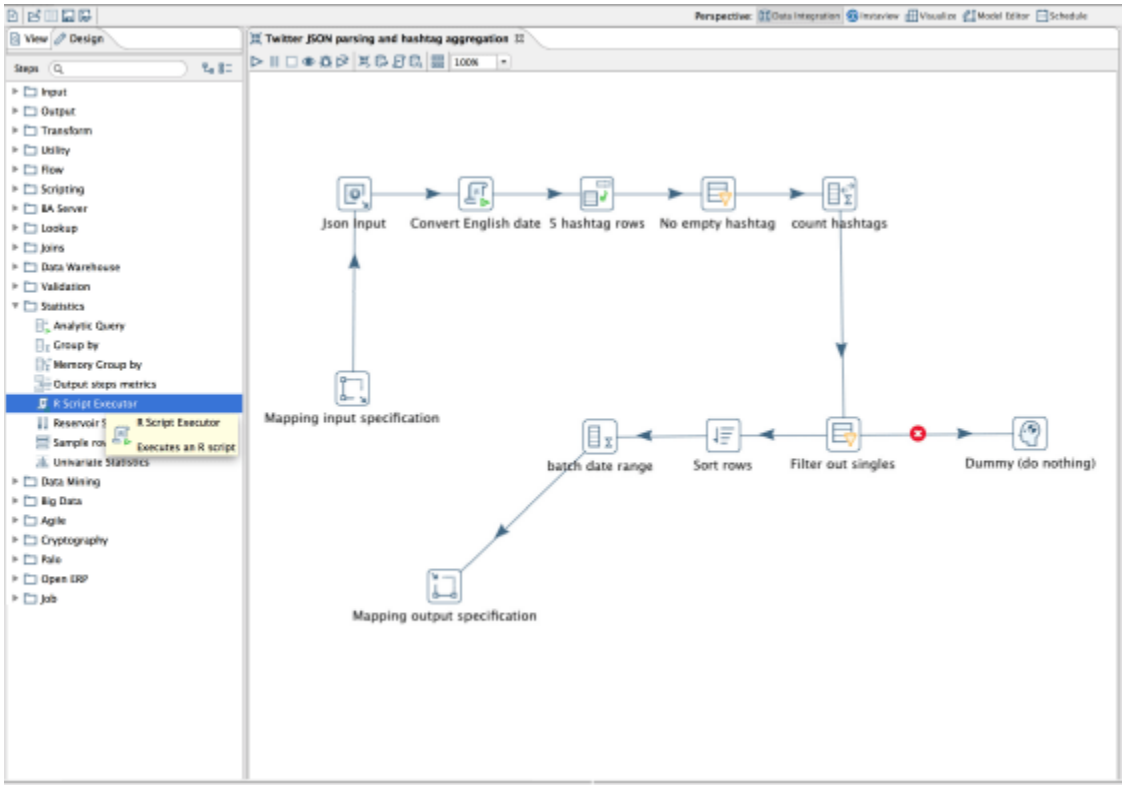


Figura 31 - Vista herramienta *Pentaho Data Integration*
Fuente: (Pentaho, 2015)

Durante el desarrollo de la solución, se utiliza *Pentaho Data Integration* para realizar el proceso de extracción y transformación final de datos, de tal manera que le sean suministrados los datos a las herramientas utilizadas para desarrollar la aplicación de inteligencia de negocio, y así poder generar los indicadores especificados en el primer paso.

4.6.3. *CTools*

Es un conjunto de herramientas y componentes que trabajan sobre *Pentaho*, creado y mantenido por *Webdetails*, para permitir la creación de cuadros de mandos avanzados. Este conjunto de herramienta cuenta con las siguientes características:

- Herramientas multiplataforma de código abierto y de libre uso, para la creación de soluciones de Inteligencia de Negocio.
- Enfocado en *Pentaho* y su comunidad, *CTools* trabaja tanto en *Pentaho Enterprise* como en *Comunnity Editions*.
- Los cuadros de mando de *CTools* se presentan perfectamente en dispositivos de escritorio tradicionales, móviles y largos formatos.
- Toda la interfaz y el ambiente son totalmente personalizables, así puede modificarlo hasta obtener el producto deseado.
- La plataforma de código abierto prospera en participación y colaboración, lo cual significa que todos los usuarios pueden dar sus propios aportes.
- *CTools* está bajo un ciclo de publicación rápido, siendo éste actualizado y revisado cada seis (6) semanas.

Adicionalmente, *CTools* dispone de varios *plugins* que apoyan la elaboración de los indicadores, los cuales son descritos a continuación:

4.6.3.1. *Community Dashboard Framework* (CDF)

Es un *framework* de código abierto que permite la creación de cuadros de mando altamente personalizables, por encima del server *Pentaho Business Intelligence*. CDF está basado en estándares de desarrollo web, como *CSS*[14], *HTML5* y *JavaScript*[15], aprovechando algunos *frameworks* utilizados, como *jQuery* o *Bootstrap*. CDF es una solución que combina datos con una capa de visualización, creado especialmente para los desarrolladores.

4.6.3.2. *Community Data Access* (CDA)

Es un *plugin* de *Pentaho*, diseñado para acceder los datos con una gran flexibilidad. CDA permite el acceso a cualquiera de los muchos datos fuentes de *Pentaho*, y permite:

- Unir diferentes fuentes de datos solo editando el archivo *XML*.
- Hacer cache a consultas, proveyendo un gran mejoramiento en el rendimiento.
- Entregar datos en diferentes formatos (CSV, XLS, etc.) a través de *Pentaho User Console*.
- Organizar y paginar los datos del lado del servidor.
- Ser utilizado, sin preocuparse por los detalles, como un *plugin* independiente (*standalone*) en el Servidor *Pentaho BI* o en una combinación de CDE/CDF.

[14] Lenguaje de hojas de estilo utilizado para describir la presentación de un documento en lenguaje de marcado, como HTML.

[15] Lenguaje de programación de alto nivel, dinámico, débilmente tipado e interpretado, que ha sido estandarizado en ECMAScript.

4.6.3.3. *Community Dashboard Editor* (CDE)

CDE, y toda la tecnología detrás de éste (dígase CDF, CDA y CCC), permiten el desarrollo y lanzamiento de los cuadros de mando avanzados de *Pentaho*. CDE nació para simplificar los procesos de creación, edición y representación de los cuadros de mando *CTools*, y es una herramienta completa y poderosa, que combina *front-end* con fuentes de datos y componentes personalizados transparentemente (ver Figura 32).

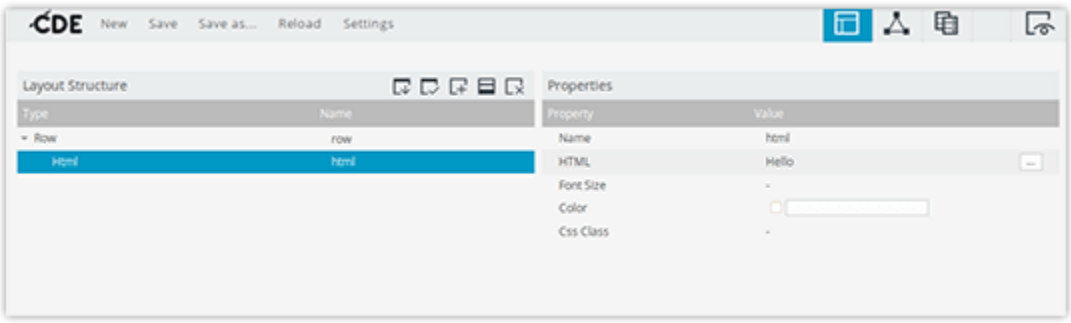


Figura 32 - Vista herramienta *CDE*
Fuente: (Webdetails, 2013)

4.6.3.4. *Community Charts Components* (CCC)

La librería de gráficos *CTools*, la cual fue construida encima de *Protovis*, un conjunto de herramientas de visualización poderosa, gratuita y de código abierto. Su objetivo es proveer el camino a los desarrolladores para incluir en sus cuadros de mandos los tipos básicos de gráficos, sin perder el principio fundamental: extensibilidad.

Debería preferir CCC sobre otros tipos de gráficos, por las propiedades de las gráficas heredadas de *Protovis*, es decir, los gráficos CCC se ven bien, son flexibles y permiten la interacción, además de poseer una inmensa capacidad de personalización.

4.6.4. *Sparkl - Pentaho Application Builder*

Es un *plugin* que permite la creación de otros *plugins*, dentro de *Pentaho*, de manera más sencilla, ya que anteriormente era necesario tener conocimiento sobre Java para poder desarrollarlos. Con el uso de *Sparkl* es necesario tener conocimiento sobre *CTools* y *Pentaho Data Integration* para poder construir aplicaciones usables para sus consumidores o clientes (ver Figura 33).

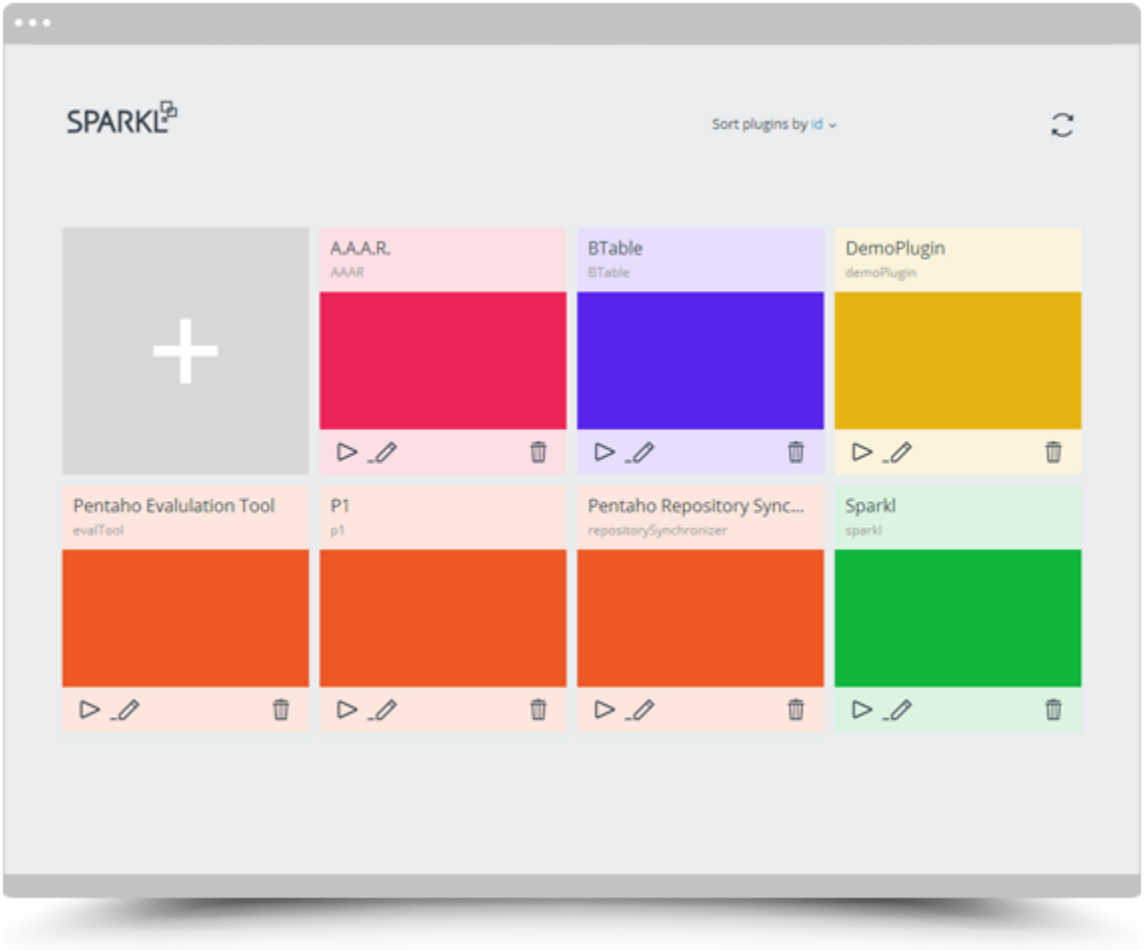


Figura 33 - Vista creación *plugins* en *Sparkl*
Fuente: (Webdetails, 2013)

Hay dos (2) secciones cuando se edita una aplicación *Sparkl*, metadatos asociados al *plugin* y la definición de elementos (ver Figura 34). El grupo de elementos está básicamente compuesto por dos (2) tipos: cuadros de mando (*dashboards*), para el *front-end* de la aplicación, y puntos finales (*endpoints*), para la lógica del *back-end*, explicados a continuación:

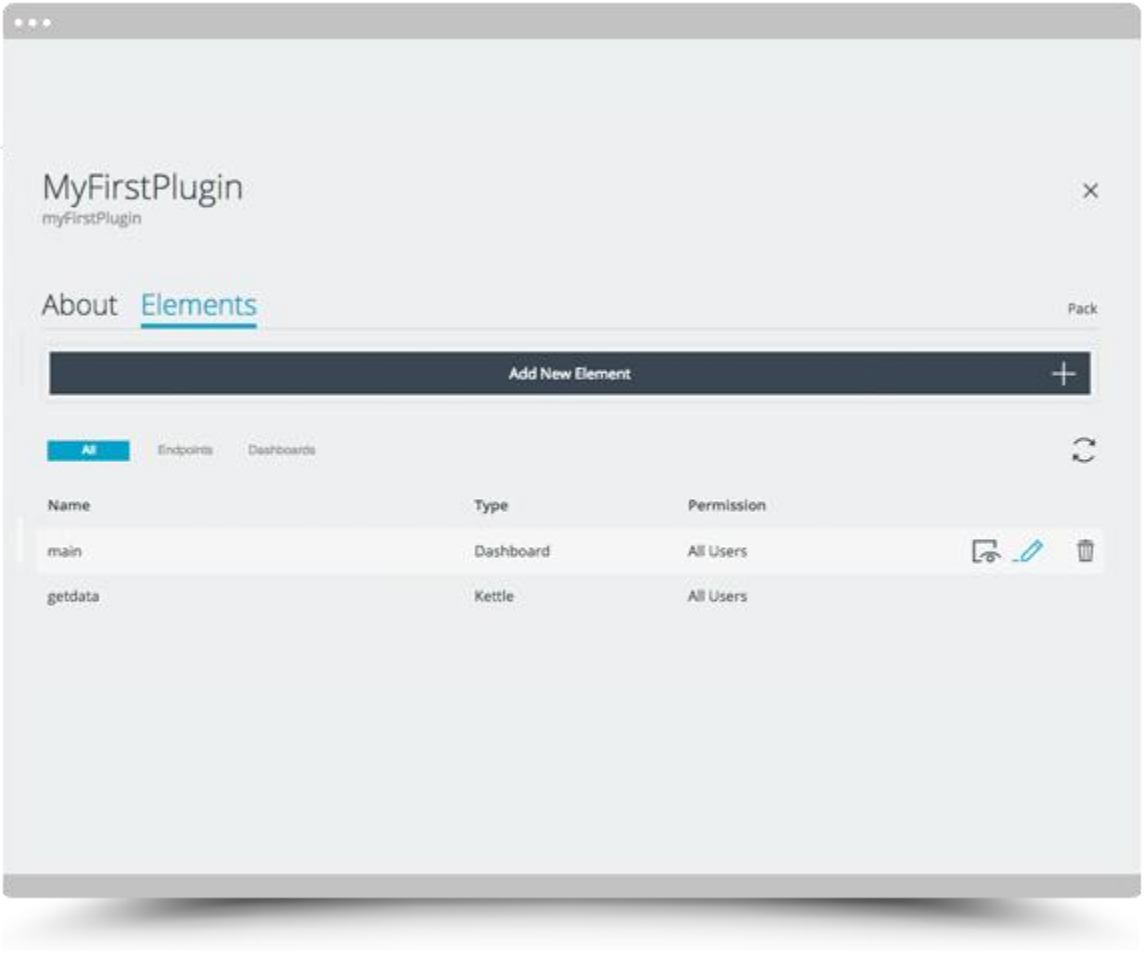


Figura 34 - Vista *Elements* de *Sparkl*
Fuente: (Webdetails, 2013)

4.6.4.1. Cuadros de mando (*Dashboards*)

Filtrando los tipos de elemento *dashboard*, podemos ver cuantas interfaces tiene el *plugin*. Las interfaces de usuarios son generadas a través de CDE, y podemos ir de *Sparkl* a CDE si se desea editar el *dashboard*, al igual que si deseamos crear un nuevo *dashboard*.

4.6.4.2. Puntos finales (*Endpoints*)

Son los encargados de gestionar las fuentes de datos, además de la lógica asociada a su procesamiento y trasformación, con el fin de poder visualizarlas a través del *dashboard*. Todo ésto es posible a través del uso de *Pentaho Data Integration*.

Durante el desarrollo de la solución, todas estas herramientas fueron integradas (ver Figura 35) para poder generar los indicadores definidos en la sección 4.1. Inicialmente, se utiliza el *plugin Sparkl* junto con la herramienta *Kettle* de *Pentaho Data Integration*, permitiendo generar todos los *endpoints* necesarios, y que van a ser utilizados, por los indicadores. Seguidamente, estos *endpoints* son utilizados por las

herramientas que componen *CTools* (*CCC*, *CDE* y *CDA*) a través de *Sparkl*, para poder desarrollar los indicadores con sus gráficos correspondientes.

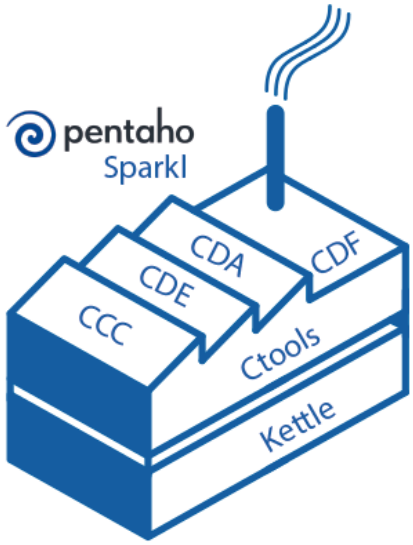


Figura 35 - Herramientas utilizadas en la Solución BI
Fuente: Elaboración propia adaptado de http://goo.gl/KX28CT

4.7. Desarrollo de la Aplicación de Inteligencia de Negocio

Para el desarrollo de la solución, el primer paso fue definir como se visualizarían los indicadores que la aplicación se encargará de calcular, basándonos en los usuarios que se mencionan en el Capítulo 1, especialmente los Administradores y el Director. Como planteó en el Capítulo 3, específicamente en la actividad correspondiente al Desarrollo de la Aplicación de Inteligencia de Negocio, ésta se realizó siguiendo un diseño basado en prototipos, ágil, iterativo e incremental.

Para el funcionamiento de la solución, fue necesario el desarrollo de varios procesos ET (extracción y transformación de los datos), dado que la herramienta a utilizar para el desarrollo de la solución requiere que los datos tengan un formato estructurado, tal como una tabla de *Excel*, característica que el repositorio donde se almacenan los datos no posee. Es necesario recalcar que los ET están fuertemente vinculados a los datos que se visualizan en la interfaz, por lo tanto, cada vez que se definen o modifican los filtros o indicadores, el ET asociado debe modificarse o crearse igualmente. Para crear los ET, se utilizó *Pentaho Data Integration*.

Es importante recalcar que, durante todo el proceso de desarrollo de la solución, fue necesario reunirse continuamente con los usuarios finales de la aplicación. En nuestro caso, y por razones de disponibilidad, las reuniones se llevaron a cabo únicamente con el usuario Director. Con éste se discutieron aspectos de diseño y usabilidad, así como también ideas para el desarrollo de la solución.

Se realizaron tres (3) prototipos principales, los cuales se describen brevemente a continuación:

4.7.1. Desarrollo del primer prototipo

En este prototipo se realizaron los primeros tres (3) procesos ET, con el fin de poder mostrar los resultados iniciales asociados al desarrollo de la solución. En la Figura 36, se detallan los pasos asociados a un ET que alimenta a un indicador.

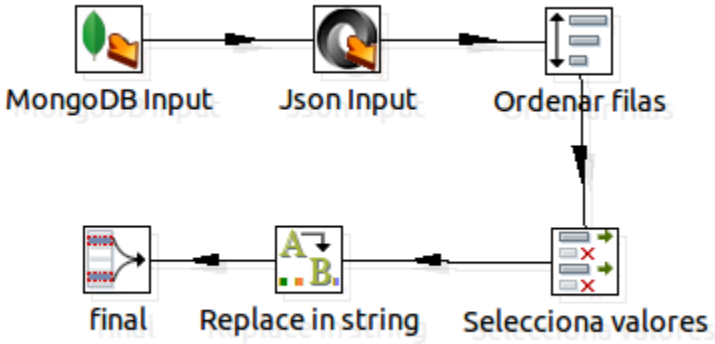


Figura 36 - ET correspondiente al Indicador Cantidad de *MIME Type* (Prototipo 1)
Fuente: Elaboración propia

Se utiliza el *plugin* CDE de *Pentaho* para desarrollar la interfaz de usuario, con el fin de mostrar dos (2) gráficos que se construyen a partir de los datos de los indicadores implementados en este prototipo, como se puede visualizar en la Figura 37. Además, en esa misma figura se puede observar un filtro, el cual provee un parámetro para la búsqueda de los datos de los indicadores antes mencionados.

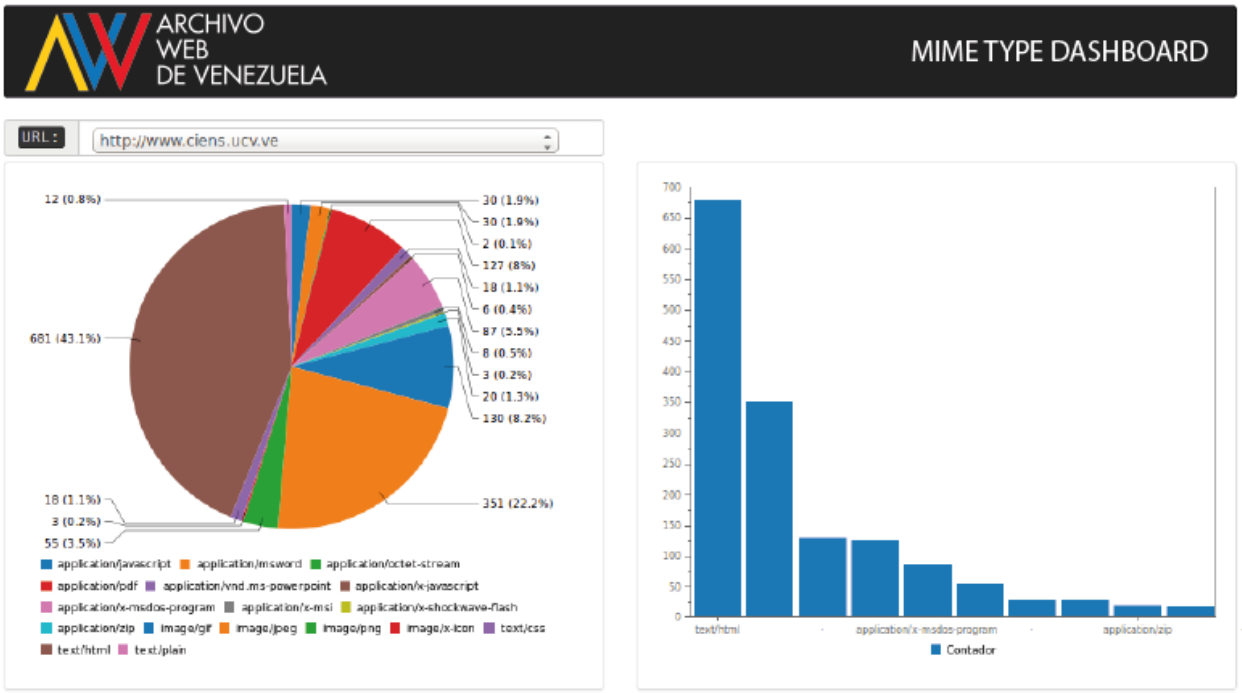


Figura 37 - Cuadro de mando de Cantidad de *MIME Type* (Prototipo 1)
Fuente: Elaboración propia

4.7.2. Desarrollo del segundo prototipo

En este prototipo, se generaron los diferentes cuadros de mando relacionados al resto de los indicadores. Para ésto, se utilizan las operaciones *slice* y *dice*[16], lo que requirió definir las dimensiones involucradas, además de elaborar sus respectivos ET.

El indicador del prototipo anterior fue modificado, debido a que no cumplía con los requerimientos del usuario. Este cambio se puede visualizar en la Figura 38.

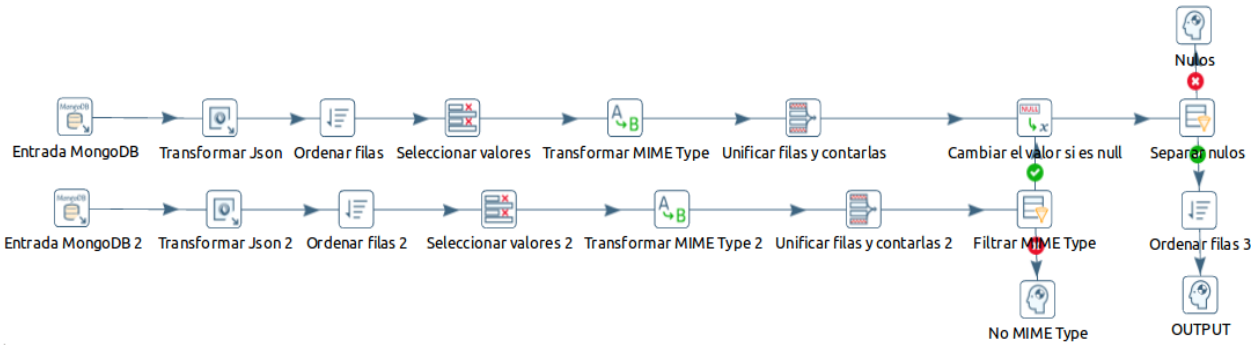


Figura 38 – ET correspondiente al indicador Cantidad de *MIME Type* (Prototipo 2)
Fuente: Elaboración propia

[16] Estrategia para segmentar, visualizar y entender los datos cortando un segmento largo de datos en partes más pequeñas hasta alcanza el nivel de detalle correcto para el análisis.

Como se puede observar en la Figura 39, se modificó el filtro, ya que se necesitaba otro tipo de parámetro diferente al anteriormente utilizado. En esta figura, también se puede visualizar un cambio en el estilo de los gráficos de la interfaz de la solución del prototipo.

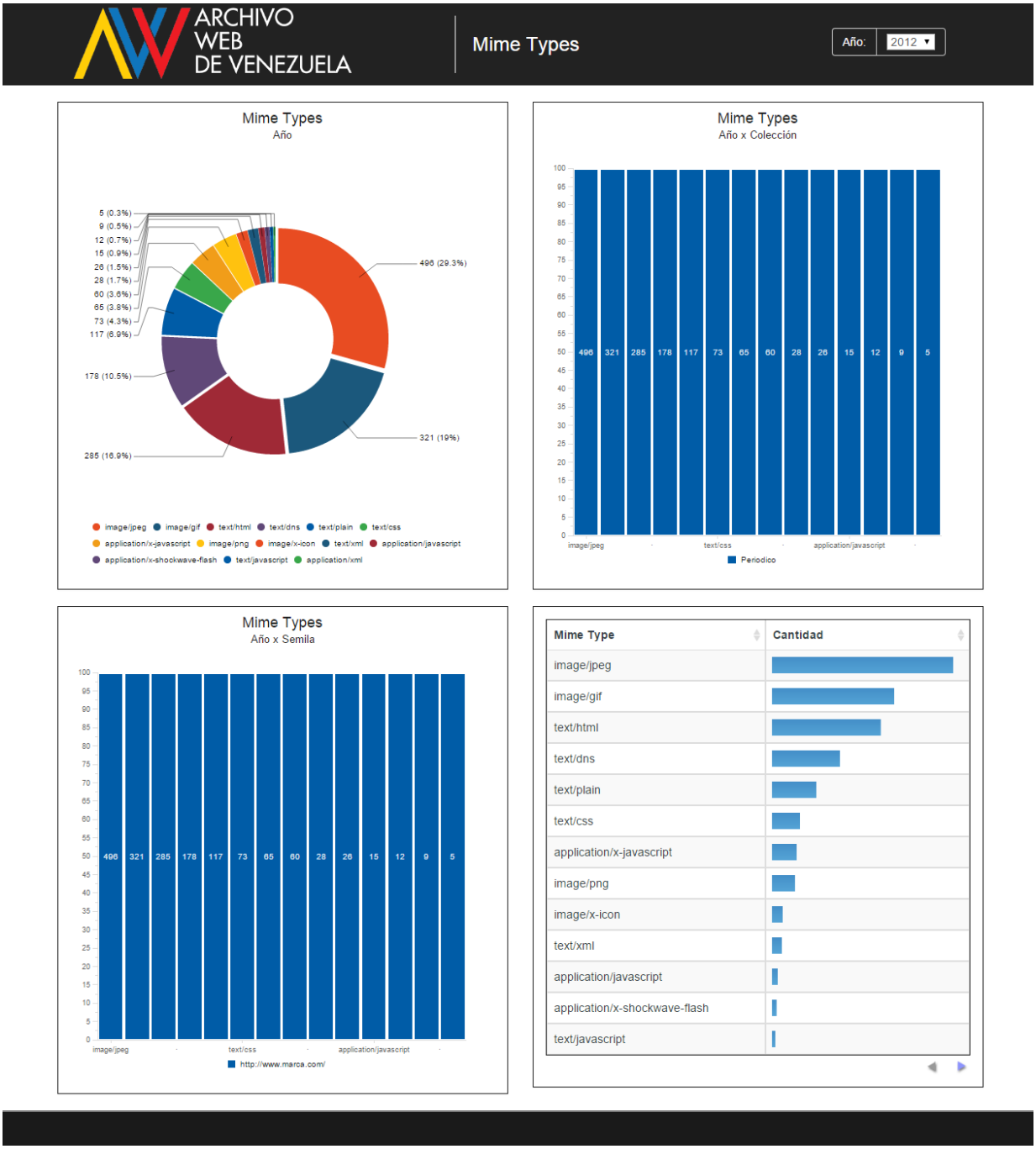
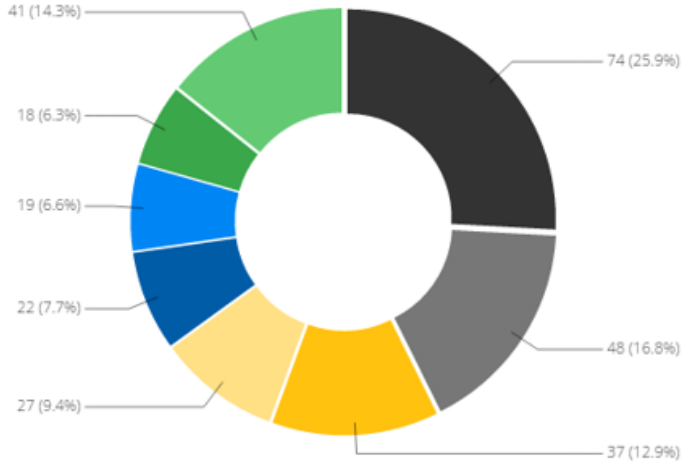


Figura 39 - Cuadro de mando de Cantidad de *MIME Type* (Prototipo 2)
Fuente: Elaboración propia

Dicho cambio se debe a la utilización de código *Javascript* adicional, el cual se puede apreciar del lado derecho en la Figura 40, mientras que del lado izquierdo podemos ver el resultado final del gráfico.




```
new pvc.PieChart({
    canvas: 'cccPieExample4',
    width:  600,
    height: 400,

    // Data source
    crosstabMode: false,

    // Main plot
    valuesVisible: true,
    valuesFont: 'lighter 11px "Open Sans"',
    explodedSliceRadius: '2%',
    slice_innerRadiusEx: '50%',

    // Panels
    legend: false,

    // Chart/Interaction
    selectable: true,
    hoverable:  true,
    tooltipClassName: 'light',

    // Color axes
    colors: [
        '#333333', '#777777', '#FFC20F', '#FFE085',
        '#005CA7', '#0086F4', '#39A74A', '#63CA73'
    ]
})
.setData(relational_03_b)
.render();
```

Figura 40 – Modificación del gráfico de torta y código asociado
Fuente: Elaboración propia

4.7.3. Desarrollo del tercer y último prototipo

En este prototipo se integraron los cuadros de mando, utilizando el *plugin Sparkl* de *Pentaho*. Esta modificación se debió a que los cuadros de mando construidos con el *plugin* CDE, en el prototipo anterior, eran independientes unos de los otros. Además, cada cuadro de mando desarrollado con CDE era un archivo independiente, el cual *Pentaho* no es capaz de reconocer como una aplicación. Es por ésto que, con el fin de que *Pentaho* reconociera esta solución como una única aplicación donde se pudieran visualizar los distintos indicadores, se construyeron los mismos con *Sparkl*.

Para conectar los distintos cuadros de mando se implementó un menú, el cual agrupa los indicadores según su tema. Se incorporaron, a la interfaz, botones que permiten exportar el cuadro de mando en formato PDF y exportar los datos asociados a un gráfico en los formatos XLS o CSV, lo cual se puede observar en la Figura 41.

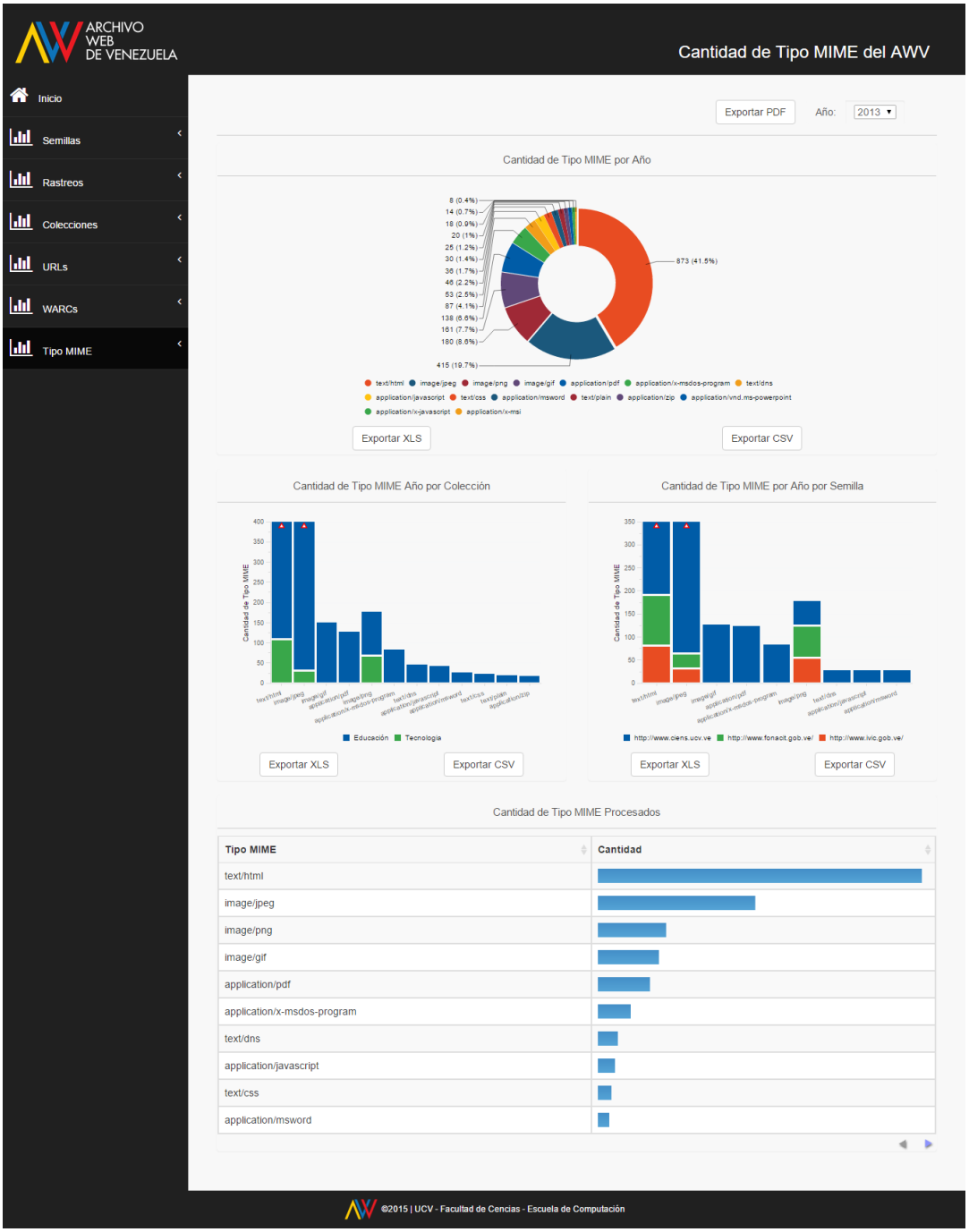


Figura 41 - Cuadro de mando de Cantidad de *MIME Type* – Anual (Prototipo 3)
Fuente: Elaboración propia

En las interfaces que tienen una gran cantidad de valores, se tomaron los primeros N valores más significativos, con el fin de mejorar la legibilidad de los gráficos que este prototipo permitirá visualizar. Además, se desarrollan interfaces para cada indicador como vistas generales de los mismos, es decir, sin ningún filtro o parámetro, como se puede ver en la Figura 42.

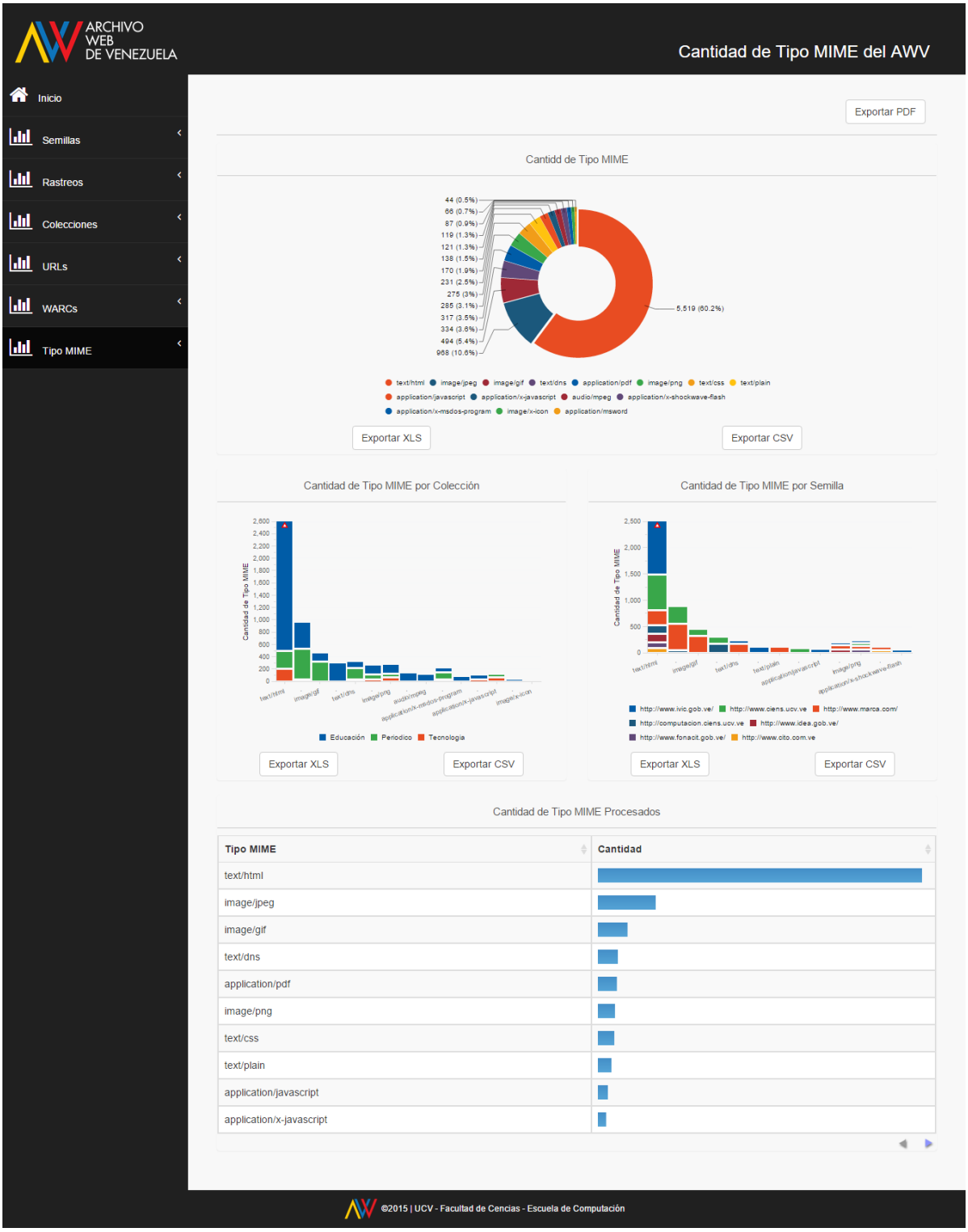


Figura 42 - Cuadro de mando de Cantidad de *MIME Type* – General (Prototipo 3)
Fuente: Elaboración propia

4.8. Pruebas

Con el fin de verificar que los datos que se visualizan en los gráficos de los cuadros de mando son correctos, nos dirigimos al directorio *Reports*, el cual se encuentra dentro del directorio generado al finalizar el rastreo de una página web.

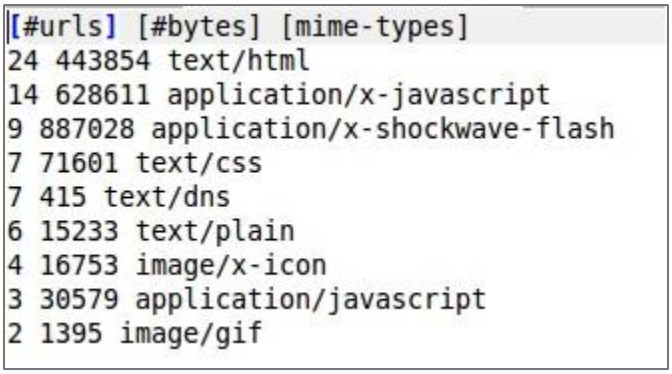


Figura 43 - Reporte generado por *Heritrix* de *MIME types*
Fuente: Elaboración propia

Dentro de este directorio, se encuentran un conjunto de reportes asociados a dicho rastreo, como el que se aprecia en la Figura 43, y estos reportes contienen el detalle de lo que se puede apreciar en los cuadros de mando, es decir, para verificar que los datos sean correctos, se suman los detalles de los reportes y se compara el resultado con el que se muestra en el gráfico correspondiente.

Al realizar la comparación entre los datos contenidos en los reportes generados por Heritrix y los resultados obtenidos en los indicadores, se observa que la información es consistente, es decir, el valor mostrado en los indicadores es igual a los obtenidos por la herramienta Heritrix.

En la Tabla 21, se pueden visualizar los reportes relacionados al *MIME Type*, generados por la herramienta Heritrix por de cada rastreo realizado en el año 2012, y la suma de sus totales. Adicionalmente, en la Tabla 22, se pueden apreciar los datos que se obtienen de los indicadores Cantidad de Tipo *MIME* y Costo en Bytes de los Tipo *MIME*, correspondientes a los rastreos del año 2012. Comparando estas tablas, se puede concluir que la mayoría de los datos tienen el mismo valor, por lo que la calidad de los datos es aceptable.

Tabla 21 - Reportes *MIME Type* del año 2012 generados por Heritrix

Nombre	Cantidad	Tamaño en Bytes
Rastreo 1		
image/jpeg	463	16067445
image/gif	307	1362007
text/html	245	6584281
text/dns	105	14338
text/plain	74	77731

Tabla 21 - Reportes *MIME Type* del año 2012 generados por Heritrix

Nombre	Cantidad	Tamaño en Bytes
text/css	70	1056264
application/x-javascript	57	1355905
image/png	52	616247
text/xml	26	2095200
image/x-icon	23	234311
application/javascript	13	403373
application/x-shockwave-flash	10	355056
text/javascript	7	237639
application/xml	4	1895
Rastreo 2		
text/dns	72	10001
text/plain	45	46155
text/html	41	1090745
image/jpeg	33	1581232
image/gif	14	43574
application/x-javascript	8	329755
image/png	8	184485
image/x-icon	5	30637
text/css	3	31292
application/javascript	2	31568
application/x-shockwave-flash	2	58772
text/javascript	2	83813
application/xml	1	435
Rastreo 3		
text/dns	1	56
Total		
image/jpeg	496	17648677
image/gif	321	1405581
text/html	286	7675026

Tabla 21 - Reportes *MIME Type* del año 2012 generados por Heritrix

Nombre	Cantidad	Tamaño en Bytes
text/dns	178	24395
text/plain	119	123886
text/css	73	1087556
application/x-javascript	65	1685660
image/png	60	800732
text/xml	26	2095200
image/x-icon	28	264948
application/javascript	15	434941
application/x-shockwave-flash	12	413828
text/javascript	9	321452
application/xml	5	2330

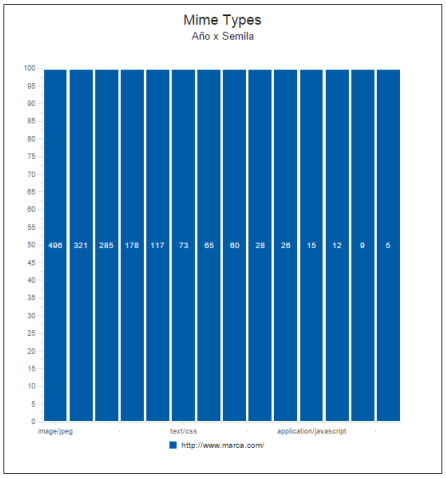
Fuente: Elaboración propia

Tabla 22 - Indicador Cantidad de *MIME Type* y Costo en *bytes MIME Type* del 2012

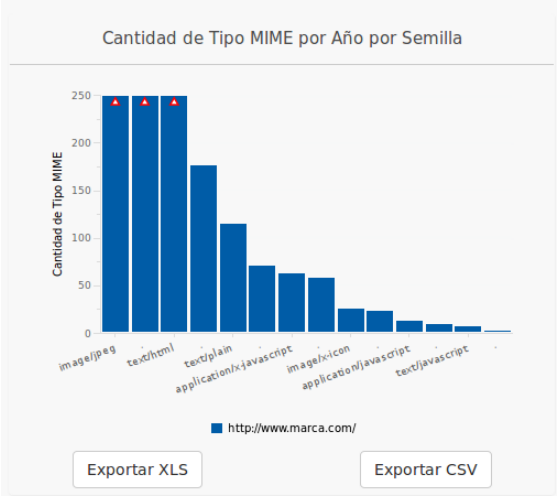
Nombre	Cantidad	Tamano en Bytes
image/jpeg	496	17648677
image/gif	321	1405581
text/html	285	7620959
text/dns	178	24395
text/plain	117	120516
text/css	73	1087556
application/x-javascript	65	1685660
image/png	60	800732
image/x-icon	28	264948
text/xml	26	2095200
application/javascript	15	434941
application/x-shockwave-flash	12	413828
text/javascript	9	321452
application/xml	5	2330

Fuente: Elaboración propia

Además, se realizaron varias reuniones con el usuario final para realizar pruebas de aceptación, donde fueron revisadas y validadas las presentaciones de los indicadores, para garantizar que éstos representaran información útil a los usuarios, entre estos se solicitó el cambio de la representación de los indicadores que tenían una gran diferencia entre los valores mayores y los menores como se muestra en la Figura 44 donde se observa la representación anterior (a) y la solicitada por el usuario (b).



a) Representación anterior



b) Representación solicitada

Figura 44 – Representación de los Indicadores en la prueba de aceptación

Conclusiones y Recomendaciones

Al culminar este Trabajo Especial de Grado (T.E.G.), se desarrolló exitosamente una solución de Inteligencia de Negocio para el Prototipo de Archivo Web de Venezuela, utilizando los metadatos ya almacenados en el prototipo y analizando las herramientas disponibles en el mercando para poder construir la solución acorde a la cantidad de metadatos almacenados y a las necesidades de los usuarios del Prototipo.

Debido al uso de elementos que no son tradicionales en el ámbito de las soluciones de Inteligencia de Negocio, fue necesario realizar una adaptación de la metodología de Kimball con la incorporación de elementos, como el modelo dimensional documental, que nos permitieron establecer una serie de actividades que nos guiaron a través del diseño y construcción de la solución de inteligencia de negocio, pudiendo avanzar rápida y ordenadamente durante el desarrollo, dando como resultado un producto usable y que satisface las necesidades de los diferentes usuarios definidos, además de hacer al Prototipo un sistema que usa tecnologías de punta, debido a que a la fecha, no se encontró información de otros Archivos Web que implementen indicadores utilizando como fuente los metadatos generados por los rastreos, así como tampoco que integren a su Archivo Web una sección de indicadores sobre los diferentes procesos del sistema.

Esta solución de Inteligencia de Negocio ofrece a los usuarios del Prototipo una serie de indicadores que los ayudará a tomar decisiones objetivas sobre el Prototipo, así como también facilitar la tarea de definir planes de preservación a lo largo del tiempo, ya que estos indicadores muestran el estado actual del Prototipo y su desarrollo a través del tiempo.

Durante el desarrollo de la solución, se presentaron algunas limitantes al momento de construir la solución. La primera fue la escasa documentación de las diferentes tecnologías utilizadas durante el desarrollo, en específico las herramientas *CTools* y *Sparkl*. En el caso de los WARCs, también se presentó una limitante por la poca documentación sobre su formato. Esto ocasionó que, al inicio, tuviéramos una curva de aprendizaje un poco elevada, pero cuando se comprendió el uso de las distintas herramientas, la conexión entre ellas y se realizaron los primeros indicadores, el desarrollo de la solución fue más rápido.

También, se consiguieron limitantes al momento de probar el correcto funcionamiento de la aplicación. Durante éste proceso, se pudo notar que la carga inicial de los indicadores es muy lenta, sin embargo, una vez consultado el indicador durante la sesión, éste es almacenado en caché, por lo que su visualización es más rápida. A su vez, cuando hay un gran volumen de datos y la diferencia entre los grupos de datos es elevada, los valores más pequeños no tienen una representación correcta de su verdadero valor, ya que éstos son considerados como poco significativos o nulos, generando cambios en la definición inicial de los indicadores ya que se tuvo que tomar una cantidad determinada de valores a ser mostrados, es decir, un Top N.

A través de pruebas de aceptación con el usuario, realizadas durante el desarrollo de la solución, se obtuvieron diversos comentarios y sugerencias sobre la aplicación, así como también un gran interés por parte de los usuarios, lo cual generó motivación adicional para entregar un producto con altos estándares y que agregara valor al novedoso proyecto que es el Prototipo de Archivo Web, objetivo que fue logrado una vez finalizadas las pruebas de aceptación con los usuarios.

En el documento ISO/TR 14873:2013, se definen una vasta cantidad de indicadores y datos estadísticos que permiten medir el funcionamiento y los diferentes procesos involucrados en el archivado web, pero la información almacenada actualmente en el Prototipo de Archivo Web no es suficiente para poder generar todos los indicadores definidos en dicho documento. Como trabajos futuros, se plantea continuar expandiendo y creando módulos para el Prototipo de Archivo Web de Venezuela, que permitan almacenar información referente a la gestión del Prototipo o la gestión del personal relacionado al Archivo Web, así como también se recomienda almacenar información relacionada al portal de acceso, es decir, información sobre las visitas realizadas por usuarios registrados o la actividad realizadas por éstos; de esta manera se pueden generar indicadores adicionales que ya han sido definidos, y que permitirán al Prototipo seguir creciendo e innovando en el área.

En la aplicación actualmente no se pueden crear reportes bajo demanda de los usuarios, por lo que se propone como trabajo futuro, seleccionar tecnologías que permitan construir cubos lógicos, para poder crear estos reportes.

Para finalizar, la información cosechada (metadatos) y los indicadores generados a partir de ellos, permiten facilitar la gestión del Archivo Web, pero hasta el momento no se sabe con certeza si se requiere tal nivel de detalle o si es económicamente factible, ya que aún ningún Archivo Web ha llegado al siguiente paso: la ejecución de estrategias efectivas para la migración y emulación de recursos en formatos obsoletos. Mientras tanto, los usuarios directores y administradores pueden desarrollar estrategias, definir objetivos y métodos para cumplir el objetivo de preservar los recursos de los Archivos Web, y para esto se apoyarán en los indicadores desarrollados en esta solución de Inteligencia de Negocio.

Bibliografía

Acosta, A. E. (13 de Diciembre de 2005). AgilUs: Construcción ágil de la Usabilidad. Caracas, Distrito Capital, Venezuela. Obtenido de Genasig: http://www.ciens.ucv.ve:8080/genasig/sites/interaccion-humano-comp/archivos/234_CLEI_Acosta_Paper.pdf

AEC, A. E. (2013). *AEC*. Obtenido de http://www.aec.es/web/guest/centro-conocimiento/indicadores

Beck, K., Beedle, M., Bennekum, A. v., Cockburn, A., Cunningham, W., Fowler, M., . . . Thomas, D. (2001). *Principles behind the Agile Manifesto*. Obtenido de Manifesto for Agile Software Development: http://www.agilemanifesto.org/iso/en/principles.html

Beltrán, J. (2006). *Indicadores de Gestión: Herramientas para lograr la competitividad.* Bogotá: 3R Editores.

Burner , M., & Kahle, B. (1996). *Internet Archive*. Obtenido de http://archive.org/web/researcher/ArcFileFormat.php

Cano, J. L. (2007). *Business Intelligence: Competir con Información.*

CCSDS. (2002). *Reference Model for an Open Archival Information System.* Washington DC.

Cho, J., & Garcia-Molina, H. (2000). The Evolution of the Web and Implications for an Incremental Crawler. Stanford, California, Estados Unidos de América.

Consultative Committee for Space Data Systems. (Enero de 2012). Reference Model for an Open Archival Information System. Washington, DC, USA.

Francia, S. (2012). *MongoDB and PHP*. Estados Unidos de América: O'Reilly Media, Inc. Obtenido de https://www.safaribooksonline.com/library/view/mongodb-and-php/9781449324827/ch01.html

Goel, V. (02 de Abril de 2011). *Web Archive Metadata File Specification*. Obtenido de Internet Research: https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Metadata+File+Specification

IIPC. (2012). *Netpreserve*. Obtenido de Netpreserve: http://netpreserve.org/

Inmon, W. H. (1996). *Building the Data Warehouse.*

ISO 28500. (2009). *ISO.* Obtenido de https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-1:v1:en

ISO Working Group. (15 de 06 de 2001). *ISO/IEC 9126-1:2001.* Obtenido de International Organization for Standardization: http://www.iso.org/iso/catalogue_detail.htm?csnumber=22749

ISO Working Group. (02 de Octubre de 2012). *Information and documentation — Statistics and Quality Indicators for Web Archiving.* Obtenido de International Internet Preservation Consortium: http://netpreserve.org/sites/default/files/resources/SO_TR_14873__E__2012-10-02_DRAFT.pdf

ISO/TR 14873. (2013). *Draft Report ISO TR 14873.*

Jack, P., & Binns, A. (2012). *Web Archive - Heritrix*. Obtenido de https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

Jimeno Bernal, J. (13 de Mayo de 2013). *PDCA Home*. Obtenido de http://www.pdcahome.com/4501/gestion-de-procesos-como-definir-indicadores-y-cuadros-de-mando/

Kabchi, M., & Martínez, M. (2013). *Definición de las estrategias para el desarrollo del Módulo de Acceso a los Contenidos Web Preservados en formato WARC para el Prototipo de Archivo Web para la Preservación del Patrimonio Web de Venezuela.* Caracas.

Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit*.

Laudon , K. C., & Laudon, J. P. (2012). *Sistemas de Información Gerencial.* New York: Always Learning Pearson.

Lavoie, B. (2004). *The Open Archival Information System Reference Model: Introductory Guide.* Ohio: OCLC Online Computer Library Center.

Llull P., J. (2005). Evolución del concepto y de la significación social del patrimonio cultural. *Arte, Individuo y Sociedad, 17*, 175-204.

Loshin, D. (2012). *Business Intelligence: The Savvy Manager's Guide.* Waltham: Morgan Kaufmann.

Masanès, J. (2006). *Web Archive.* New York: Springer.

Mondragón, A. (2014). *¿Qué son los indicadores?*

MongoDB, Inc. (03 de Marzo de 2015). *Model Relationships Between Documents*. Obtenido de The MongoDB 3.0 Manual: http://docs.mongodb.org/master/applications/data-models-relationships/

MongoDB, Inc. (2015). *NoSQL Database Explained*. Obtenido de mongoDB: https://www.mongodb.com/nosql-explained

National Information Standards Organization. (2004). *Understanding Metadata.* Bethesda, Maryland, United States of America: NISO Press.

Ospina Torres, M. H. (2014). *Un Marco de Referencia para la Implementación de Archivos Web.* Universidad Central de Venezuela. Caracas: .

Ospina, M. (Septiembre de 2011). Modelo de Archivo de la Web para Venezuela. Caracas, Distrito Capital, Venezuela.

Ospina, M., Martinez, M., Kabchi, M., & León, C. (2014). *Desarrollo de una Aplicación para Acceder a Contenidos de un Archivo Web en Formato WARC.* Caracas.

Ponniah. (2001). *DataWarehousing Fundamentals.* USA: John Wiley & Sons, Inc.

RAE. (2001). *Real Academia Española*. Obtenido de http://www.rae.es/

Rebiun. (s.f.). *Informe del objetivo operacional 1.2.1: Guía de Recursos para la Preservación Digital.* España.

Rivero, L., & García, J. (2013). *Implementación de los móulos de adquisición y almacenamiento de un prototipo para el archivado de sitios Web en Venezuela.* Caracas.

Roe, C. (01 de Marzo de 2012). *ACID vs. BASE: The Shifting pH of Database Transaction Processing*. Obtenido de Dataversity Education: http://www.dataversity.net/acid-vs-base-the-shifting-ph-of-database-transaction-processing/

Sadalage, P. (3 de Octubre de 2014). *NoSQL Databases: An Overview*. Obtenido de ThoughtWorks: http://www.thoughtworks.com/insights/blog/nosql-databases-overview

Stern, H. (31 de Mayo de 2011). *Web Archive Transformation (WAT) Specification, Utilities, and Usage Overview*. Obtenido de Web Archive Jira: https://webarchive.jira.com/wiki/plugins/servlet/mobile#content/view/14484029

UNESCO. (Marzo de 2003). Directrices para la Preservación del Patrimonio Digital. Australia: Preparado por la Biblioteca Nacional de Australia.

Webdetails. (s.f.). *Webdetails a pentaho company*. Obtenido de Webdetails a pentaho company: http://www.webdetails.pt/

Anexo A Gráficos de los Indicadores implementados

A continuación, se presentan los gráficos desarrollados dentro de la solución de Inteligencia de Negocio, así como la pantalla inicial de la aplicación de usuario. Los gráficos serán mostrados a nivel general y después por tiempo, de acuerdo al cálculo del indicador al cual representan.

1. Página inicial

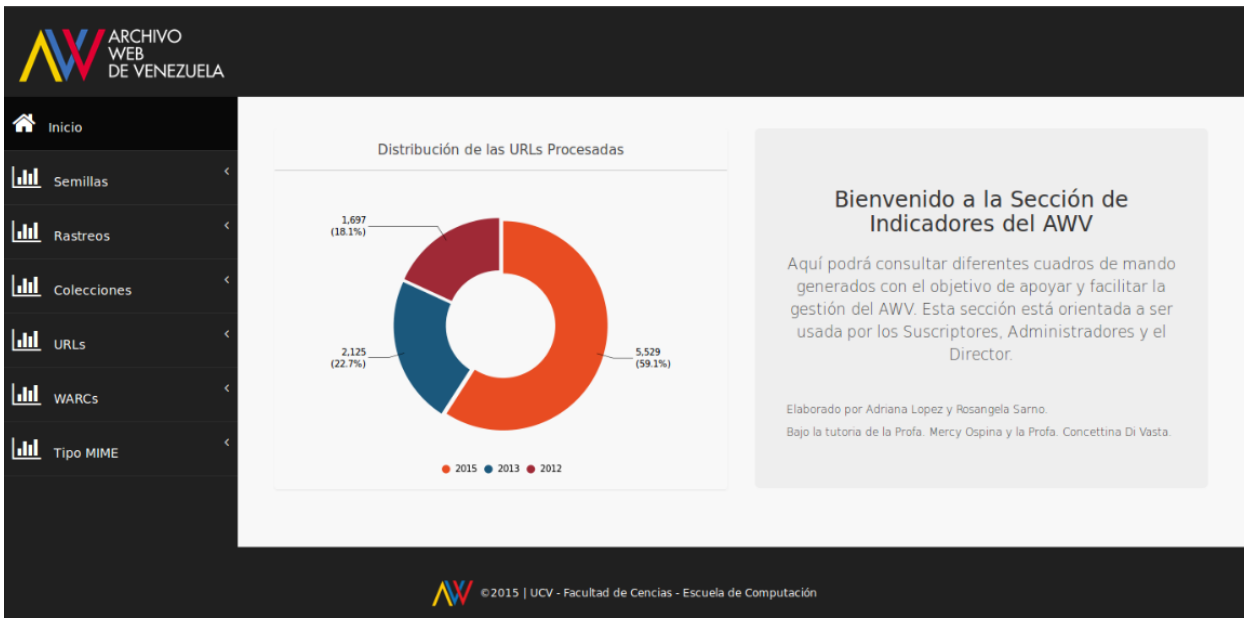


Figura 45 - Vista página Inicio de la Aplicación de Usuario
Fuente: Elaboración propia

2. Cantidad de semillas

Formula	Unidad	Criterio de clasificación	Representación
Conteo de semillas	#	Por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)

2.1.General

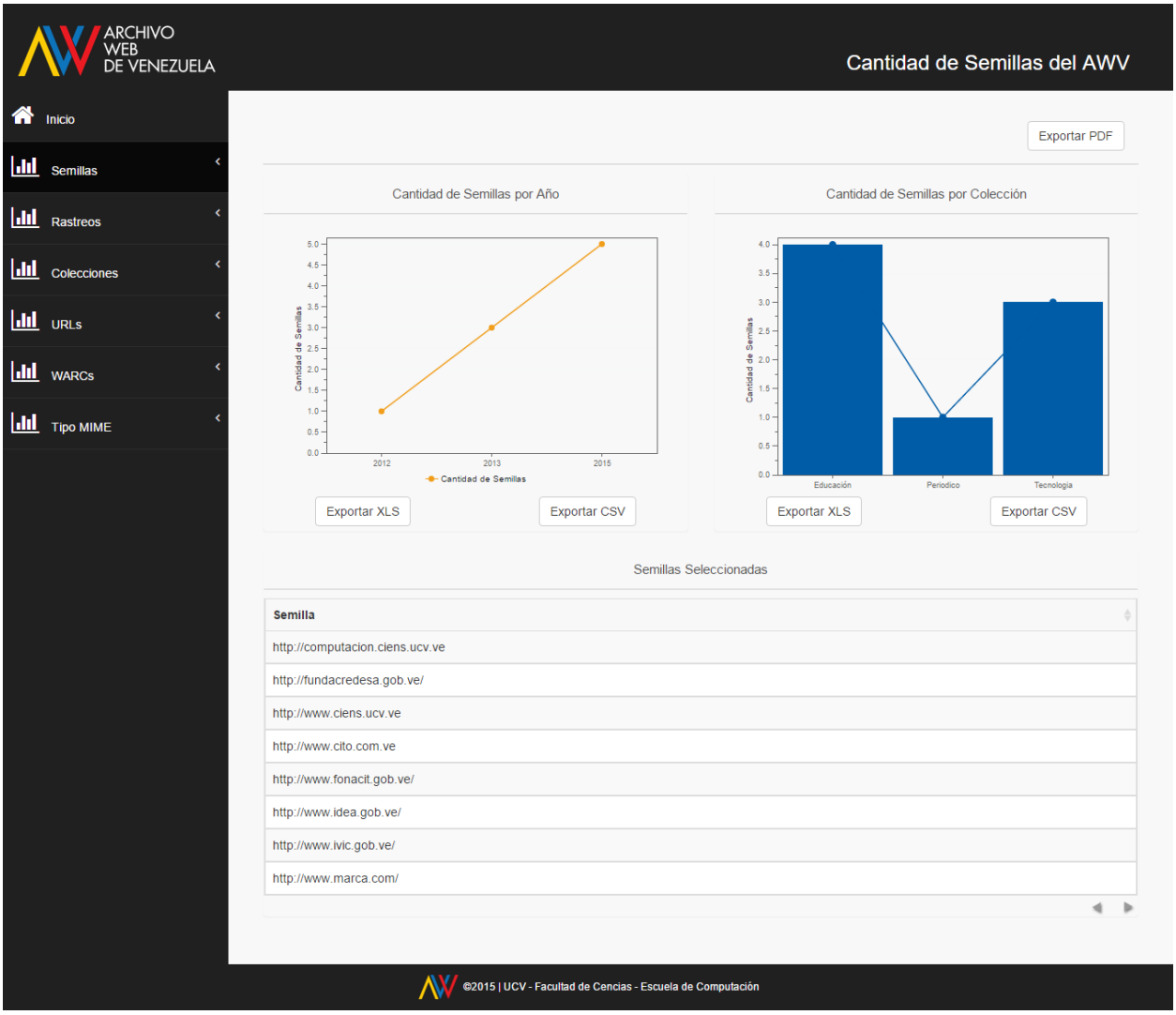


Figura 46 - Vista de Indicador Cantidad de Semillas – General
Fuente: Elaboración propia

2.2. Por año

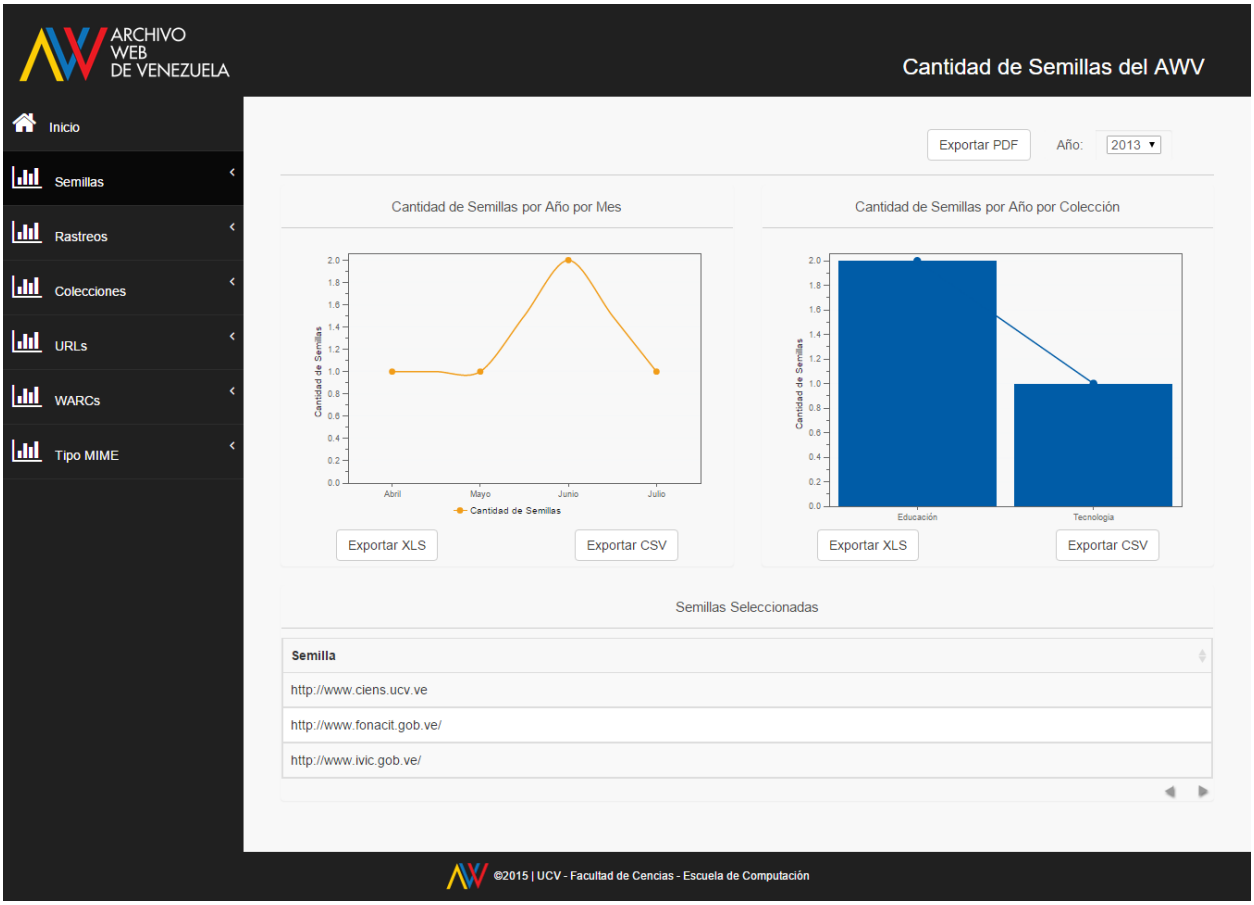


Figura 47 - Vista de Indicador Cantidad de Semillas - Año 2013
Fuente: Elaboración propia

3. Cantidad de rastreos

Formula	Unidad	Criterio de clasificación	Representación
Conteo de rastreos	#	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)

3.1. General

Figura 48 - Vista de Indicador Cantidad de Rastreos – General
Fuente: Elaboración propia

3.2. Por año

Figura 49 - Vista de Indicador Cantidad de Rastreos - Año 2015
Fuente: Elaboración propia

4. Cantidad de colecciones

Formula	Unidad	Criterio de clasificación	Representación
Conteo de colecciones	#	Por fecha (mes y año)	Gráfico histórico (barra/línea)

4.1. General

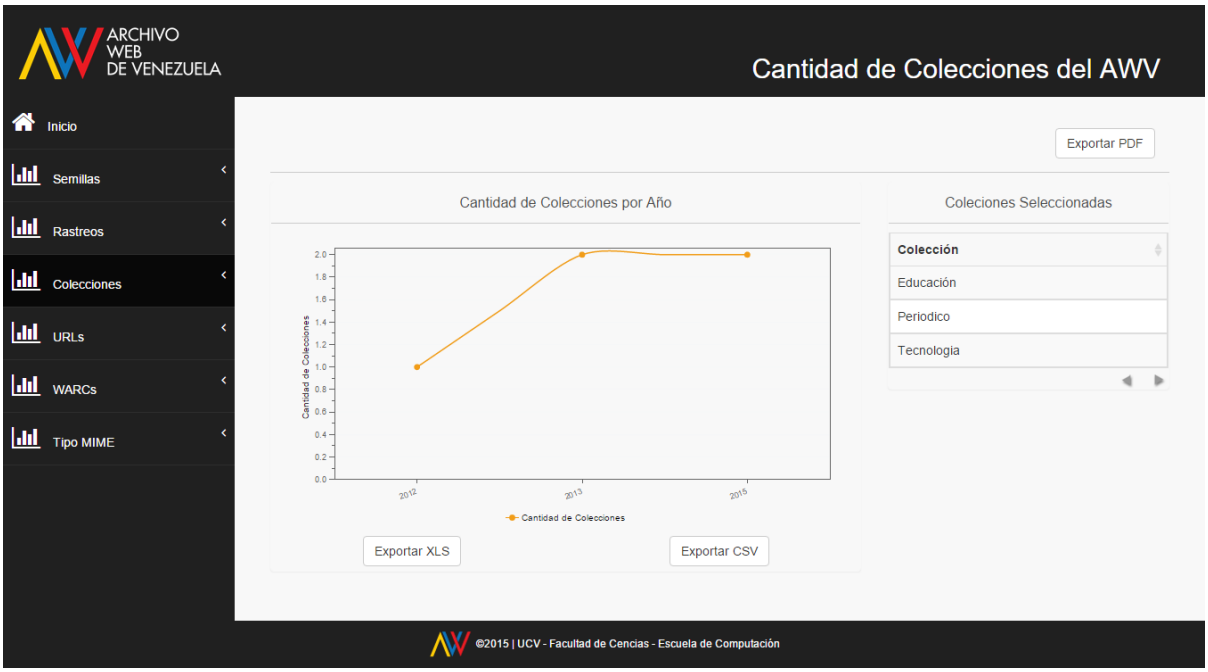


Figura 50 - Vista de Indicador Cantidad de Colecciones - General
Fuente: Elaboración propia

4.2.Por año

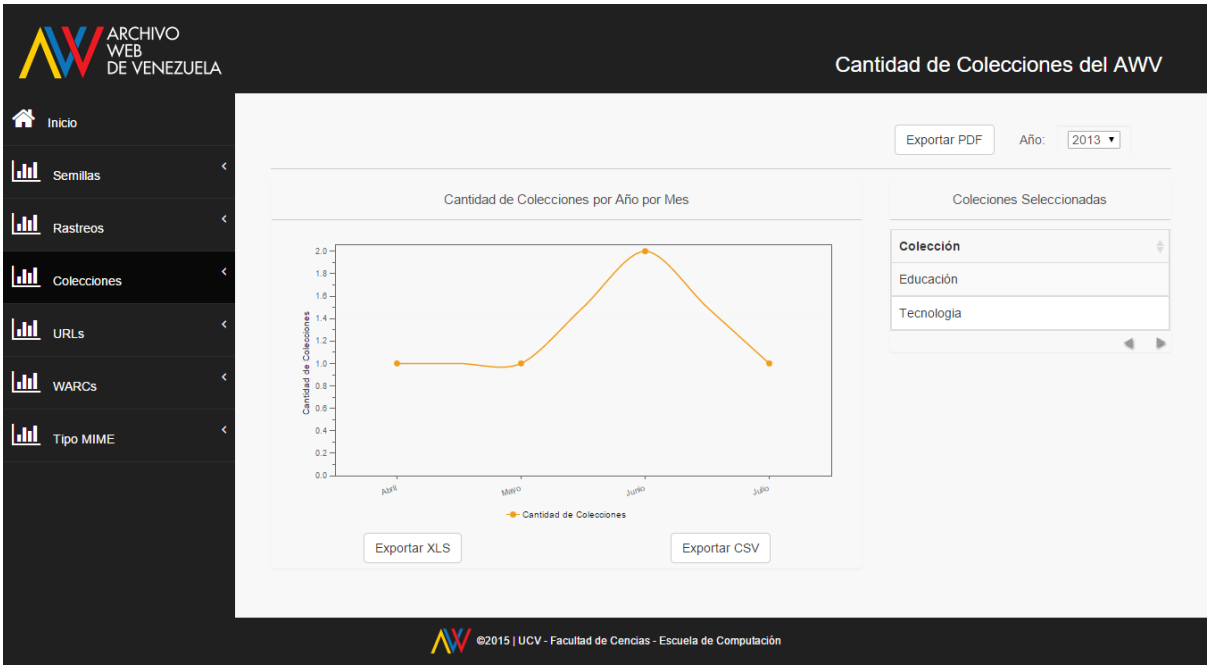


Figura 51 - Vista de Indicador Cantidad de Colecciones – Año 2013
Fuente: Elaboración propia

5. Cantidad de URLs

Formula	Unidad	Criterio de clasificación	Representación
Conteo de URL	#	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)

5.1. General

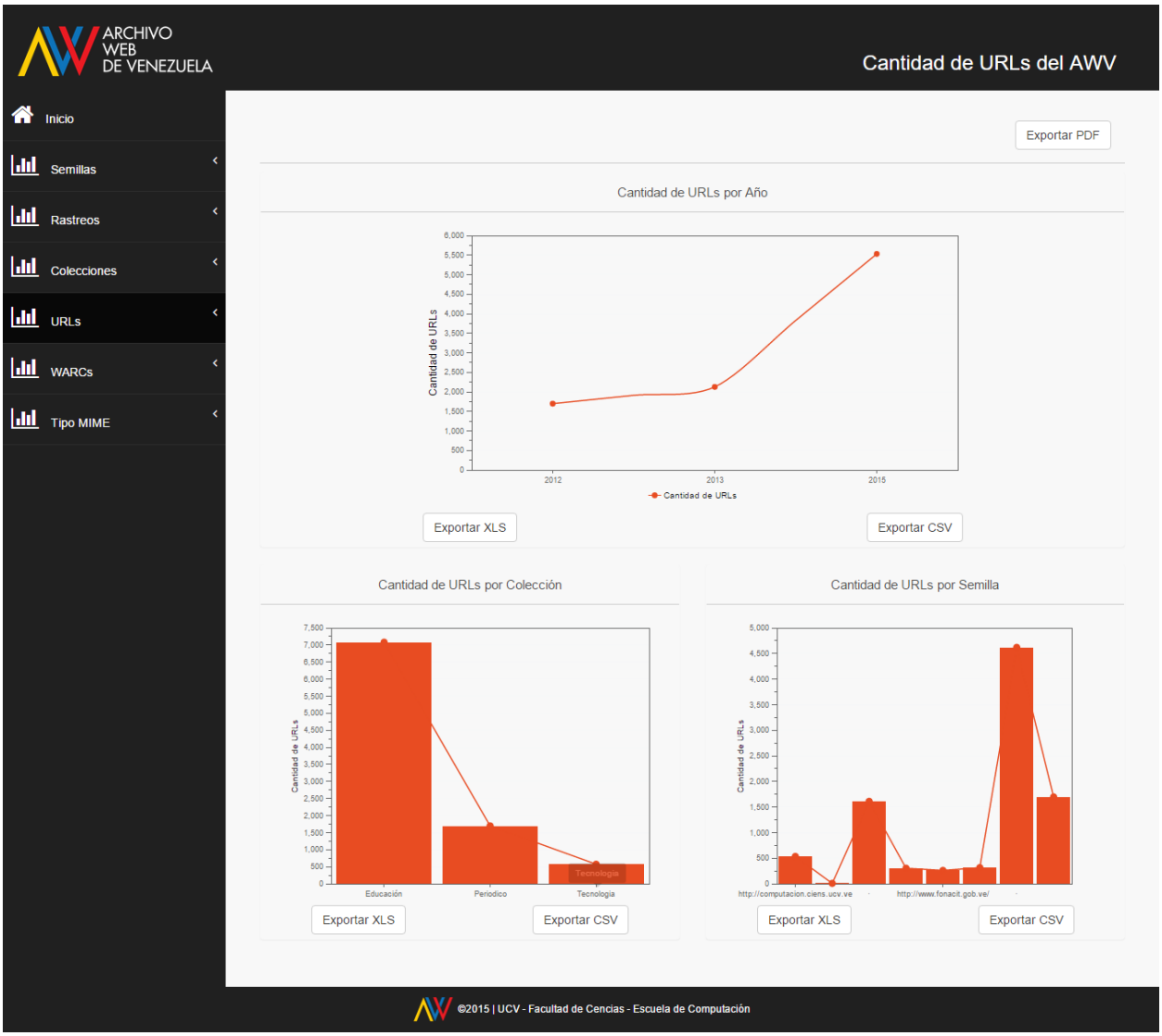


Figura 52 - Vista de Indicador Cantidad de URLs – General
Fuente: Elaboración propia

5.2. Por año

Figura 53 - Vista de Indicador Cantidad de URLs - Año 2013
Fuente: Elaboración propia

6. Distribución de URLs por código de estatus

Formula	Unidad	Criterio de clasificación	Representación
Conteo URL por código estatus/ Conteo URL	%	Por semilla, por colección, por fecha (año)	Gráficos de torta, gráfico de barras

6.1. General

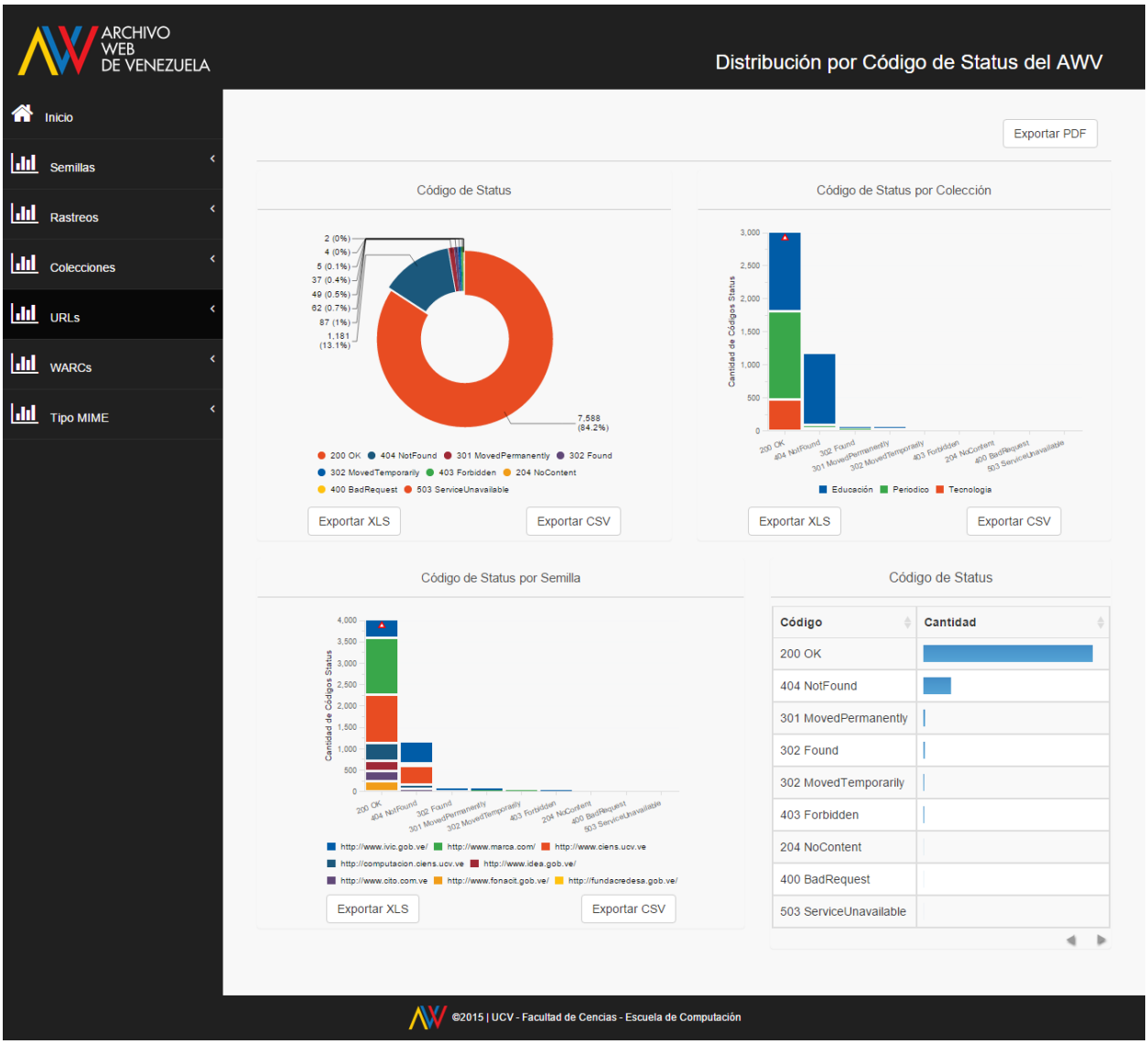


Figura 54 - Vista de Indicador Distribución de URLs por Código Estatus – General
Fuente: Elaboración propia

6.2. Por año

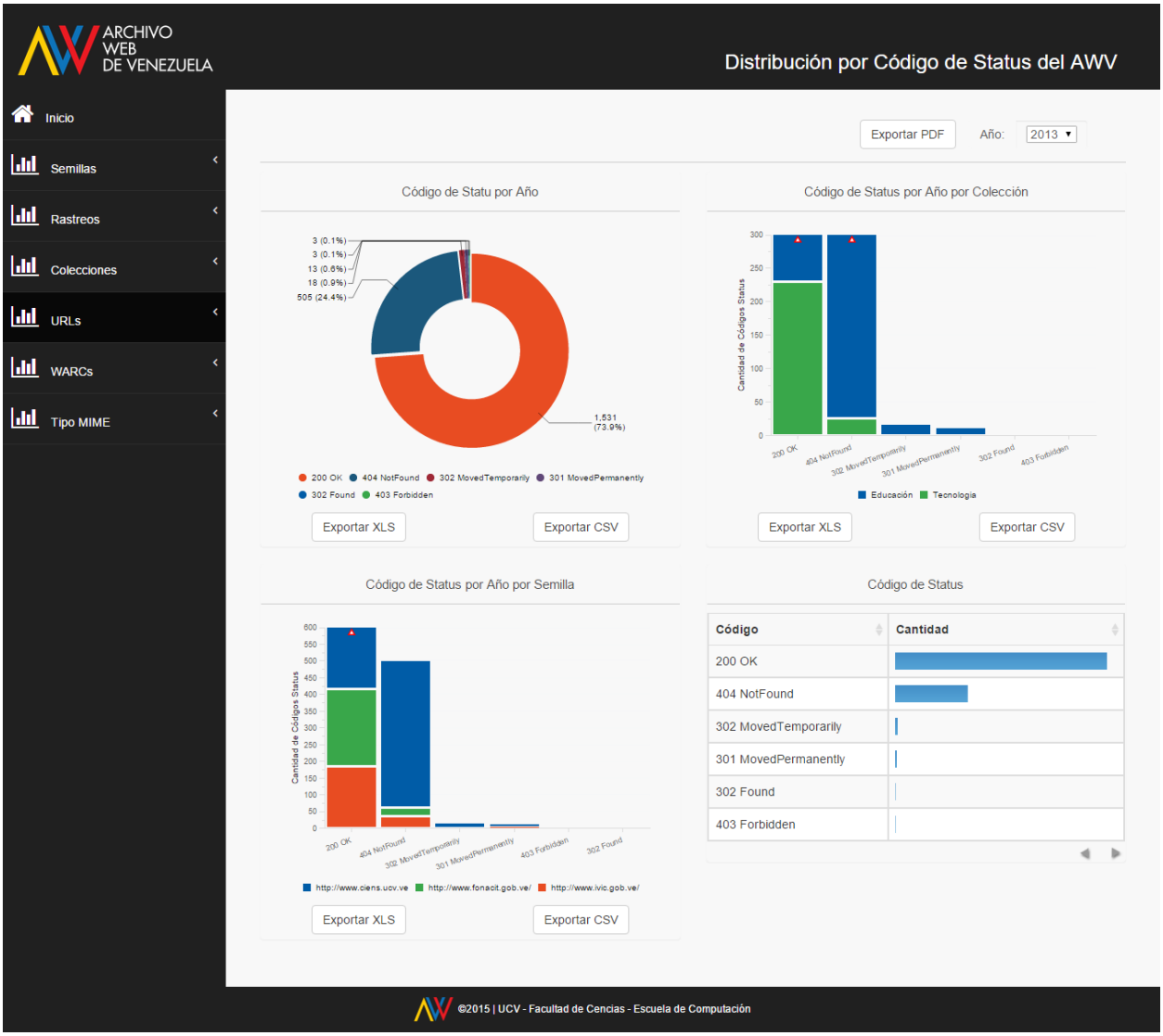


Figura 55 - Vista de Indicador Distribución de URLs por Código Estatus - Año 2013
Fuente: Elaboración propia

7. Cantidad de WARCs

Formula	Unidad	Criterio de clasificación	Representación
Conteo de WARCs	#	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)

7.1. General

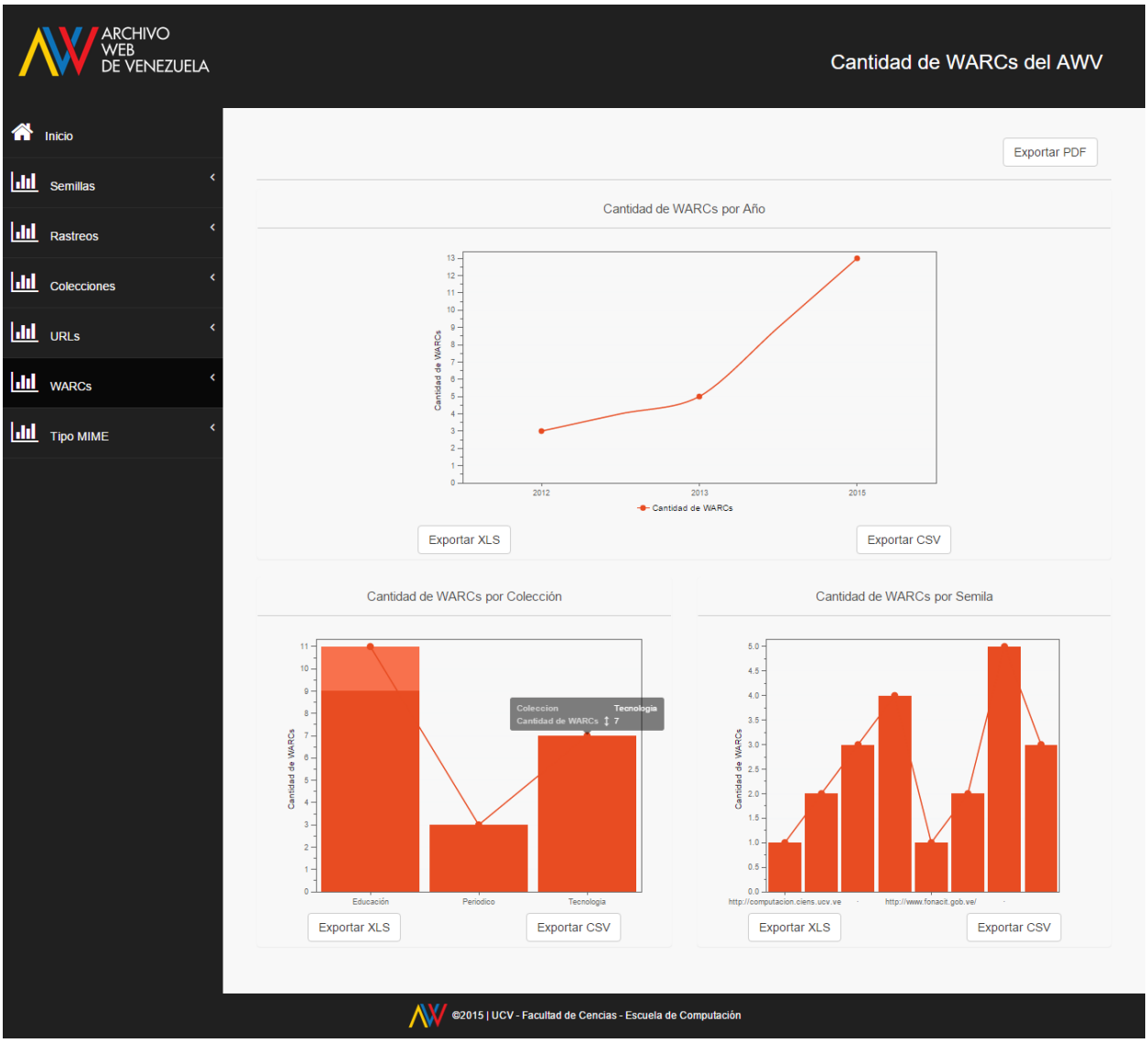


Figura 56 - Vista de Indicador Cantidad de WARCs – General
Fuente: Elaboración propia

7.2. Por año

Figura 57 - Vista de Indicador Cantidad de WARCs - Año 2015
Fuente: Elaboración propia

8. Duración promedio rastreo

Formula	Unidad	Criterio de clasificación	Representación
∑duración rastreos/cantidad rastreos	Minutos	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)

8.1. General

Figura 58 - Vista de Indicador Duración promedio rastreo – General
Fuente: Elaboración propia

8.2. Por año

Figura 59 - Vista de Indicador Duración promedio rastreo - Año 2013
Fuente: Elaboración propia

9. Tamaño del Archivo Web

Formula	Unidad	Criterio de clasificación	Representación
∑bytes rastreados	Bytes	Por semilla, por colección, por fecha (mes y año)	Gráfico histórico (barra/línea)

9.1. General

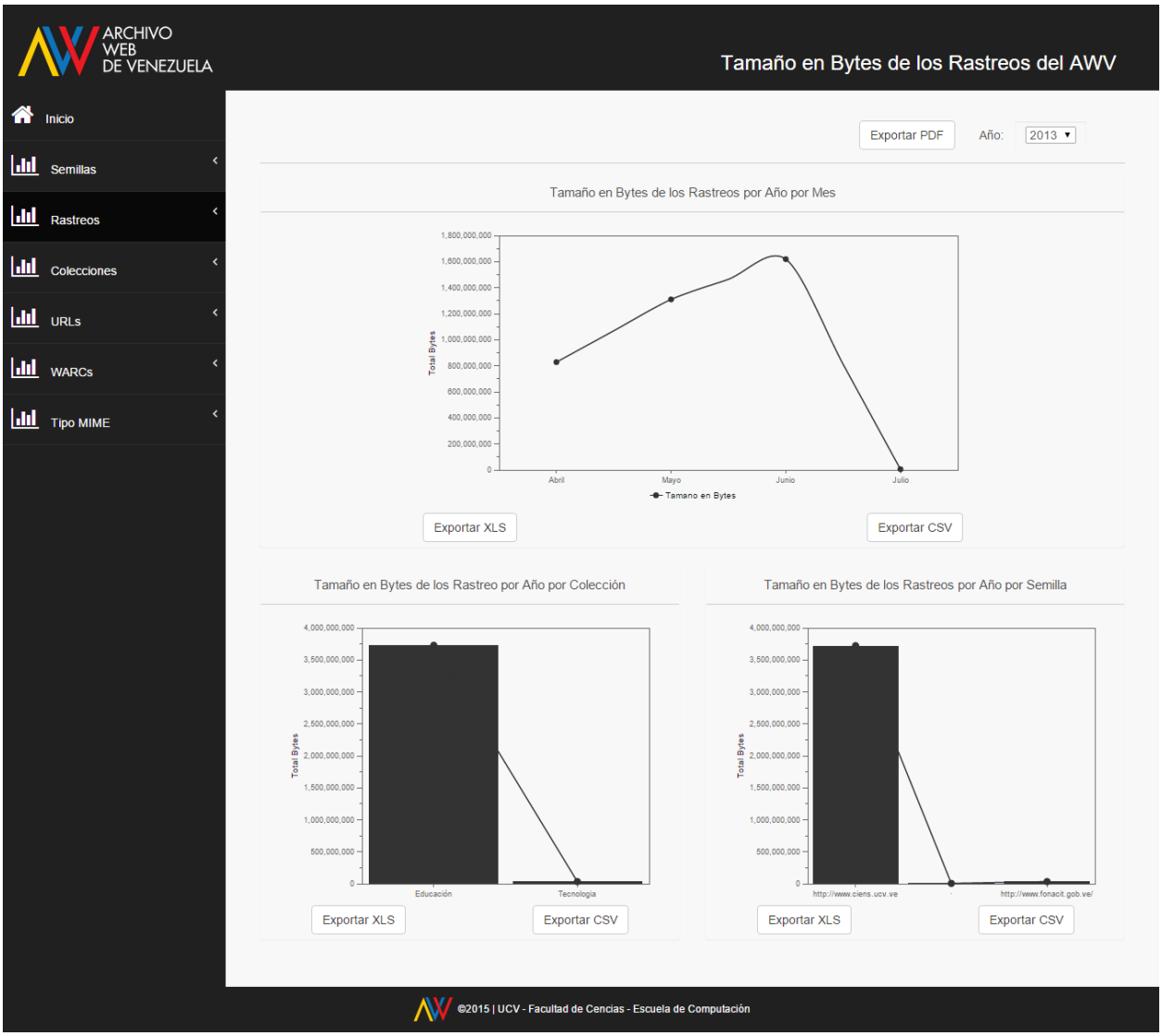


Figura 60 - Vista de Indicador Tamaño del Archivo Web – General
Fuente: Elaboración propia

9.2. Por año

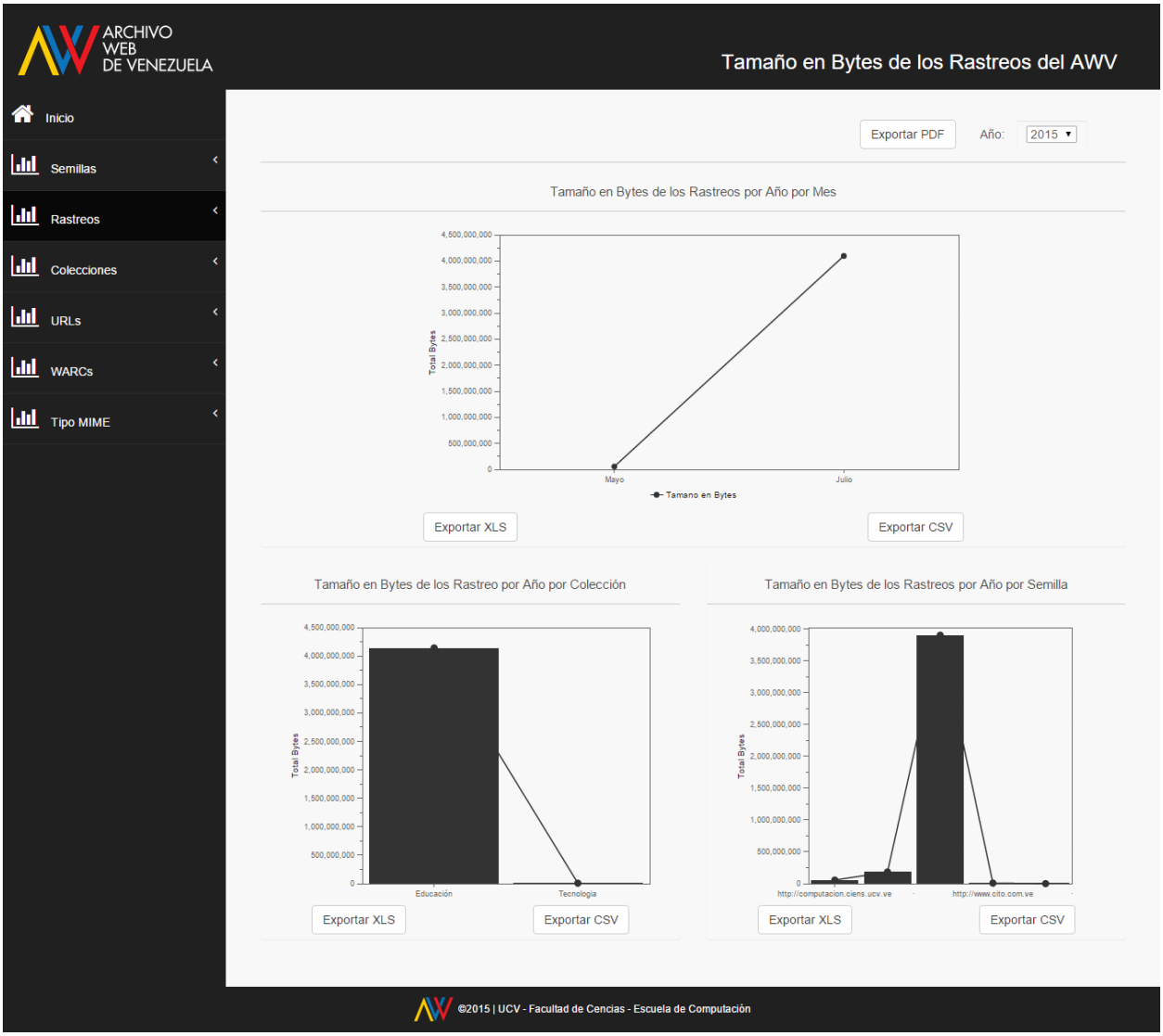


Figura 61 - Vista de Indicador Tamaño del Archivo Web - Año 2015
Fuente: Elaboración propia

10. Distribución de URLs

Formula	Unidad	Criterio de clasificación	Representación
Conteo de URL del archivo web	%	Por semilla, por colección, por fecha (mes y año)	Gráficos de torta

10.1. General

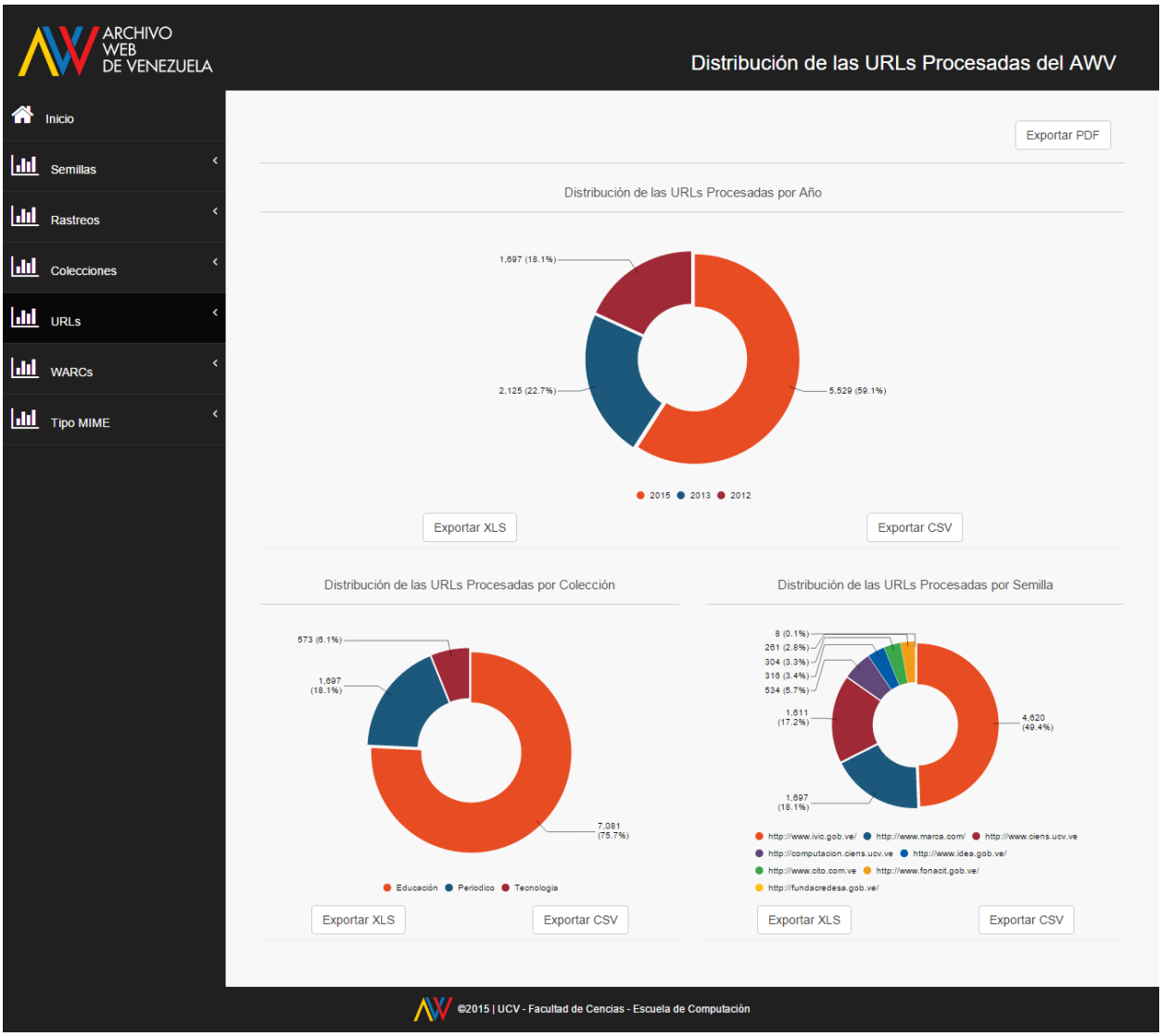


Figura 62 - Vista de Indicador Distribución de URLs – General
Fuente: Elaboración propia

10.2. Por año

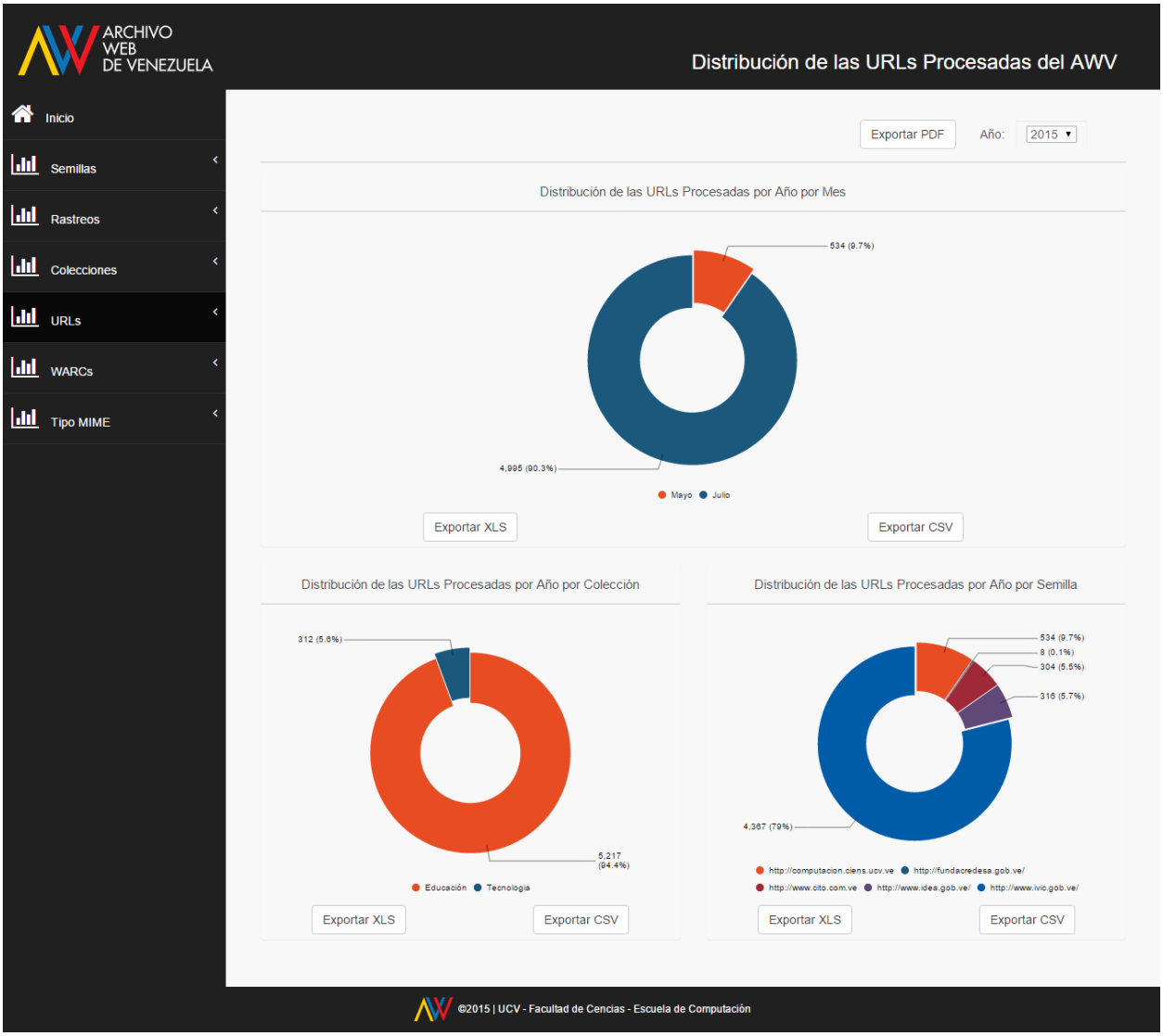


Figura 63 - Vista de Indicador Distribución de URLs - Año 2015
Fuente: Elaboración propia

11. Distribución por Tipos de formatos

Formula	Unidad	Criterio de clasificación	Representación
Cantidad de bytes de formato/total URL formatos	%	Por semilla, por colección, por fecha (año)	Gráficos de torta, gráfico de barras

11.1. General

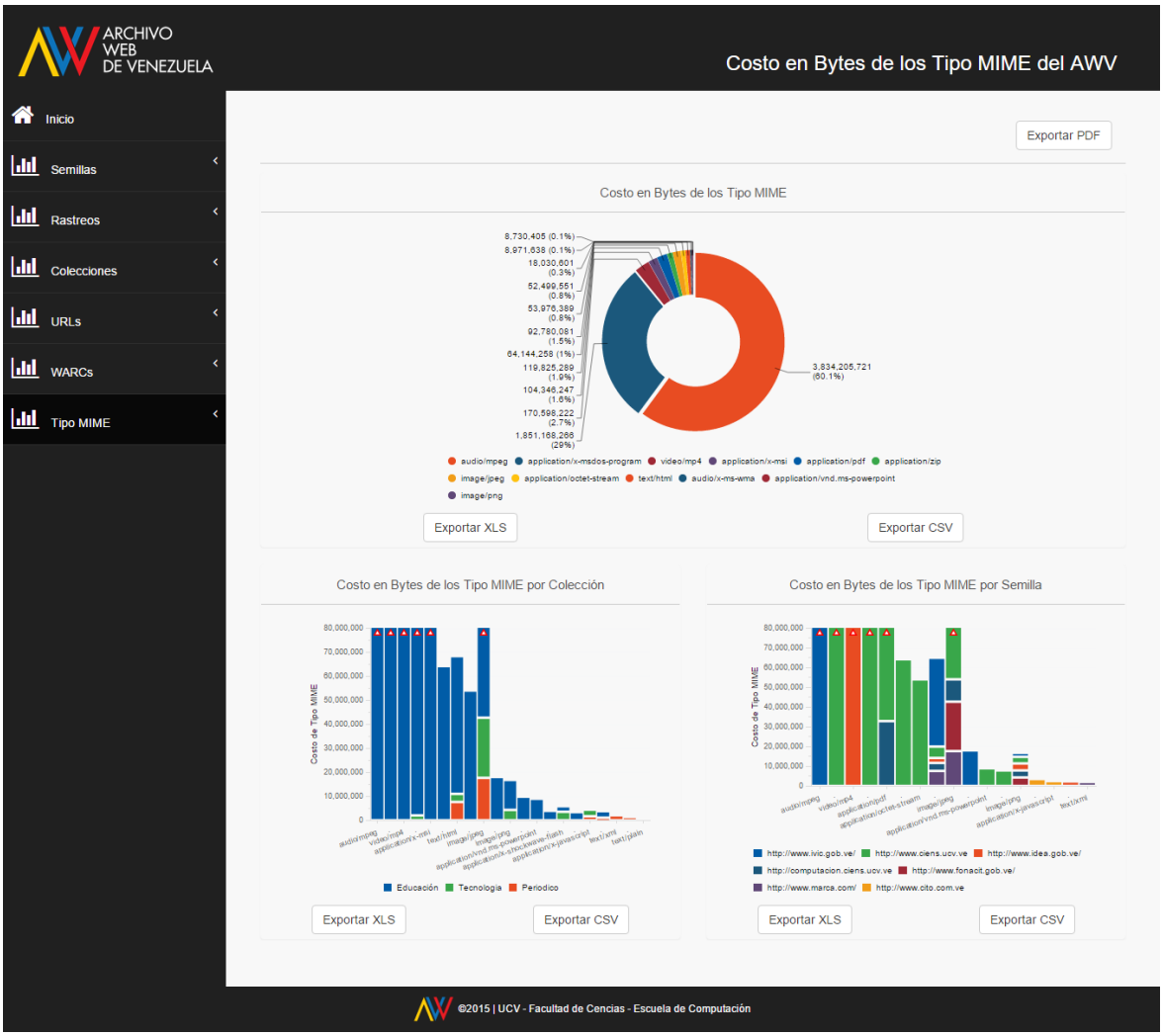


Figura 64 - Vista de Indicador Distribución por Tipos de formatos – General
Fuente: Elaboración propia

11.2. Por año

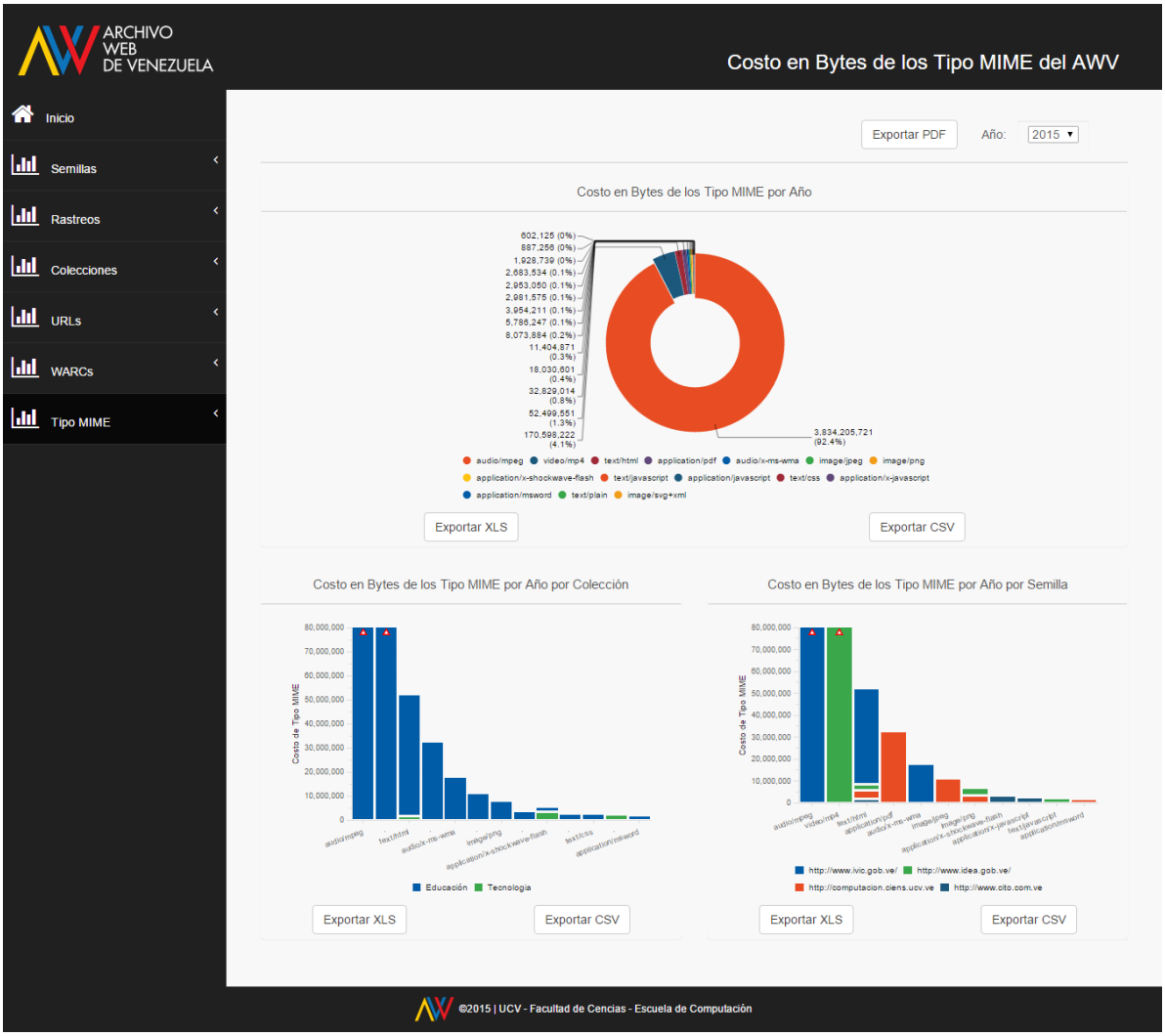


Figura 65 - Vista de Indicador Distribución por Tipos de formatos - Año 2015
Fuente: Elaboración propia

12. Cantidad de URLs por formato

Formula	Unidad	Criterio de clasificación	Representación
Conteo de recursos por formato	%	Por semilla, por colección, por fecha (año)	Gráficos de torta, gráfico de barras

12.1. General

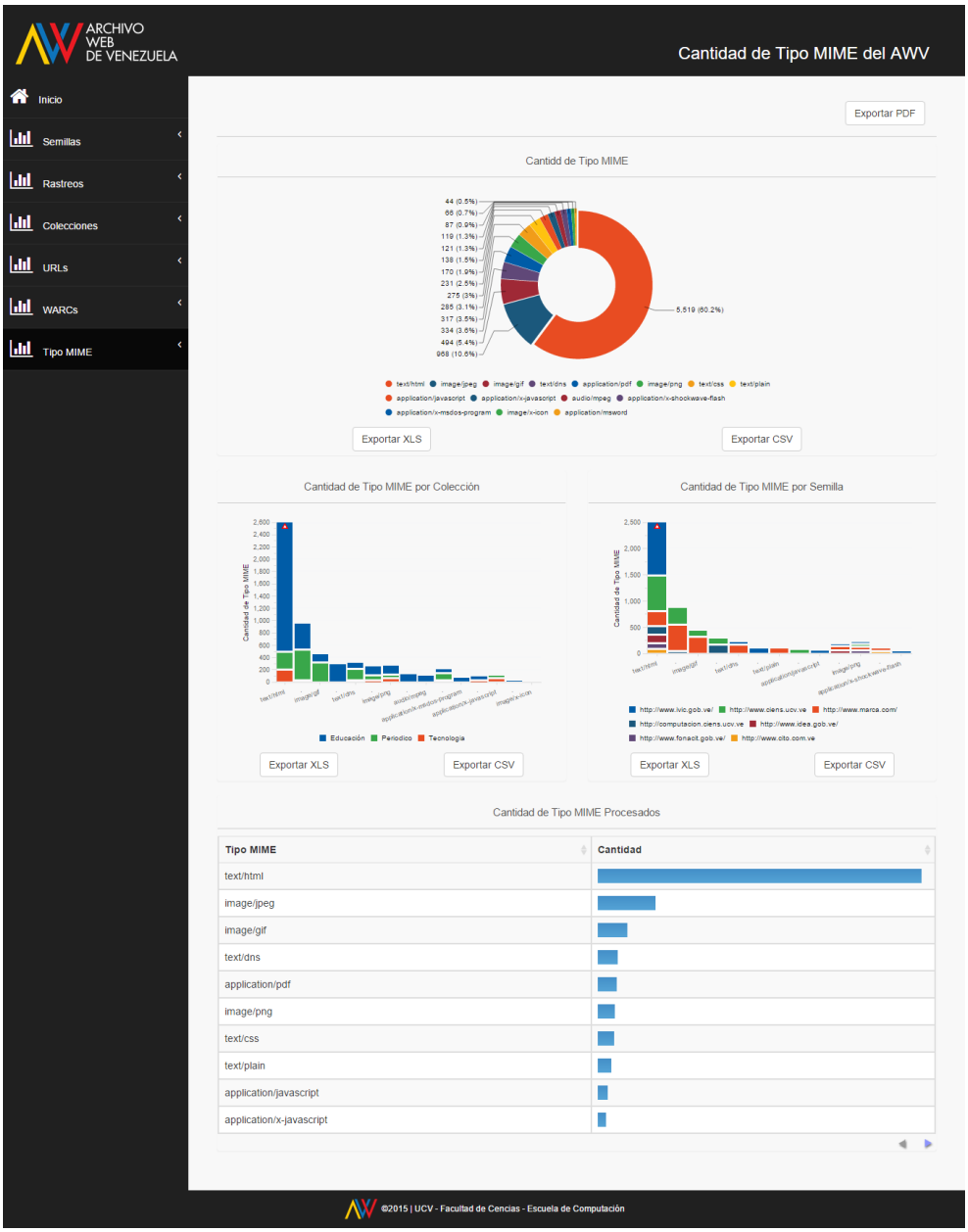


Figura 66 - Vista de Indicador Cantidad de URLs por Tipos de formatos – General
Fuente: Elaboración propia

12.2. Por año

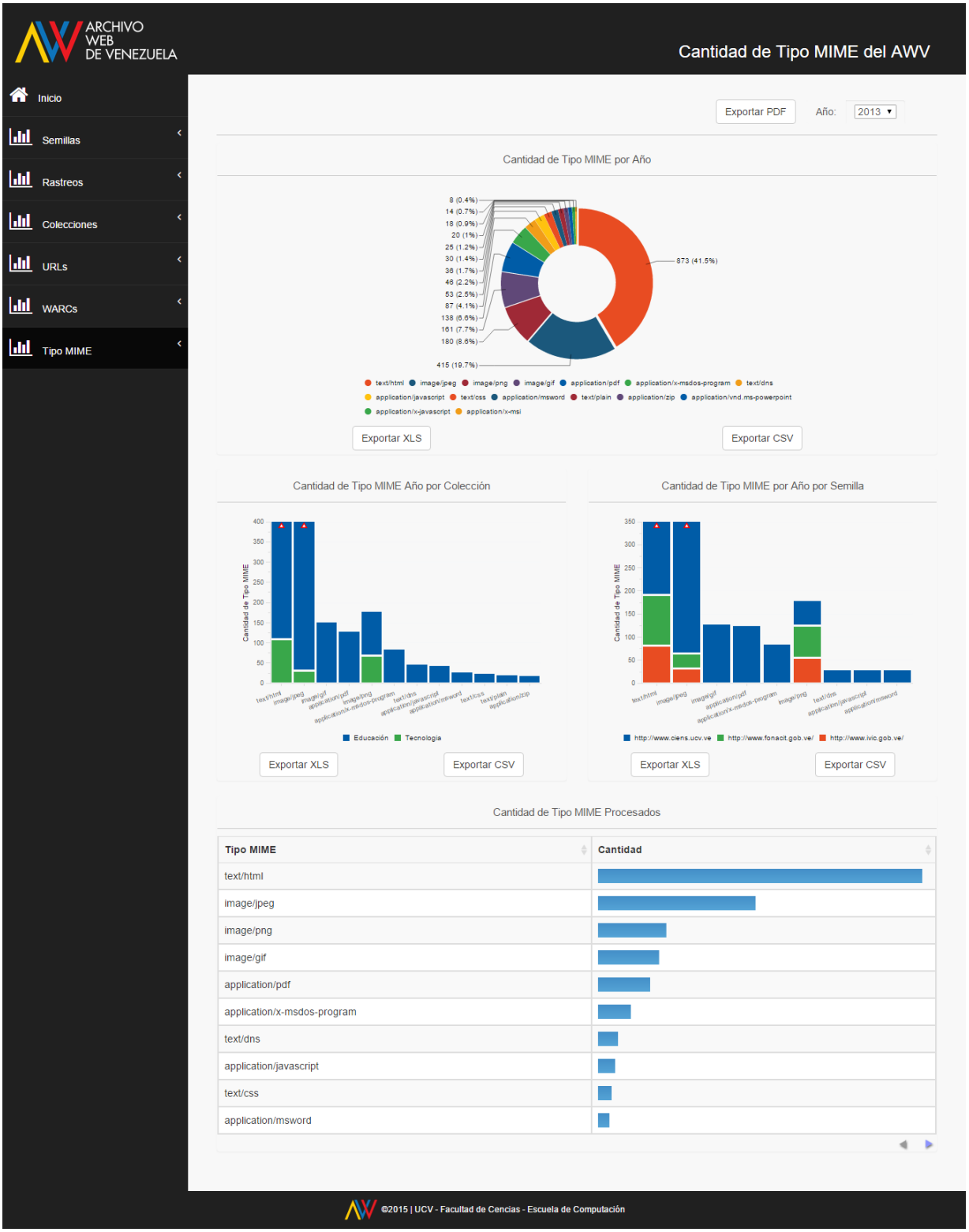


Figura 67 - Vista de Indicador Cantidad de URLs por Tipos de formatos - Año 2013
Fuente: Elaboración propia

13. Cobertura cronológica

Formula	Unidad	Criterio de clasificación	Representación
Fecha inicio rastreo	Fecha	Por semilla, por colección, por fecha	Gráfico de barra

13.1. General

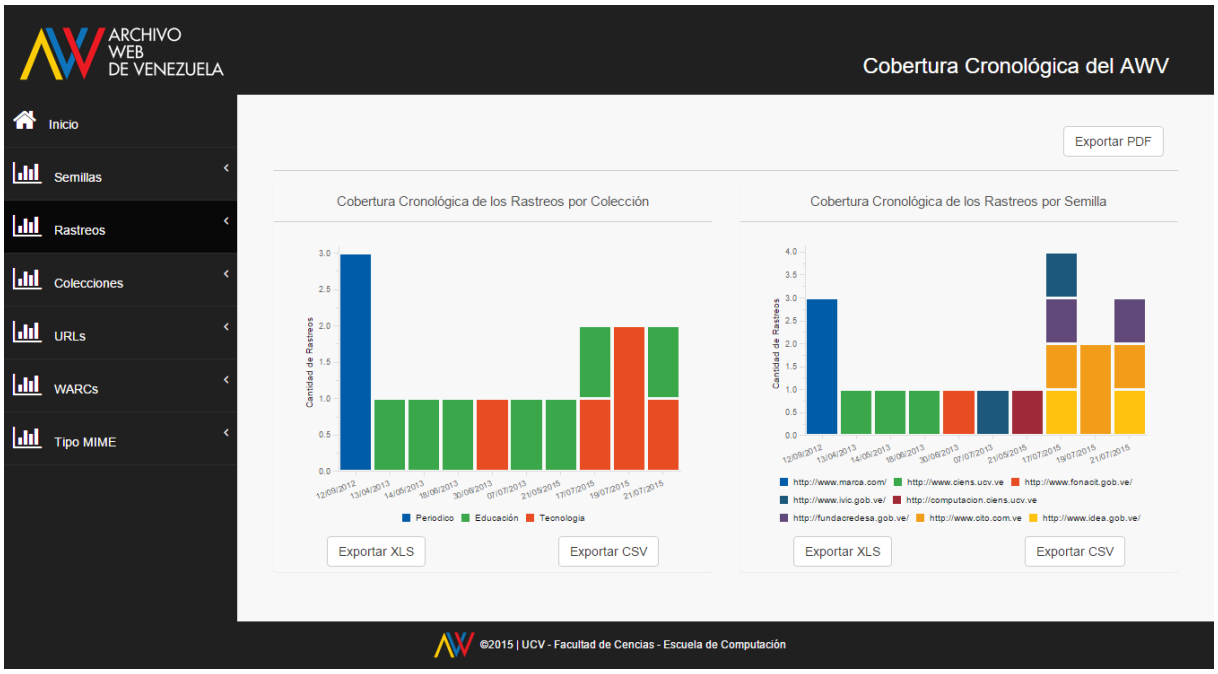


Figura 68 - Vista de Indicador Cobertura cronológica – General
Fuente: Elaboración propia

13.2. Por año

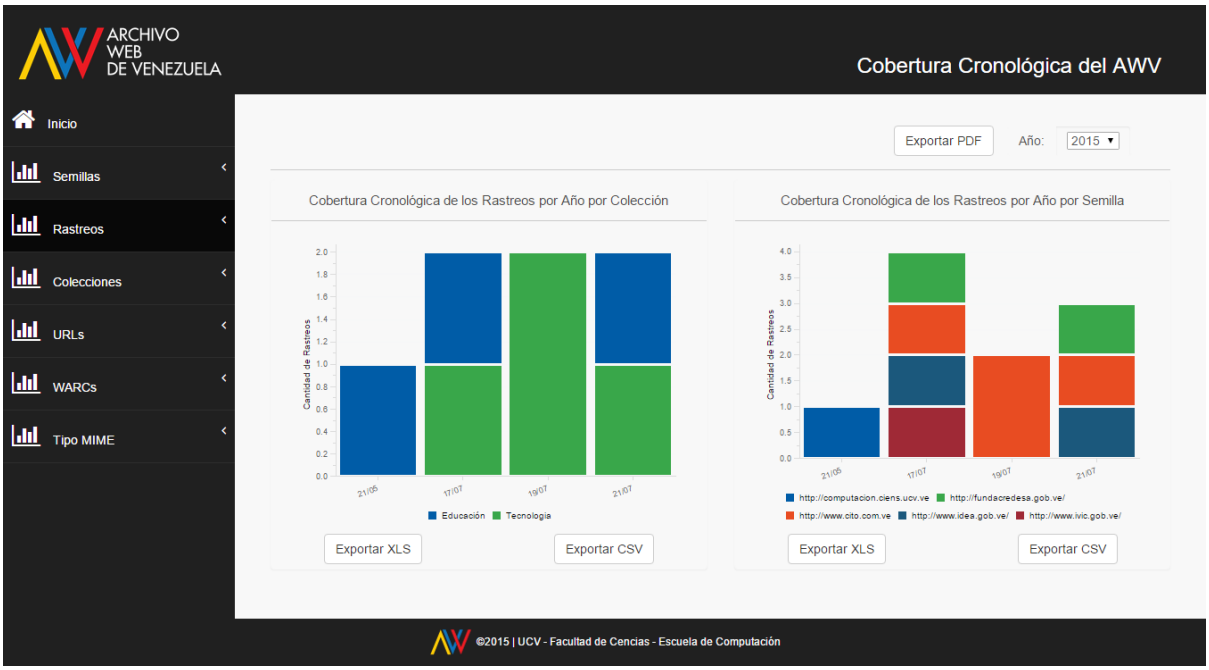


Figura 69 - Vista de Indicador Cobertura cronológica - Año 2015
Fuente: Elaboración propia

14. Costo de objetivo recolectado

Formula	Unidad	Criterio de clasificación	Representación
∑bytes rastreados/conteo URL	Bytes	Top 10 por semilla, por colección, por fecha (año)	Gráficos de torta, gráfico de barras

14.1. General

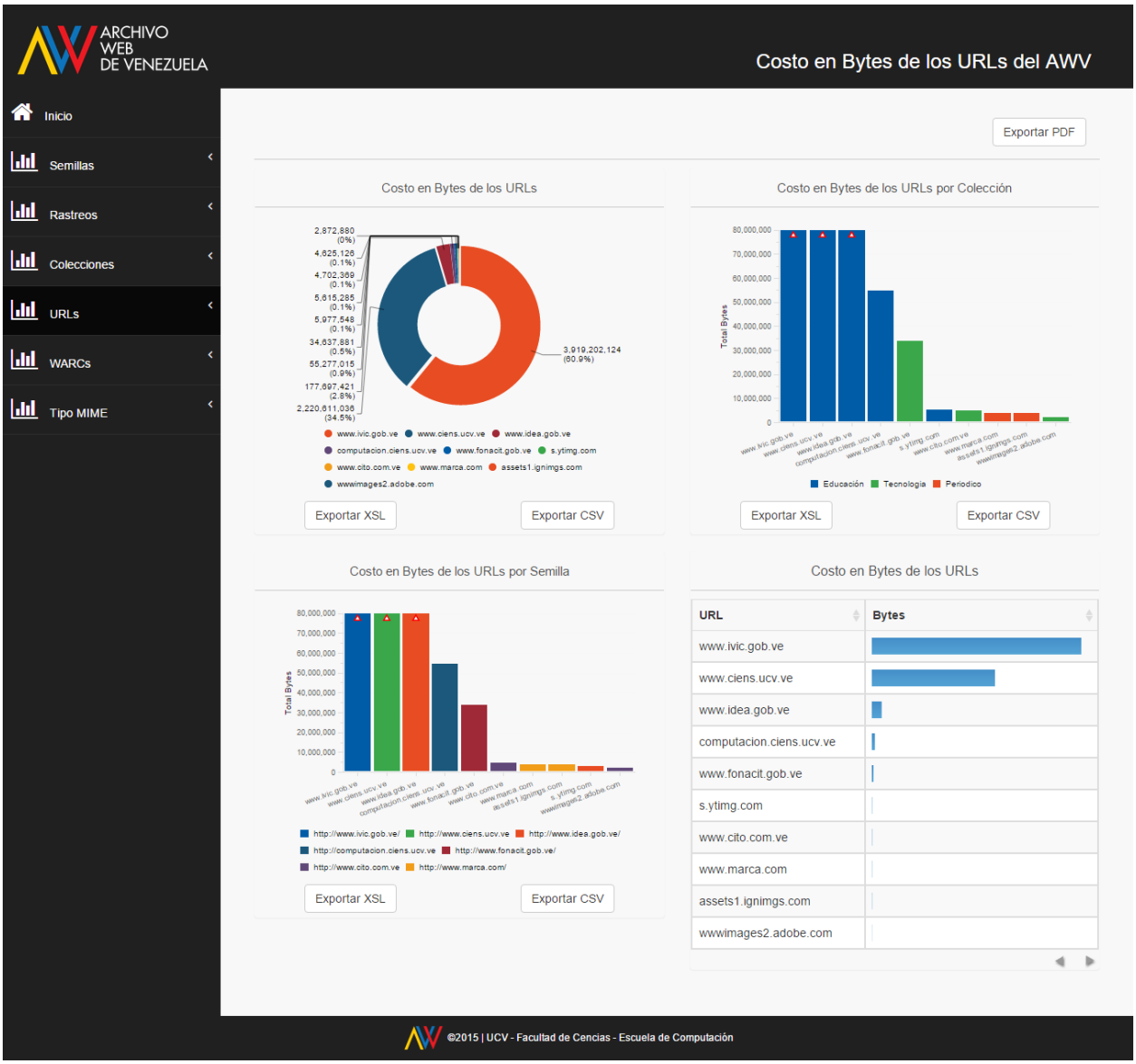


Figura 70 - Vista de Indicador Costo de Objetivo recolectado – General
Fuente: Elaboración propia

14.2. Por año

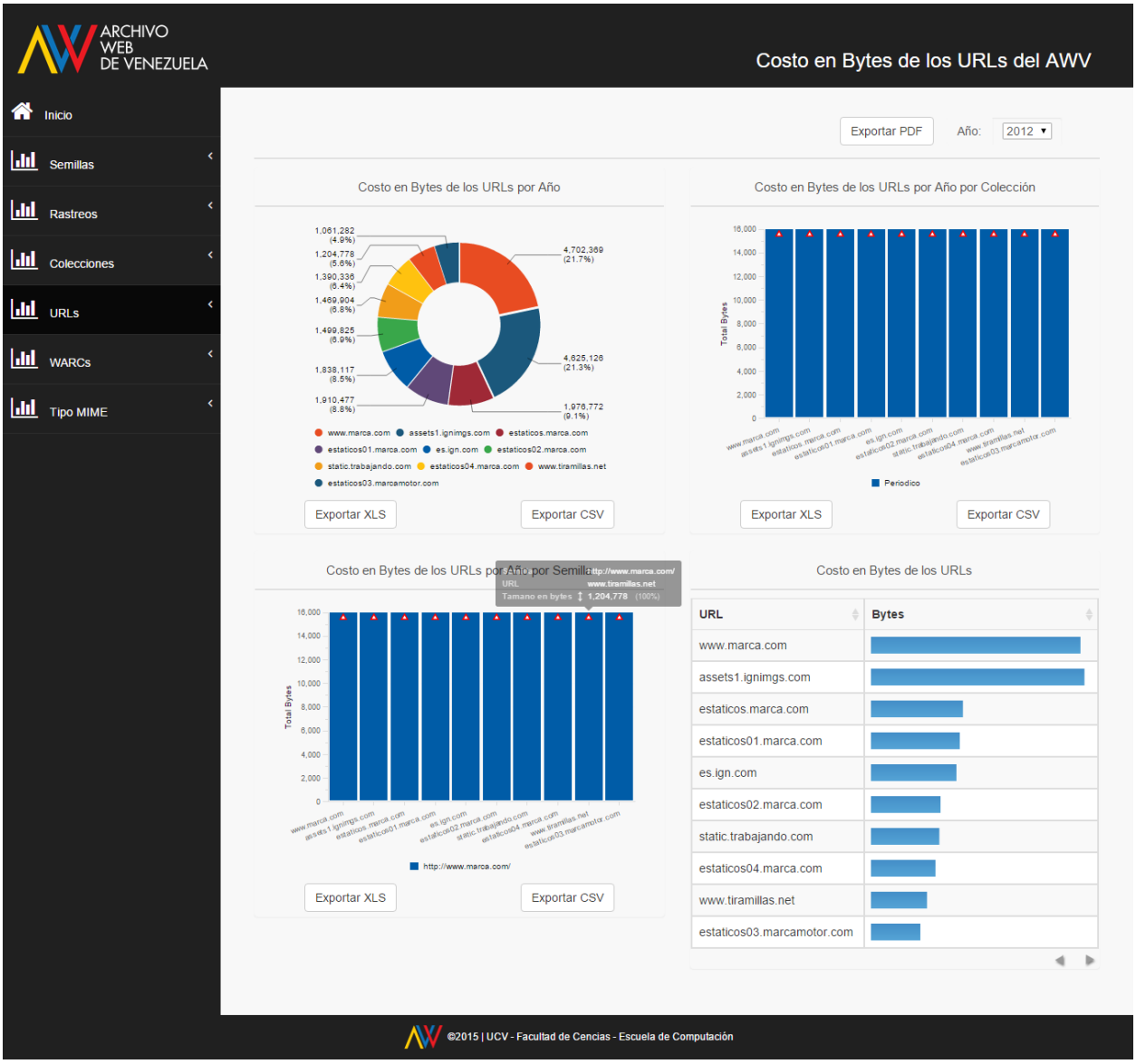


Figura 71 - Vista de Indicador Costo de Objetivo recolectado – General
Fuente: Elaboración propia