Redes Escolares y Portales Educativos de América Latina (REPEAL)

Línea 1: Gestión y Monitoreo de Portales Educativos

Informe Técnico Nº4: "Propuesta Metodológica para un Componente Automático del Observatorio RELPE"

Junio 2007

Institución: Fundación Evolución
Investigadora: Maria Teresa Noguera, noguera@fundacionevolucion.org.ar
Asistente de Investigación: Daniel Finquelievich, danielf@fundacionevolucion.org.ar

TABLA DE CONTENIDOS

ANEXOS

RESUMEN

La Red Latinoamericana de Portales Educativos (RELPE) reúne a los portales educativos nacionales de dieciocho países de la región. La evolución de esta red y en particular sus actividades de intercambio de contenidos digitales educativos ha planteado la necesidad de producir una base de información pertinente, sistemática y confiable que permita constituir un Observatorio capaz de proveer conocimientos para la gestión y el monitoreo de los avances de la red. El objetivo de este trabajo es presentar una propuesta metodológica para la producción de dicha base de información basada en datos pasibles de ser obtenidos y procesados por procedimientos automáticos.

La base de información que comprende este componente automático del observatorio se centra en las siguientes dimensiones: características intrínsecas de los portales tales como la tecnología, el funcionamiento, el contenido y la estructura de los mismos, el uso de los portales y las relaciones que estos mantienen en un cierto contexto de publicación virtual. Para la definición y operacionalización de las métricas relacionadas a estas dimensiones se revisaron los desarrollos en el ámbito de Data Mining y Web Mining.

I. INTRODUCCIÓN

La Red Latinoamericana de Portales Educativos (RELPE), que reúne a dieciocho países, ha sido creada con el fin de constituir una comunidad de intercambio y colaboración para la integración efectiva de las TIC en los procesos educativos. Uno de sus ejes de trabajo ha sido el de constituir un sistema distribuido de almacenamiento y libre circulación de contenidos educativos digitales aportados por sus miembros. Cada portal constituye un nodo que forma parte de una red, la que permite difundir recursos valiosos en el ámbito educativo de la región.

La creación de RELPE ha planteado la necesidad de constituir un Observatorio de Portales que permita estudiar los avances realizados por estos e identificar prácticas exitosas. Este observatorio tendrá como propósito contribuir a generar conocimiento que apoye la gestión y el desarrollo de los propios portales como así también permitir el monitoreo de los resultados del proyecto regional a lo largo del tiempo.

La propuesta para la constitución del Observatorio de Portales RELPE abarca un componente automatizado y uno no-automatizado. En este trabajo se aborda el primer componente que tiene por misión constituir una base de información pertinente, sistemática, confiable y comparable referente a sus miembros. Esta base se constituirá en un insumo para caracterizar la situación de la red y plantear estudios en profundidad que indague acerca de las prácticas que hacen posible los resultados obtenidos.

A continuación se presentan una serie de definiciones referidas a las dimensiones que deben ser medidas así como criterios que definen los procedimientos de recolección y procesamiento de datos que minimicen los posibles sesgos y que deberían aplicarse de manera común a todos los casos.

II. ESTUDIOS DE LA WEB - REVISION DE LITERATURA

La revisión del estado del arte en cuanto a los estudios de la Web presenta un panorama complejo en cuanto al alcance y los propósitos de las métricas utilizadas para caracterizarla (Dhyani, Keong NG et al., 2002). Por un lado es posible identificar un grupo de estudios orientados a caracterizar la Web en general o bien una gran porción de ella (Kleinberg, 1998; S Chakrabarti, Dom et al., 1999; Baeza-Yates & Castillo, 2005; Bordignon & Tolosa, 2006; Tolosa, Bordignon et al., 2006). Estos se realizan en la mayoría de los casos para contribuir al desarrollo de motores de búsqueda. Existe un campo mas reciente que es el estudio de la Web en tanto red social (Scott, 2000; Lauw, Lim et al., 2005).

Es posible reconocer otro grupo de investigaciones enfocadas a caracterizar las unidades que componen la Web (ej: dominios, sitios o páginas) desde distintos aspectos, ya sea con el propósito de contribuir a la mejora sitios Web o bien de permitir la personalización de los mismos (Srivastava, Cooley et al., 2000; Dhyani, Keong NG et al., 2002; Poblete & BaezaYates, 2006). Estos estudios se basan en los métodos de Data Mining y Web Mining que comprenden tres tipos principales considerando la naturaleza de los datos analizados: la minería de uso, la minería de contenido y la minería de estructura (Srivastava, Cooley et al., 2000).La minería de uso (Web Usage Mining) es la aplicación de técnicas de minería de datos al análisis de las secuencias de vistas de páginas (click-stream) a fin determinar patrones de uso (Cooley, 2000). La minería de contenido (Web Content Mining) busca caracterizar a las páginas de un sitio principalmente en función del texto que contienen (Da Graca Guerra & Modarelli, 2003). La minería de estructura busca caracterizar la organización de dicho contenido (Web Structure Mining) (Botafogo, Rivlin et al., 1992). Los estudios de la estructura interna de los sitios han demostrado el impacto de esta característica sobre la posibilidad del usuario de efectivamente encontrar contenido dentro de un sitio, y sobre la *usabilidad* del mismo (Díaz, 2003; Miller & Remington, 2004; Adhikari & Lemone, 2007). La estructura del sitio influye a su vez en la posibilidad del mismo de ser *ubicable* y *visible* para los usuarios y para los motores de búsqueda responsables indexar su contenido (Baeza Yates, 2004).Si bien no es muy citado en la bibliografía, existe un grupo de estudios conocido como de "minería de consultas" (Query Mining) que permite a los sitios, que cuentan con motores de búsqueda, encontrar información útil a partir de las consultas de los usuarios que quedan registradas y que reflejan sus intereses (Poblete, 2004).

Estas aproximaciones han dado lugar a una serie de herramientas de uso comercial así como investigación comercial y académica. La mayoría de las herramientas en el ámbito comercial se basan en el parseo (parsing), limpieza y análisis de los archivos LOG para producir principalmente estadísticas de uso. En los trabajos de investigación abundan distintas especializaciones abocadas al análisis de estructura y de contenido Web combinadas con el análisis de uso. Dentro de este grupo de estudios encontramos distintas técnicas tales como: producción de estadísticas que describen las propiedades de los sitios y ayudan a mejorar los procesos de autoría, producir reglas de asociación,

clustering, clasificaciones, patrones de secuencias y modelado de dependencias utilizados para conocer las preferencias de los usuarios y su comportamiento (McEneaney, 1999; Poblete, 2004; Bordignon & Tolosa, 2006; Poblete & BaezaYates, 2006; Adhikari & Lemone, 2007). Existe un grupo de trabajos que buscan representar el comportamiento de navegación de los usuarios recurriendo a métodos de modelización a fin de evaluar la usabilidad de los sitios (Miller & Remington, 2001, 2004).

Resulta interesante señalar un último grupo de estudios que se diferencia de los anteriores por que no utilizan procedimientos automáticos sino que adaptan las técnicas tradicionales de evaluación de software y se abocan en particular al estudio "aplicaciones Web". El marco WebQEM aporta la visión del estudio de un sitio en tanto "artefacto" analizando la "calidad del producto" y la "calidad del producto en uso"[1]. Si bien no es posible tomar las técnicas de estos estudios por su orientación a la evaluación basada en procedimientos no automáticos, en el presente trabajo tomaremos estas dos ideas que permitirán definir para el contexto de los portales educativo dos dimensiones de estudio: las "características intrínsecas" de los portales y las características de los portales en relación a sus usuarios y su entorno, que llamaremos "características de los portales en uso".

III. MODELO CONCEPTUAL

La literatura existente sobre los portales y los portales educativos en particular permiten definirlos como puntos únicos de acceso a conjuntos de Sitios Web que proveen información y servicios, o funcionalidades puntuales[2], a un cierto grupo de usuarios (Franklin, 2004). Los portales educativos cumplen en mayor o menor medida con dos funciones básicas. Por un lado son Portales Conectores (*Networking Portals*) en la medida que proveen un punto de entrada para acceder a diferentes herramientas y recursos (información, links a otros sitios, newsletters, grupos de discusión, suscripciones, etc)[3]. Así mismo, cumplen la función de ser puntos de entrada a la navegación a fin de que los usuarios puedan enfocar los procesos de búsqueda y acceso a los contenidos valiosos evitando perderse entre el masivo volumen de recursos presentes en Internet. Asimismo pueden ser considerados también Portales Organizacionales (*Organizacional Portals*) en tanto transmiten filosofías, políticas y publicaciones que representan por ejemplo a los ministerios de educación. Un tipo especifico de Portal Conector son los portales Basados-en-Recursos (*Resources Based Portals*) dado que cumplen la función de ofrecer contenidos de diferentes temáticas sean adquiridos y/o creados, herramientas de búsqueda y conexiones a otras fuentes relevantes (Butcher, 2002).

[1] *Ver Anexo 1*
[2] *Para hacer una diferencia con las aplicaciones Web*
[3] *Desde la clasificación desarrollada por Elena García García, E. Portales educativos. Consideraciones de tipo general corresponden a una combinación de portales informativos, de encuentro y banco de recursos.*

En este trabajo se reconoce a los portales educativos RELPE desde su función de *Portales Conectores-Basados en Recursos*, que siendo a la vez portales nacionales tienen la misión de alcanzar a una audiencia compuesta por la totalidad de la comunidad educativa del país. A partir de esta definición podemos resumir que cumplen tres funciones básicas: (i) proveen recursos educativos de diferente naturaleza, (ii) buscan alcanzar a una audiencia objetivo compuesta por docentes, alumnos, directivos de todos los niveles educativos del país (algunos pueden estar más atendidos que otros), y (iii) constituirse en puntos de entrada que permite enfocar las conductas de búsqueda y navegación de los usuarios tanto al interior como al exterior del sitio.

A partir de la revisión de la literatura se trabajó en identificar diversos aspectos que pudieran dar cuenta de las cuestiones planteadas por cada función de los portales. Posteriormente, estos aspectos identificados fueron agrupados en dimensiones según puede observarse en la Tabla 1.

Tabla 1 – Funciones y Dimensiones de los Portales Educativos

Funciones		**Cuestiones asociadas**		**Aspectos**	**Dimensiones**
(i)	Provisión de Recursos	*¿Qué se provee y cómo?*		Contenido Estructura Interna Tecnología Funcionamiento	**(a) Características Intrínsecas**
(ii)	Cobertura	*¿A quiénes alcanza?*	*¿Cual es la proporción de la audiencia alcanzada?*		No automatizable
(iii)	Punto de Entrada	*¿Cuál es la conducta de Navegación Interna?*	*¿Cuál es la actividad general del portal?* *¿Cuánto es usado?*	Trafico	**(b) Portal en Uso**
			¿Cuáles contenidos se usan?	Uso de contenidos	
			¿Cómo es usado el portal?	Comportamiento de los usuarios	
		¿Cuál es la conducta de Navegación hacia el portal?		Comportamiento de Entrada	
		¿Cómo se relaciona el portal con el espacio Web?		Estructura Externa	**(c) Estructura Externa**

Por ejemplo, para la caracterización de los portales según los recursos que proveen se incluyeron los aspectos de: Contenido, Estructura Interna, Tecnología y Funcionamiento. Estos aspectos fueron agrupados en la dimensión denominada *Características Intrínsecas de los Portales*.

Las cuestiones vinculadas a la "Cobertura" y "Alcance de la Audiencia" no fueron consideradas en este trabajo por no poder ser abordadas con fuentes de datos automatizables (En el Anexo 2 se presenta un detalle de las posibilidades para el estudio de audiencia de los portales).

Por último, la función de los portales de constituirse en Puntos de Entrada a la navegación fue dividida en dos dimensiones. La dimensión *Uso del Portal* incluye los aspectos de: Tráfico, Comportamiento de los Usuarios y Uso de los Contenidos. En tanto la dimensión *Estructura Externa* busca dar cuenta de la relación del portal con otros sitios en la Web.

a. Características Intrínsecas

Desde una visión operativa los contenidos digitales son los datos reales, en general multimediales (textos, imágenes, etc.), que almacenan los sitios Web y que son provistos a los usuarios (Poblete, 2004). Debe notarse que al hablar de recursos en este trabajo se los está circunscribiendo a lo que se conoce como "contenidos educativos", excluyéndose a los servicios como e-mail, foros, etc. El termino "contenido educativo" se refiere en consecuencia a la "información digitalizada" que está dirigida a miembros de una comunidad educativa para ser utilizados según propósitos de enseñanza, aprendizaje y/o evaluación (Libedinsky, 2007).

La estructura interna de los portales puede ser definida como la organización que estos contenidos adoptan al interior de un sitio. Se adopta en particular la perspectiva de organización entre-páginas, la que se establece a partir de las URLs y los links que los conectan (Srivastava, Cooley et al., 2000). Este aspecto es de suma relevancia puesto que define algunos aspectos de la usabilidad de un sitio y que tiene efectos directos sobre la posibilidad de que los contenidos sean encontrados por los usuarios.

Otro aspecto que permite caracterizar a los portales educativos es la tecnología utilizada para su construcción. en tanto sitios o conjunto de sitios Web, pueden ser caracterizados, siguiendo a Powell, Jones et al (1998)[4], por su *Complejidad*; según si predominan las páginas estáticas (simples) o las páginas dinámicas (complejas). Así mismo, siguiendo a Bordignon & Tolosa (2006), puede incluirse la *Orientación* según la proporción relativa de documentos no-HTML (audio, software, documentos, etc). Siguiendo a estos últimos autores se agrega el aspecto del funcionamiento, el cual está relacionado con la integridad de lo que se ofrece y el tiempo de respuesta de acuerdo a las condiciones de acceso de los usuarios. La inclusión de los aspectos de *Tecnología* y *Funcionamiento* permite describirlos en tanto "artefactos Web" y se justifica en que constituyen en todos los casos requisitos mínimos para el éxito en proveer recursos a los beneficiarios así como en la evidencia que lo vincula con la accesibilidad y la usabilidad de los mismos.

b. Portal en Uso

La dimensión del producto portal educativo en uso, de acuerdo a la literatura revisada, permite describir en que medida alcanzan los portales sus propósitos de llegar a los

[4] *Ver detalles en Anexo 3.*

usuarios y es un resultado de las características intrínsecas del producto junto a la relación que este mantiene con el entorno virtual y las condiciones contextuales de los usuarios (acceso).

Estudiar al portal en uso implica considerar las actividades desarrolladas por los usuarios en el entorno Web. En este trabajo no se aborda el uso de los portales desde la perspectiva de las diferentes aplicaciones que los usuarios puedan darles en la práctica docente o de aprendizaje. Esta dimensión real del uso esta vedada por las restricciones que impone contar exclusivamente con fuentes de información automatizables. Operativamente el uso se define como cualquier actividad de un usuario único identificado[5], por ejemplo a partir de un análisis combinado de los registros de actividad del servidor (IP, Path), URL irrepetible y cookie persistente[6]. La caracterización del uso del portal hace que se deba considerar además el comportamiento de navegación hacia el interior del portal. En lo que se refiere a la navegación interna el uso del portal podrá caracterizarse de acuerdo a los perfiles de comportamiento de los usuarios, puntos de acceso, caminos recorridos, contenidos más utilizados, etc.

La inclusión del aspecto de *Tráfico* en esta dimensión busca lograr medidas de: Nivel de Demanda del portal, Nivel de Asiduidad de la Demanda, Nivel de Actividad de los usuarios y Duración de la Actividad. En cuanto al *Uso de los Contenidos* se incluyen medidas de Diversidad y Riqueza de los tópicos utilizados. Por ultimo se busca describir el *Comportamiento* de los usuarios más frecuentes en términos de: organización y linealidad, dos variables asociadas con la comprensión que los usuarios desarrollan en su interacción con los recursos hipermediales (McEneaney, 2000), y el comportamiento de entrada al portal.

c. Estructura Externa

Los portales educativos, forman parte del universo de sitios y páginas que conforman la Web. Baeza-Yates & Castillo (2005) comprobaron que los sitios Web independientemente del contexto económico, histórico o cultural se organizan en base a una ley de potencias y no a una distribución aleatoria. En estudios de la Web de dominio .edu para países de América latina se encontró que también se cumple la ley de potencias (Bordignon & Tolosa, 2006; Tolosa & Bordignon, 2006; Tolosa, Bordignon et al., 2006). Esta ley implica que existen pocos sitios con muchos enlaces y muchos sitios con pocos enlaces. Los sitios con muchos enlaces hacia otros sitios constituyen *Hubs* (Kleinberg, 1998) que tienen la función de conducir a la audiencia, mientras que los que reciben muchos enlaces son llamados *Authorities* que tienen la función de ser referentes de un tema gozando de alto prestigio y popularidad. La estructura de relaciones o links entre sitios (o dominios) encierra un gran monto de "juicio humano" acerca de la autoridad de

[5] *El "usuario único" se trata de un [equipo+browser] únicos. Durante el procesamiento se descartar robots, arañas y spiders que realizan actividades automáticas en los sitios Web.*
[6] *Una alternativa a la metodología [IP + Path + URL irrepetible + cookie] es utilizar la metodología de taggs combinadas con cookies.*

los mismos. Cabe preguntarse, ¿Qué rol están cumpliendo los portales educativos integrantes de RELPE?, ¿son buenos hubs?, ¿son buenas authorities?

El estudio propuesto en este apartado se vincula con entender las relaciones y los roles que los portales educativos mantienen con su entorno virtual y en especial con el sub-conjunto de Sitios Web Educativos. Para este análisis resulta relevante la dimensión de dicho espacio Web y la forma de determinarlo. El análisis de esta propiedad implica la construcción de un grafico de redes de relaciones entre el sitio y un cierto espacio Web. Siguiendo la aproximación de los trabajos de Baeza-Yates & Castillo (2005) y Bordignon & Tolosa (2006) es posible caracterizar los espacios Web educativos[7] como si fueran redes.

Representar a los portales educativos por las relaciones de autoridad (authoritativeness) o conducción (hubness) que mantienen con su entorno implica describirlos por sus propiedades de "ubicuidad" es decir, por las propiedades de ser ubicables y visibles para usuarios reales tanto como para robots de búsqueda. En la medida que los portales posean la propiedad de ser ubicables, serán accesibles a quienes navegan la Web. Este aspecto resulta fundamental para comprender el uso que se hace de los contenidos y servicios que ofrecen y como elemento para la gestión de los propios portales.

[7] *En un nivel de análisis de Dominios Web (DomainGraph) estudiaron la distribución de los grados en las redes que se formaron.*

IV. PROPUESTA METODOLOGICA DEL COMPONENTE AUTOMATICO

La propuesta metodológica que se presenta a continuación incluye los siguientes puntos: 1º) la definición operacional de las **métricas** que se quieren obtener para cada dimensión y aspecto así como 2º) los criterios para determinar **procedimientos** confiables para la obtención[8] automática de dichos datos. Asimismo se propone que una definición completa de las métricas y los procedimientos deben lograrse luego de la implementación de un piloto que permita contar con una colección de datos a partir de los cuales conocer:

- las variaciones entre los países de la región y a lo largo del tiempo que permitirá obtener medidas de progreso de los portales a lo largo del tiempo,
- evaluar su validez y seleccionar las métricas que aportan mas información,
- establecer los valores de referencia para la interpretación de los resultados.

A partir de las redefiniciones y ajustes derivados de la fase piloto se deberá definir el diseño final de las operaciones que marcarán el funcionamiento del Observatorio. El mismo deberá estar acompañado de una estructura institucional que sustente las operaciones del mismo.

1º Definición de Métricas

Siguiendo las nociones del estándar [ISO14598-1], las métricas se definen por la escala y el método que se utiliza para la medición de un atributo (ISO/IEC, 1999). La escala es el conjunto de propiedades definidas que permite representar el estado del atributo en tanto el método de medición involucra la secuencia lógica de operaciones para la obtención de medidas.

Dimensiones	Aspectos	Métricas
(a) Características Intrínsecas	Contenido	Tamaño, Contenidos Unívocos, Edad, Diversidad y Riqueza de los Tópicos.
	Estructura Interna	Profundidad, Densidad, Navegabilidad y Linealidad.
	Tecnología	Complejidad de la Tecnología, Orientación del Portal.
	Funcionamiento	Integridad, Tiempo de Respuesta Global.
(b) Portal en Uso	Trafico	Nivel de Demanda, Asiduidad de la Demanda, Nivel de Actividad, Duración,
	Uso de contenidos	Diversidad y Riqueza de los Tópicos demandados.
	Comportamiento interno	Complejidad y Linealidad de la Navegación
	Comportamiento hacia el portal	Comportamiento de Entrada.
(c) Estructura Externa	Estructura Externa	Autorithativeness, Hubness del Portal.

[8] *Puede verse un detalle de los datos requeridos en el Anexo 8.*

En este apartado se presentan las definiciones operativas de los atributos que componen los aspectos y dimensiones consideradas para caracterizar a los portales educativos. A continuación se presentan las métricas que permiten su medición y que se incluyen por cumplir la condición de ser pasibles de ser obtenidas a partir procedimientos de recolección y procesamiento automatizados.

a. Características Intrínsecas

i. Tecnología:

- *Complejidad de la Tecnología*.
  Métrica: número de páginas dinámicas sobre el total de páginas del portal. Se entiende por páginas dinámicas a aquellas que tienen la extensión php, jsp, asp, shtml y otras que correspondan a tecnologías equivalentes.
  Objetivo: permite conocer en que medida un portal posee una tecnología más compleja.
  Denominación: CPLJ
  Interpretación: 0<= COMPL <=1, cuanto más próxima a 1 el sitio es mas complejo.
  Método de cálculo (fórmula):

$$CPLJ = \frac{PgD}{TPg}$$

PgD = número de páginas dinámicas.
TPg = número total de páginas del sitio

  Escala: cocientes.
  Unidad: proporción de páginas que utiliza una tecnología compleja.

- *Orientación del Portal*.
  Métrica: número de enlaces a documentos no-html sobre el total de enlaces del portal. Se consideran[9]: enlaces no-html aquellos que permiten acceder a material de Video (swf, mpg, avi, mov, qt), Audio (mp3, mid, wav, ram), software (deb, exe, rpm, iso), comprimidos (gz, tar, rar, zip) y documentos (pdf, doc, ps, txt).
  Objetivo: permite conocer en que medida el portal posee contenidos que no pueden ser vistos por los buscadores. Este atributo esta relacionado con la visibilidad del portal puesto que en la medida que la proporción de documento no HTML sea mayor, los buscadores tendrán menos oportunidades para indexarlos.
  Denominación: ORIENT
  Interpretación: 0<= ORIENT <=1, cuanto más próxima a 1 el sitio está más orientado a documentos no-html que dificultan la visibilidad del portal.
  Método de cálculo (fórmula):

[9] *Bordignon & Tolosa Bordignon, F., & Tolosa, G. (2006). Caracterización de Espacios Webs Educativos Sudamericanos. Enlace Informático, 5(1). analizaron también el código fuente (c, h, cc, java, js, sh)*

$$ORIENT = \frac{N\_Hlink}{Tlink}$$

N_Hlink = número de enlaces no-html del sitio
Tlink = número total de enlaces del sitio

Escala: cocientes.
Unidad: proporción de los enlaces que llevan a documentos que no son visibles para los robots de búsqueda.

ii. Funcionamiento:

- *Integridad*.
  Métrica: número de páginas con status equivalentes de "descarga correcta" sobre el total de páginas del portal. Para esto se utilizan como datos los códigos de estado devueltos cuando el robot hace la petición HTTP al servidor (ver Anexo 4).
  Objetivo: es una medida de buen funcionamiento general del portal a partir del funcionamiento de las páginas que lo componen.
  Denominación: INTGR
  Interpretación: 0<= INTGR <=1, cuanto más próximo es el valor a 1 el sitio tienen un mayor nivel de integridad en su funcionamiento.
  Método de cálculo (fórmula):

$$INTGR = \frac{PgOK}{TPg}$$

PgOK = número de páginas descaragadas correctamente
TPg = número total de páginas del sitio

Escala: cocientes.
Unidad: proporción de páginas que tienen un funcionamiento correcto.

- *Tiempo de Respuesta global*.
  Métrica: número de páginas accedidas en tiempo igual o menor al parámetro sobre el total de páginas.
  Objetivo: evalúa al sitio según el número de páginas accedidas en tiempo igual o menor a un parámetro definido por una medida de tiempo para el nivel más mas bajo de conectividad (ej: 8 segundos para un modem de 28.8Kb – Ver Anexo 5). El tiempo de descarga de las páginas es obtenido por un robot de testeo. En la medida que el sitio responda en un tiempo acorde a los niveles de conectividad de los usuarios aumenta la facilidad de uso (usabilidad) y las posibilidades de ser usado (uso). Se busca tener una medida de funcionamiento relacionada con la tecnología del usuario.
  Denominación: T_RESP
  Interpretación: 0<= T_RESP <=1, cuanto más próximo es el valor a 1 significa que mayor es la proporción de páginas que responden al estándar mínimo.
  Método de cálculo (fórmula):

$$T\_RES = \frac{PgTe}{TPg}$$

PgTe = número de páginas descargadas en tiempo estándar
TPg = número total de páginas del sitio

Escala: cocientes.
Unidad: proporción de páginas que tienen un tiempo de respuesta aceptable.

iii. Contenido del Sitio:

- *Tamaño de las Páginas*:
  Métrica: tamaño promedio de las páginas del portal que se obtiene por la suma de los tamaños individuales de las páginas sobre el número de páginas que lo componen. El tamaño de las páginas se mide en KB, teniéndose en cuanta solo las páginas HTML, sin incluir imágenes ni otros objetos. (El software de captura de datos puede ser configurado para bajar hasta un cierto límite de KB. El límite de seguridad especificado no debe causar sesgo de la métrica).
  Objetivo: permite conocer el tamaño de las páginas del portal, lo que incide sobre la visibilidad o factor de recuperación del sitio por parte de los robots de búsqueda (Bordignon & Tolosa, 2006).
  Denominación: TAMAÑO.
  Interpretación: cuanto > es el valor < es el factor de recuperación.
  Método de cálculo (fórmula):

$$TAMAÑO = \frac{\Sigma_{i=1,n} Tm(i)}{TPg}$$

Tm (i)= Tamaño página (i)
TPg = número total de páginas del sitio

  Escala: cocientes.
  Unidad: KB por página.

- *Número de Contenidos Digitales*
  Métrica: Es una medida de la abundancia de contenidos en la colección del portal. Cada contenido está identificado unívocamente por su URL generado automáticamente por un CMS por lo que contiene según la clasificación RELPE el meta-name [DC:identifier][10]. Para calcular esta métrica se extrae esta información y se cuenta el número de identificadores unívocos.
  Objetivo: Conocer la cantidad de contenidos unívocos que posee un portal.
  Denominación: CONT
  Interpretación: Cuanto > es el valor, mayor es la cantidad de contenidos.
  Método de cálculo (fórmula):

$$CONT = CtU$$

CtU= Total de Contenidos Unívocos

[10] *Ver Documento Técnico Nº2"Descripción de la Solución Tecnológica para el Intercambio de Contenidos en el Contexto RELPE"* http://ww2.relpe.org/Relpe/documentos

Escala: cocientes.
Unidad: número de contenidos unívocos.

- *Edad Promedio del contenido*:
  Métrica: sumatoria de los días de vida del contenido sobre el número de contenidos unívocos. Los días de vida del contenido se obtienen por la diferencia calculada entre la fecha de publicación del contenido que figura en el meta-name [DC:date] y la fecha al momento del calculo de la métrica. Se establece el control de eliminar las fechas anteriores a la creación de la Web y aquellas que hagan referencia al futuro.
  Objetivo: conocer el tiempo de vida del contenido.
  Denominación: EDAD
  Interpretación: cuanto > es el valor en días > es el tiempo que ha permanecido el mismo.
  Método de cálculo (fórmula):

$$EDAD = \frac{\sum_{i=1,n} Dv(i)}{CtU}$$

Dv (i) = Dias de vida del contenido (i)
CtU= Total de Contenidos Unívocos

  Escala: intervalo.
  Unidad: días de vida por contenido.

- *Diversidad de Tópicos*
  Métrica: Se propone utilizar el índice de diversidad de Simpson utilizado en investigación demográfica para datos categóricos.
  Como se indicó anteriormente cada contenido está identificado unívocamente por su URL generado automáticamente por un CMS y contiene también un campo que identifica el tipo de recursos educativo según la clasificación RELPE [RELPE:Type][11]. Se extrae esta información para conformar una colección de tópicos del portal. Cada tópico puede ser pensado como una especie y el conjunto de tópicos RELPE puede ser definido como un ecosistema.
  Objetivo: permite obtener una medida de la abundancia proporcional de los tópicos en la colección de contenidos del portal.
  Denominación: D(T)
  Interpretación: 0<= D(T) <=1, cuanto más próximo es el valor a 1 significa que los tópicos presentan mayor heterogeneidad.
  Método de cálculo (fórmula):

$$D(T) = 1 - \sum_{p=1,n} p(i)^2$$

p(i)= porcentaje de individuos u objetos en el tópico (i).
N= número de tópicos RELPE

  Escala: intervalo.
  Unidad: grados de heterogeneidad.

[11] *Ver criterios de catalogación RELPE en Documento Técnico Nº2"Descripción de la Solución Tecnológica para el Intercambio de Contenidos en el Contexto RELPE"* http://ww2.relpe.org/Relpe/documentos

- *Riqueza de Tópicos*.
  Métrica: número de tópicos utilizados por un portal sobre el total de tópicos presentes entre los criterios de catalogación RELPE.
  Objetivo: permite obtener una medida de cobertura de los tópicos RELPE en cada portal.
  Denominación: R(T)
  Interpretación: proporción de los tópicos RELPE utilizados por el portal.
  Método de cálculo (fórmula):

$$R(T)= \Sigma_{i=1,n}\, t(i)/T(R)$$

t(i)=tópico *i* de la clasificación RELPE
T(R)=Total de tópicos presentes en los criterios de catalogación de RELPE

  Escala: cocientes.
  Unidad: proporción de tópicos utilizados.

iv. Estructura Interna:

- *Profundidad del portal.*
  Métrica: promedio de enlaces que se deben seguir para llegar a una página desde la página principal (profundidad 0) del sitio.
  Objetivo: este factor incide sobre la visibilidad o factor de recuperación del sitio por parte de los robots de búsqueda. El software de recolección de links se puede limitar en cuanto a la profundidad que analiza para optimizar recursos de procesamiento. Tolosa & Bordignon (2006) limitaron la recolección de datos a quince niveles de profundidad para páginas estáticas y a cinco para dinámicas. Tomando en cuenta su estudio, la distribución de la profundidad de los sitios Web en los diferentes países de Sudamérica se podría limitar a una profundidad de diez. Los niveles más profundos se encuentran en los portales de Argentina, Perú y Uruguay (Petricek, Escher et. al (2006) lo limitan a dieciocho para páginas estáticas).
  Denominación: PROF
  Interpretación: cuanto > el valor, > es la profundidad del portal.
  Método de cálculo (fórmula):

$$PROF=\frac{\Sigma_{i=1,n}\, Path0(i)}{Npath0}$$

Path0 (i) = número de enlaces a seguir desde el nivel 0 para el path (i).
N Path0 = Total de Path establecidos desde el nivel 0

  Escala: cocientes.
  Unidad: enlaces por path desde el origen.

- *Densidad* [12](Usabilidad):
  Métrica: para caracterizar la densidad Bordignon et. al proponen utilizar como medida de densidad el número de enlaces existentes normalizado por el total de enlaces posibles (es un valor entre 0 y 1).
  Objetivo: permite caracterizar a una red en función de las relaciones que mantienen sus páginas entre si. (sin considerar la dirección de la relación). Cuanto más corta es la distancia de los path, es decir cuanto más cohesionado un sitio mayor cantidad de links posee y por tanto mayor es su densidad (Petricek, Escher et al., 2006).
  Denominación: DENS
  Interpretación: 0<= DENS <=1, cuanto más próximo es el valor a 1 significa que mayor es la cohesión del portal.
  Método de cálculo (fórmula):

$$DENS = \frac{Tlink}{T(t)link}$$

  Tlink = número total de enlaces del sitio
  T(t)link = número total de enlaces posibles teóricamente en el sitio

  Escala: cocientes.
  Unidad: proporción de enlaces sobre el total teórico.

- *Navegabilidad* (Usabilidad):
  Métrica: Esta métrica es calculada a partir de una matriz de distancia que representa a un grafo dirigido. El grafo se forma sobre la base de los enlaces jerárquicos luego de que se eliminaron los enlaces de referencias cruzadas (método breadth-first spanning tree algorithm). A partir de la matriz de distancias que se conforma es posible calcular la "converted distance" o la suma de las distancias (Out-Distance o In-Distance) de cada nodo del grafo respecto a los demás. La navegabilidad o "compactness" consiste en la diferencia del valor máximo posible y el valor real de la "converted distance" sobre la diferencia entre el valor máximo y mínimo del mismo factor. ((Dhyani, Keong NG et al., 2002).
  Objetivo: Un alto grado de compactness indica una buena navegabilidad lo que significa que cada nodo puede ser fácilmente accedido desde otros nodos (Pahl, 2001). Un grafo completamente desconectado tiene un valor de cero mientras que uno totalmente conectado tiene un valor de uno.
  Denominación: Cp
  Interpretación: los valores extremos deben ser evitados para lograr un adecuado nivel de lectura y navegabilidad.
  Método de cálculo (fórmula):

$$Cp = \frac{Max - \sum_{i=1} \sum_{j=1} C\,ij}{Max - Min}$$

  C = matriz de distancia.
  Max = (N2 – N) * K
  Min = (N2 – N)
  Nn = número de nodos (ej: páginas)
  K = valor constante

  Escala: intervalo.

[12] *Ver Betwenness, Eigenvector*

Unidad: proporción de nodos con referencias cruzadas en el grafo.

- *Linealidad* (Usabilidad).
  Métrica: la linealidad de un sitio se mide a través del "stratum". Esta métrica se basa en el concepto de prestigio. El prestigio de un nodo es la diferencia entre su status (o la suma de distancia a todos los demás nodos) y su contra-status (de la suma de distancias desde los otros nodos). El prestigio absoluto de todos los nodos es la suma de los valores absolutos del prestigio de cada nodo. El stratum se obtiene normalizando el prestigio absoluto de todos los nodos por el prestigio absoluto lineal del grafo (LAP). El LAP se obtiene elevando el número de nodos (N) al cubo (si es impar se calcula N al cubo – N) dividido N-1 (Dhyani, Keong NG et al., 2002).
  Objetivo: esta métrica revela cuan organizado es un sitio y cuantas elecciones tiene que hacer el usuario durante la navegación (Pahl, 2001).
  Denominación: St
  Interpretación: cuanto mayor es el valor del stratum más lineales son los sitios y menos decisiones deben enfrentar los usuarios. Un valor bajo esta indicando que la cantidad de decisiones que enfrenta un usuario es muy alta siendo el siendo el sitio poco estructurado. Los sitios altamente lineales son tediosos de navegar.
  Método de cálculo (fórmula):

$$St = \frac{\Sigma_{i=1,n} | Ptg |(i)}{LAP}$$

$$LAP = \frac{N3}{4} \quad \text{Si N es par}$$

$$LAP = \frac{N3 - N}{4} \quad \text{Si N es inpar}$$

Ptg = el prestigio es la diferencia entre el Status y Contra-Status de un nodo.
Status = suma de las distancias del nodo (i) respecto a los demás nodos.
Contra-Status = suma de las distancias desde los demás nodos hacia el nodo (i).
LAP = Linear Absolute Prestige

  Escala: intervalo.
  Unidad: proporción de nodos que siguen un esquema lineal en el grafo.

b. Portal en Uso

i. Tráfico:

- *Nivel de Demanda del portal*.
  Métrica: se mide a través del número de visitas[13] recibidas por el portal en un cierto periodo de tiempo descontando las visitas provenientes de agentes

[13] *Autores como Villena, J., J. Gonzalez, et al. (2002) definen a la Visita como Sesión de Usuario en Servidor, quienes definen a la Sesión de Usuario como la agrupación de todas las páginas que visita un mismo usuario durante la visita a un sitio Web.*

automáticos (crawlers y arañas). La visita puede ser definida como el conjunto de secuencias de vistas de página, click-stream, requeridas a un portal en particular durante una sesión de usuario (Lavoie & Frystyk Nielsen, 1999). Se considera que una sesión ha concluido cuando se produce un período de inactividad superior a 30 min. y no existe una sucesión lógica de las páginas visitadas cuando se supera ese período de inactividad.
Objetivo: permite tener una medida gruesa de demanda del portal por parte de los usuarios sin considerar la actividad que realicen dentro del mismo.
Denominación: DEMAN
Interpretación: cuanto > es el número de visitas > es la demanda.
Método de cálculo (fórmula):

$$DEMAN = Vt$$

Vt = Total de Visitas para el periodo (t)

Escala: cocientes.
Unidad: número de visitas.

- *Nivel de Asiduidad de la Demanda.*
Métrica: número promedio de visitas al portal por visitante único para un cierto periodo de tiempo descontando las visitas provenientes de robots, crawlers y arañas. El usuario único se define por la identidad del equipo+browser+cookie y no como un individuo real.
Objetivo: permite tener una medida de cuan asidua es la demanda al portal ya que considera la repetición de visitas por parte de un mismo usuario en un periodo de tiempo.
Denominación: ASID.
Interpretación: cuanto > es el valor, > asidua es la demanda.
Método de cálculo (fórmula):

$$ASID = \frac{Vt}{Visitantes}$$

Vt = Total de Visitas para el periodo (t).
Visitantes = total de visitantes únicos, entendido como la combinación de parámetros (ej: "equipo+browser+cookie") para el periodo (t)

Escala: cocientes.
Unidad: número promedio de visitas por visitantes.

- *Nivel de Actividad de los usuarios del portal*.
Métrica: se mide a través del promedio de vistas de página por visita calculado a partir de la suma de vistas de páginas o clic-stream sobre el número total de visitas en un cierto periodo de tiempo. Se asume que se han filtrado los requerimientos realizados por robots, arañas, etc.
Objetivo: permite describir la demanda al portal por medio de una medida de la actividad que realizan los usuarios del portal.
Denominación: ACTIV
Interpretación: cuanto > es el valor, > es el requerimiento que realizan los usuarios de las páginas del portal.

Método de cálculo (fórmula):

$$ACTIV=\frac{\Sigma i=_{1,n} VsPg(i)}{Vt}$$

VstPg = número de vistas de páginas en la visita (i)
Vt = Total de Visitas para el periodo (t).

Escala: cocientes.
Unidad: número de vistas por página.

- *Duración de la Actividad.*
  Métrica: se mide a través del promedio de duración de las visitas calculado a partir de la suma de la duración de las vistas de páginas individuales sobre el número total de visitas en un cierto periodo de tiempo. Se asume que se han filtrado los requerimientos realizados por robots, arañas, etc.
  Objetivo: permite describir la demanda al portal por medio de una medida de la duración de la actividad que realizan los usuarios del portal.
  Denominación: DURAC
  Interpretación: cuanto > es el valor, > es el requerimiento que realizan los usuarios de las páginas del portal.
  Método de cálculo (fórmula):

$$DURAC=\frac{\Sigma_{i=1,n} DVt(i)}{Vt}$$

D Vt = suma de los tiempos de duración de las vistas de páginas en la visita (i).
Vt = Total de Visitas para el periodo (t).

Escala: intervalo.
Unidad: número de minutos que dura la visita promedio.

ii. Uso de los Contenidos:

- *Diversidad de Tópicos Usados*
  Métrica: Se propone utilizar el mismo índice de diversidad de Simpson señalado para el análisis de contenido. La diferencia consiste en que se toman la información de [RELPE:Type][14] proveniente de las URLs registradas entre los datos de uso.
  Objetivo: permite obtener una medida de la diversidad de los tópicos utilizados de acuerdo a los criterios que conforman la colección de contenidos del portal
  Denominación: D(Tu)
  Interpretación: 0<= D(Tu) <=1, cuanto más próximo es el valor a 1 significa que los tópicos utilizados presentan mayor heterogeneidad.
  Método de cálculo (fórmula):

$$D(Tu)=1-\sum_{p=i}^{N} (pi)^2$$

p(i)= porcentaje de individuos u objetos en el tópico (i) utilizado.
N= número de tópicos RELPE

[14] *Ver criterios de catalogación RELPE en Documento Técnico Nº2"Descripción de la Solución Tecnológica para el Intercambio de Contenidos en el Contexto RELPE"* http://ww2.relpe.org/Relpe/documentos

Escala: intervalo.
Unidad: grados de heterogeneidad.

- *Riqueza de Tópicos Usados*.
  Métrica: número de tópicos requeridos por los usuarios de un portal sobre el total de tópicos presentes entre los criterios de catalogación RELPE.
  Objetivo: permite obtener una medida de proporción de los tópicos RELPE efectivamente utilizados por los usuarios del portal.
  Denominación: R(Tu)
  Interpretación: cuanto > es el valor, > es el uso de los tópicos RELPE.
  Método de cálculo (fórmula):

$$R(Tu)= \Sigma_{i=1,n} Tu(i)/T(R)$$

t u=tópico utilizado de la clasificación RELPE que haya recibido peticiones de los usuarios
T(R)=Total de tópicos presentes en los criterios de catalogación de RELPE

  Escala: cocientes.
  Unidad: proporción de tópicos utilizados.

iii. Comportamiento de los usuarios:

Se basa en la caracterización previa de los tipos de usuario de acuerdo a su nivel de actividad (ej: livianos o pesados[15]). El nivel de actividad se establece por el número de visitas realizadas en un periodo de tiempo por un usuario único. En algunos estudios el valor de corte es de 19 días (Bolger & Mörn, 2004). Las visitas seleccionadas según este criterio pueden ser representadas a través del concepto de Path o ruta de navegación. El Path se define como el grupo de datos de registro de las páginas vistas por un sujeto durante una sesión y son la base para obtener las métricas de comportamiento del usuario. Para esto se propone utilizar adaptaciones de las métricas de Compactness y Stratum vistas para el análisis de la Estructura Interna del sitio. Aplicadas al Path de navegación se llaman Path Compactness (P*Cp*) y Path Stratum (P*St*) (McEneaney, 2000).

- *Complejidad de la navegación*:
  Métrica: la diferencia del valor máximo posible y el valor real de la "converted distance" sobre la diferencia entre el valor máximo y mínimo del mismo factor para un grafo con dirección. La "converted distance" se obtiene de la suma de las distancias (Out-Distance o In-Distance) de cada nodo respecto los demás que se calcula a partir de la matriz de distancias. La matriz de distancias, en este caso, se construye a partir del grafo conformado por la actividad de navegación de los usuarios.
  Objetivo: permite medir el grado complejidad del Path seguido por los usuarios.
  Denominación: PCp

[15] *Ver metodología ComScore.*

Interpretación: 0<= PCp <=1, los estudios han demostrado que los sujetos que tienen éxito en procesos de búsqueda y en pruebas de comprensión presentan paths con altos valores de P*Cp*.
Método de cálculo (fórmula):

$$PCp = \frac{Max' - \sum_{i=1} \sum_{j=1} C'_{ij}}{Max' - Min'}$$

C' = Matriz de Distancia correspondiente al grafo que representa la navegación de los usuarios.
Max' = (N2 – N) * K
Min' = (N2 – N)
Nn' = número de nodos (ej: páginas)
K' = valor constante

Escala: intervalo.
Unidad: nivel de complejidad de la navegación de los usuarios.

- *Linealidad de la navegación (PSt)*:
  Métrica: Esta métrica se basa en el concepto de prestigio de un nodo, que es la diferencia entre su status (o la suma de distancia a todos los demás nodos) y su contra-status (de la suma de distancias desde los otros nodos). El prestigio absoluto de todos los nodos es la suma de los valores absolutos del prestigio de cada nodo. El stratum se obtiene normalizando el prestigio absoluto de todos los nodos por el prestigio absoluto lineal del grafo (LAP). El LAP se obtiene elevando el número de nodos (N) al cubo (si es impar se calcula N al cubo – N) dividido N-1. Estos valores se obtienen en este caso, a partir del grafo conformado por la actividad de navegación de los usuarios.
  Objetivo: permite medir la linealidad u organización del Path seguido por los usuarios.
  Denominación: PSt
  Interpretación: 0<= PSt <=1. Los estudios han demostrado que los sujetos que tienen éxito en procesos de búsqueda y en pruebas de comprensión presentan paths con bajos valores de P*St*.
  Método de cálculo (fórmula):

$$PSt = \frac{\sum_{i=1,n} | Ptg' |(i)}{LAP'}$$

$$LAP' = \frac{N3}{4} \quad \text{Si N es par}$$

$$LAP' = \frac{N3 - N}{4} \quad \text{Si N es inpar}$$

Ptg' = el prestigio es la diferente entre el Status y Contra-Status de un nodo.
Status' = suma de las distancias del nodo (i) respecto a los demás nodos.
Contra-Status' = suma de las distancias desde los demás nodos hacia el nodo (i).
LAP' = Linear Absolute Prestige

  Escala: intervalo.
  Unidad: nivel de linealidad de la navegación de los usuarios.

- *Comportamiento de entrada*:
  Métrica: número de visitas que tienen origen externo (provienen de remitentes externos al portal, como ser otros sitios o robots de búsqueda) sobre el total de visitas.

Objetivo: conocer la proporción de visitas que se originan fuera y dentro del propio portal como aproximación al reconocimiento y afiliación de los usuarios.
Denominación: ENT.
Interpretación: 0<= ENT <=1. Cuanto < es el valor, > es la afiliación de los usuarios, quienes recurren al portal para iniciar la búsqueda.
Método de cálculo (fórmula):

$$ENT = \frac{Vt\ ext}{DEMAN}$$

Vt ext = número de visitas con origen externo

Escala: cocientes.
Unidad: porcentaje de visitas con comportamiento de entrada exógeno.

c. Estructura Externa

- *Popularidad* – authoritativeness (visibilidad):
  Métrica: posición del portal en una distribución de nodos según el *grado entrante* (Ge) o número de enlaces que reciben desde los otros sitios considerados. El número de enlaces que recibe cada nodo se calcula a partir de la matriz de relaciones construida en base al sub-grafo S*e* (ver parágrafo IV. c. más adelante)
  Objetivo: Permite conocer en que medida los autores están dando una recomendación acerca del contenido del portal cuando general links hacia este, por lo que el número de enlaces entrantes se convierte en un indicador de popularidad o prestigio.
  Denominación: Ge
  Interpretación: Cuanto > el valor, > es la popularidad del portal
  Método de cálculo (fórmula):

$$Ge(i) = \Sigma_{j=} Cij$$

C = Matriz de relaciones
i = eje i de la matriz C
j= eje j de la matriz C

  Escala: ordinal.
  Unidad: posiciones respecto al primer lugar.

- *Centralidad* – Hubness :
  Métrica: posición del portal en una distribución de nodos según el *grado saliente* (Gs) o número de enlaces que emite hacia los otros sitios considerados. El número de enlaces que emite cada nodo se calcula a partir de la matriz de relaciones construida en base al sub-grafo S*e* (ver parágrafo IV. c. más adelante)
  Objetivo: Permite conocer en que medida el portal se constituye en un buen Hub que orienta a su audiencia hacia otros sitios (S Chakrabarti, Dom et al., 1999).
  Denominación: Gs
  Interpretación: Cuanto > el valor, > es la capacidad del portal de conducción de su audiencia.

Método de cálculo (fórmula):

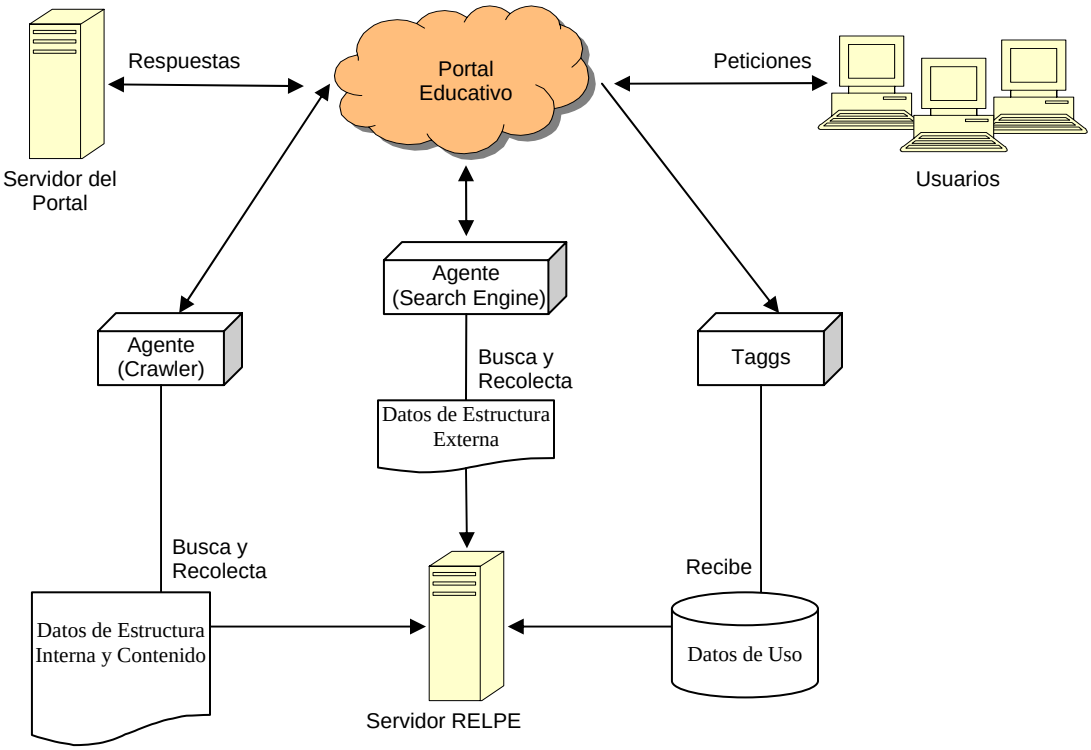
$$Gs(i) = \sum_{J} {}_{j=} Cij$$

C = Matriz de relaciones
i = eje i de la matriz C
j= eje j de la matriz C

Escala: ordinal.
Unidad: posiciones respecto al primer lugar.

2º Obtención de Datos

A continuación se presentan algunos criterios para la obtención de datos validos que sirvan para el cálculo de las métricas definidas. Algunos de estos criterios se corresponden con herramientas por lo que las mismas serán tomadas para ejemplificar lo que no implica una recomendación para la implementación.



Como primer criterio general proponemos que, cualquiera sea el caso de esquema de implementación resultante, se debe garantizar la uniformidad de los procedimientos de obtención de datos para todos los casos. Con este fin se recomienda el desarrollo previo de una estructura institucional que pueda centralizar los procedimientos de obtención de datos. Esta propuesta conlleva el beneficio de evitar que los procedimientos de obtención de datos queden a cargo de las mismas unidades que serán estudiadas. En este sentido, se estima que una aproximación basada en "agentes" y "taggs" permitiría obtenerlos de forma automática con mínima intervención de las unidades de análisis.

La obtención de las métricas requiere de tres grupos de datos, para cada uno de los cuales es posible definir criterios de obtención. A continuación se presentan estos criterios para cada grupo:

a. Datos de Estructura y Contenido del Portal

La obtención de datos de estructura interna del portal puede ser realizada por medio de un agente (crawler) que recorra los dominios correspondientes a los Portales Educativos RELPE[16] para obtener las listas de URLs y links asociados. Se recomienda que el agente debiera iniciar su operación en la página principal y hasta una profundidad de quince niveles para páginas estáticas y a cinco para dinámicas. Estos criterios pueden ser redefinidos luego del piloto. A partir de los listados de URL pueden obtenerse los meta-tags [DC:identifier], [DC:date], [RELPE:Type].

b. Datos de Uso del Portal

La "metodología de Taggs" o "etiquetas" por la cual se inserta líneas de código en las páginas del portal que se activan cuando los usuarios acceden al mismo enviando la información de uso a una base de datos. Esta metodología evita la mayoría de los sesgos que se presentan en la recolección de datos de uso. La obtención de métricas de uso presenta una serie de problemas de validez y confiabilidad según cuáles sean: (a) la fuente de los datos y (b) el procedimiento utilizado para su recolección. En este trabajo se recomienda utilizar el registro mediante la metodología de "etiquetas" o "Taggs". Las etiquetas son líneas de código que ejecutan una acción de llamada cuando la página es bajada al browser y envían información del visitante a una base de datos central que los dispone inmediatamente para el análisis. Se trata de una técnica de recolección de datos orientada al cliente que se inicia cuando un usuario hace una petición de una página a un servidor y este entrega la página que contiene la llamada o link a la página de huella enviando información del visitante (NetIQ Corporation, 2004).

Esta técnica permite resolver *los problemas de identificación de visita/transacciones* haciendo que no sea necesario la utilización de URL irrepetible, re-heading propios del análisis de archivos Log, *la sub-estimación*[17] asociados a los efectos caché y Proxy así como *gestión de los datos* puesto que se genera un único "hit" o registro por cada página vista lo que disminuye considerablemente el tamaño del archivo y los datos están inmediatamente disponibles para el análisis. La implementación de esta forma de recolección de datos debe cumplir con los criterios de validez que se detallan en el Informe Técnico Nº 2 (Noguera, 2007). Esto incluye programar la transmisión del dato referrer de modo que se puedan identificar robots y arañas y las páginas de procedencia del visitante, se debe verificar que no amenace la performance del sitio. También se recomienda su uso siempre que sea posible que los portales trabajen con un CMS

[16] *Algunos de los robots de investigación de código abierto disponibles son: WIRE, Socsibot o Nutch.*
[17] También permite registrar actividad de aplicaciones flash

(Content Management System) para facilitar la gestión de las etiquetas[18] así como verificar las regulaciones existentes en cada país para el uso de esta metodología.

c. Datos Estructura Externa del Portal

Caracterizar la relación de un portal educativo con su entorno supone demarcarlo previamente. Este entorno consiste en un sub-conjunto de la Web sobre el cual se establecerán las métricas que relacionen al sitio con su entorno y servirá de base para visualizar la red macro en la que está inserto el portal. Este sub-conjunto consiste en un listado de URLs y links ordenados con cierto formato. La metodología desarrollada con este propósito por Kleinber et.al. (1998) apunta a construir un sub-grafo focalizado S (i) relativamente pequeño, (ii) rico en páginas relevantes y (iii) que contenga la mayoría de los más fuertes authorities.

Para esto se propone utilizar un robot de búsqueda (altavista, google), al cual se envía una cadena de términos que represente al contenido del portal. Esta acción arroja una colección de URLs de las cuales se extraen las primeras *t* páginas con mayor ranking[19]. Estas comprenden un conjunto raíz R o root set que satisface solo los dos primeros requisitos (i e ii).debe ser expandido para producir el conjunto S o conjunto base a partir del cual se calcula el grafo G.

La lista R cuenta con las primeras 200 referencias mejor rankeadas en el proceso de búsqueda y se expande a través de los links que entran y salen de cada referencia. Chakrabarti, Dom et al.(1998) propusieron elevar este proceso de expansión a dos etapas, de manera de incluir referencias a documentos que se encuentran a una distancia de dos o menos de R. Esta metodología evita el problema de que muchas páginas prominentes no se describen a si mismas. Para cada portal educativo RELPE se debe construir un sub-grafo S. El sub-grafo que queda definido por este procedimiento a partir de descriptores que representan a un portal educativo lo denominamos S*e* o sub-grafo de web educativa.

El procedimiento de identificación de Authorities y Hubs requiere la aplicación, hasta lograr un punto fijo, de un algoritmo que rompa la circularidad y el solapamiento entre ellos. El algoritmo desarrollado por Kleinberg (1998) propone que a través de sucesivas iteraciones se reemplace el peso de autoridad de la página *p* por la suma de los pesos de centro de las páginas que apuntan a *p*, y que as su vez reemplace el peso de centro de la página *p* por la suma de los pesos de autoridad de las páginas apuntadas por *p*.

[18] *Se debería incluir un crawler que verifique la existencia de etiquetas en las páginas*
[19] *Kleinber propone las primeras 200 páginas.*

IV. BIBLIOGRAFIA:

Adhikari, V. K., & Lemone, K. (2007). *Hypertext Structural Analysis of Nepali Educational Institution Web-sites*. Paper presented at the 16th Annual World Wide Web Conference

Baeza-Yates, R., & Castillo, C. (2005). *Link Analysis in National Web Domains.* Paper presented at the Workshop on Open Source Web Information Retrieval (OSWIR), Compiegne, France.

Baeza Yates, R. (2004). Excavando la Web. *El Profesional de la Información,, 13*(1), 4-10.

Bolger, L., & Mörn, M. P. (2004). *Internet Audience Dynamics*: Double Click.

Bordignon, F., & Tolosa, G. (2006). Caracterización de Espacios Webs Educativos Sudamericanos. *Enlace Informático, 5*(1).

Botafogo, R. A., Rivlin, E., & Shneiderman, B. (1992). Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM TransactIons on Information Systems, 10*(2), 142-180.

Butcher, N. (2002). *"Best Practice in Education Portals. Final Report"*: Prepared for The Commonwealth of Learning and SchoolNet Africa.

Chakrabarti, S., Dom, B., David Gibson, Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, et al. (1999). Mining the Link Structure of the World Wide Web.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J. (1998). *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text.* Paper presented at the Proceedings of The 7th International World Wide Web Conference.

Cooley, R. (2000). The Importance of Understanding Web Site Structure and Content when Performing Web Usage Mining.

Covella, G. J. (2005). *Medición y Evaluación de Calidad en Uso de Aplicaciones Web.* Universidad Nacional de La Plata, La Plata.

Da Graca Guerra, M., & Modarelli, M. (2003). *Determinación Automática de Perfiles de Usuarios en Internet utilizando Técnicas de Data Mining.* Universidad de Buenos Aires, Ciudad de Buenos Aires.

Dhyani, D., Keong NG, W., & Bhowmick, S. S. (2002). A Survey of Web Metrics. *ACM Computing Surveys*, *34*(4), 460-503.

Díaz, P. (2003). Usability of Hypermedia Educational e-Books. *D-Lib Magazine*, *9*(3).

Filocamo, G., & Chesñevar, C. (Artist). (2003). *Formalización de Web Mining como Conocimiento Estructurado*

Franklin, T. (2004). *Portals in Higher Education: concepts & models*: The Observatory on Borderless Higher Education.

García, E. Portales educativos. Consideraciones de tipo general

Google (Ed.). http://www.google.com/support/analytics/bin/answer.py?answer=27303&query=glosario&topic=&type.

ISO/IEC. (1999). *ISO/IEC 14598-1 - Information technology — Software product evaluation*. Genève, Switzerland.

Kleinberg, J. (1998). *Authoritative sources in a hyperlinked environment.* Paper presented at the Proceedings of The Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, California.

Lauw, H. W., Lim, E.-P., Tan, T.-T., & Pang, H.-H. (2005). *Mining Social Network from Spatio-Temporal Events.* Paper presented at the SIAM International Conference on Data Mining, Newport Beach, California.

Lavoie, B., & Frystyk Nielsen, H. (Eds.). (1999) W3C Glossary Dictionary.

Libedinsky, M. (2007).

McEneaney, J. E. (1999). *Visualizing and Assessing Navigation in Hypertext.* Paper presented at the Hypertext 99, Darmstadt, Germany.

McEneaney, J. E. (2000). *Navigational Correlates of Comprehension in Hypertext.* Paper presented at the Hypertext 2000, San Antonio, TX.

Miller, C. S., & Remington, R. W. (2001). *Modeling an Opportunistic Strategy for Information Navigation.* Paper presented at the Cogsci01.

Miller, C. S., & Remington, R. W. (2004). Modeling Information Navigation : Implications for Information Architecture. *Human-Computer Interaction, 19*(3).

NetIQ Corporation. (2004). WebTrends SmartSource Data Collection – Premier Client-side Data Collection Technology. Retrieved September 2006, from http://www.arena.no/nedlasting/dokumentasjon/wt_smartsource_r3.pdf

Noguera, M. T. (2007). *Informe Técnico Nº 2: "Problemáticas Identificadas para la Obtención de Datos Confiables para el Análisis de Portales Educativos"*. Buenos Aires: Fundación Evolución.

Olsina, L. (1999). *Metodología Cuantitativa para la Evaluación y Comparación de la Calidad de Sitios Web.* Universidad Nacional de La Plata, La Plata.

Olsina, L., Lafuente, G., & Pastor, O. (2002). Towards a Reusable Repository for Web Metrics. *Journal of Web Engineering, 1*(1), 61-73.

Pahl, C. (2001). *The Evaluation of Educational Service Integration in Integrated Virtual Courses*. Paper presented at the Proceedings of the 2001 Symposium on Applications and the Internet-Workshops

Petricek, V., Escher, T., Cox, I. J., & Margetts, H. (2006). *The Web Structure of E-Government - Developing a Methodology for Quantitative Evaluation*. Paper presented at the International World Wide Web Conference Committee (IW3C2).

Poblete, B. (2004). *Herramienta de Minería de Consultas para el Diseño del Contenido y la Estructura de un Sitio Web.* Universidad Nacional de Chile, Santiago de Chile.

Poblete, B., & BaezaYates, R. (2006). *A Content and Structure Website Mining Model*. Paper presented at the WWW 2006,.

Powell, T., Jones, D., & Cutts, D. (1998). *Web Site Engineering: Beyond Web Page Design*: Prentince Hall PTR.

Scott, J. (2000). *Social Network Analysis*. London: SAGE Publications Ltd.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations, 1*(2), 12-23.

Tolosa, G. H., & Bordignon, F. R. A. (2006). *Análisis de Enlaces en el Espacio Web de las Universidades Argentinas.* Paper presented at the Proceedings of The VII Workshop de Investigadores en Ciencias de la Computación - WICC 2006, Morón, Argentina.

Tolosa, G. H., Bordignon, F. R. A., & Lavallén, P. J. (2006). *Caracterización del Espacio Web de Perú.* Paper presented at the Proceedings of The 2nd International Congress on Librarianship and Information, Lima, Perú.

Villena, J., Gonzalez, J., Barceló, E., & Velasco, J. (2002). *Minería de Uso de la Web Mediante Huellas y Sesiones*. Paper presented at the VIIth Iberoamerican Conference on Artificial Inteligence, Iberamia 2002.

W3C. Web Content Accessibility Guidelines 1.0. Retrieved November 2006, from http://www.w3.org/TR/1999/WAI-WEBCONTENT-19990505/#content-structure

WebTrends (Ed.) (2006) Glossary of Terms.

V. ANEXOS

Anexo 1 - Estudio de las Aplicaciones Web – Marco WebQEM

La aproximación seguida por el marco WebQEM para la evaluación de calidad de sitios Web adopta una visión de productos o "artefactos de software". Desde este marco los portales educativos serian artefactos de software que pueden ser evaluados por un modelo jerárquico de requerimientos de calidad y cuenta con métricas probadas y validadas para este propósito. El análisis de la *Calidad del Producto Web* considera al menos siete dimensiones (las primeras seis definidas por el estándar ISO/IEC 9126-1, la ultima se incluye para evaluar calidad de información): *Usabilidad* concepto que tiene en cuenta la perspectiva del esfuerzo requerido para usar el producto, *Funcionalidad*, *Confiabilidad*, *Eficiencia*, *Portabilidad*, *Mantenibilidad* y *Contenido*. Los requerimientos de calidad de contenido abarcan: precisión, adecuación, accesibilidad y conformidad con las normas legales de la información (L Olsina, Lafuente et al., 2002).

La metodología WebQEM ha sido adaptada a la evaluación de Calidad en Uso de un sitio Web incluyendo las dimensiones de: *Eficacia*, *Productividad* y *Satisfacción*. Esta evaluación se realiza en función de las tareas que completan los usuarios del mismo. Esta metodología toma en cuenta a la audiencia y plantea la necesidad de definir el perfil de los usuarios para llevar adelante la evaluación. Así define dos tipos de usuarios o visitantes: los expertos y los generales, estos últimos se dividen en intencionales o casuales (Covella, 2005). La obtención de datos se realiza por medio de técnicas de video-filiación, software de captura de pantalla y cuestionarios a los usuarios. En este marco, la definición de las métricas va a depender de las características de los portales en tanto dos criterios nivel de complejidad (simple/estáticas – complejas/dinámicas) y grados de orientación (a documentos – a aplicaciones). Según estos criterios los sitios de los portales pueden ser categorizados en: estáticos, estático con formularios de entrada, sitio con Acceso de datos dinámicos, sitio creado dinámicamente, aplicación de software basada en la Web (L. Olsina, 1999).

Anexo 2 – Estudios de Audiencia

La cobertura o el alcance de la población objetivo es determinada en portales comerciales por una diversidad de métodos: paneles audiométricos, paneles RDD y tracking de usuarios. Los paneles audiométricos consisten en encuestas representativas a hogares en los que se instala un software de control y permiten obtener una gran cantidad de datos sobre el comportamiento de navegación de los usuarios y relacionar esto con variables demográficas. Los paneles RDD (Random Digital Recruitment) funcionan de manera similar al método anterior con la diferencia que la muestra no se establece sobre toda la población sino sobre quienes están en el directorio telefónico. Por ultimo, el tracking de usuarios consiste en un mero análisis de uso en el que puede realiza una identificación de usuario único por medio de técnicas de análisis que combinan los registros de actividad

de los archivos log de los servidores (IP, Path) combinados con tecnología de URL irrepetible y cookies persistentes. Esta última no permite conocer datos demográficos de los usuarios a menos que se pueda asumir que la totalidad de los usuarios del sitio se encuentran registrados y han aportado datos personales.

El método de tracking de usuarios es el único que se adecua a las restricciones que enfrenta este trabajo en cuanto a que la fuente de los datos sea automatizable (esto implica la utilización de taggs o archivos log). Sin embargo no aporta información valida acerca del nivel real de cobertura de la población objetivo puesto que en los portales RELPE. Si bien puede establecer el número de visitantes únicos que recibe un portal en un tiempo determinado, no permite conocer si detrás de este usuario identificado se encuentran una o más personas reales. El tracking seria valido si se realizara sobre los usuarios suscriptos. Pero en todos los países que implementan la suscripción estos constituyen proporciones muy bajas respecto a los tamaños conocidos de las poblaciones de docentes o de alumnos. Sin embargo el método de tracking permite caracterizar patrones recurrentes de comportamiento y de uso de los portales por lo que será utilizada en la dimensión "Portal en Uso".

Una aproximación posible para conocer el nivel de cobertura de la población constituida por las comunidades educativas seria utilizar paneles audiométricos teniendo como unidad los establecimientos educativos. Esta aproximación excede los condicionamientos de este trabajo debido a que utiliza una fuerte de información que requiere de procedimientos no-automatizables como ser: definición del marco muestral, muestreo aleatorio, instalación de software de control en las unidades y relevamiento de características de las unidades.

El muestreo se debería realizar en base a una lista de establecimientos educativos, públicos y privados, de nivel básico y medio del sistema que tienen acceso a Internet. La lista se constituiría con las bases de fuentes oficiales que serviría de marco muestra. Sobre el mismo se aplicaría un muestreo aleatorio estratificado por nivel de enseñanza. Cada país debería contar con recursos humanos e instituciones, preferentemente institutos de estadísticas nacionales, que puedan llevar adelante esta tarea. El establecimiento de este marco muestral no es una tarea menor, puesto que deben estar actualizados y deben revisarse teniendo en cuenta el nivel de crecimiento de la conectividad en el sistema educativo. Esto implica contar con datos censales de infraestructura del sistema que incluya a todas las unidades educativas del país y pueda discriminar dentro de este grupo a aquellas que cuentan con equipamiento y conectividad dedicados a la enseñanza. Sobre el establecimiento de un marco confiable sobre el cual hacer la selección descansa la validez de esta aproximación.

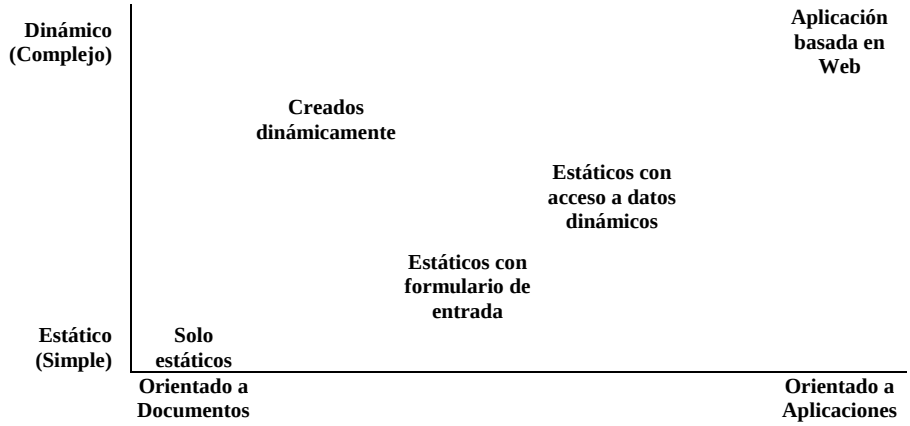
El software se debería ser instalado en cada una de las computadoras con acceso a Internet presentes en cada establecimiento registrándose en cada caso un identificador del equipo que recibió el software de control (puede ser la IP), la función del mismo (ej: solo administración, enseñanza), el número promedio de personas que tienen acceso al mismo según su rol (docente, administrativo, alumno), así como el número de veces que informan utilizar los mismos para un cierto periodo de tiempo (ej: un mes). Así mismo se

registrarían otras variables demográficas de la institución como ser: número de grados atendidos, número de alumnos por grado, género y edad de los alumnos, número de docentes, número de personal administrativo, tipo de gestión, jurisdicción, etc. Seria muy importante incluir indicadores de nivel socioeconómico en caso que esta variable no se encuentre previamente en las bases de datos de los organismos oficiales.

El software de control funcionaria enviando, vía Web, información de actividad del equipo a una base de datos externa. Al registrarse toda la actividad se podrían conocer los patrones de navegación y acceso a recursos Web que se combinarían con los datos demográficos relevados. Dado que los mismos fueron seleccionados por procedimientos que garantizan representatividad los resultados de alcance podrían expandirse a la población considerada en el marco muestral. Esta información, si se toman los recaudos correspondientes al elaborar el marco de muestreo, permitiría caracterizar de forma mas precisa el nivel de conectividad del sistema educativo.

Si bien esta aproximación resulta más factible que el establecimiento de paneles sobre toda la población, tiene el sesgo del "punto de acceso". Es posible que el uso de la Web se de además por otros puntos que no son la institución educativa. Por ejemplo en el caso del nivel universitario es probable que predominen otros puntos de acceso como el hogar o telecentros. Al captar el comportamiento de uso desde un único punto de acceso se esta produciendo un sesgo en la recolección de los datos. La actividad dentro de la escuela puede no representar la verdadera intensidad de uso por parte de los usuarios de nivel básico y medio, pero creemos que si puede estar representando una parte considerable de la actividad de uso para el aprendizaje dentro del sistema.

Anexo 3 – Caracterización de Sitios Web de acuerdo a Powell, Jones et al (1998)



Sitios:	Definición	Funcionalidad	Ejemplo
Solo estáticos	Colección de páginas estáticas	Navegación por enlaces	Documentos, o información editada y publicada en formato HTML
Estáticos con formulario de entrada	Colección de páginas estáticas con interacción básica por medio de formularios de entrada.	Comunicación en línea	Cuestionarios, Libros de Invitados, o Comentarios y Sugerencias
Estáticos con acceso a datos dinámicos	Colección de páginas estáticas con acceso a datos almacenados en bases remotas por medio de consultas	Interacción con datos presentados en formato HTML	Consulta a bases de datos del Banco Mundial
Creados dinámicamente	Colección de páginas dinámicas con contenido personalizado a las necesidades del usuario	No hay interacción	
Aplicación basada en Web	Colección de páginas dinámicas	Interacción con el usuario	Sistema de educación a distancia

Anexo 4 – Códigos de Respuesta HTTP

ID	Estado	Código HTTP	Descripción
A	Recuperaciones Exitosas	200	OK
		206	Parcial
B	Redirección a otra página	301	Movido
		302	Encontrado
		307	Re-direccionado Temporariamente
C	Fallas del lado del servidor	500	Error interno del servidor
		502	Bad Gateway
		503	No disponible
		504	No contenido
D	Requerimientos no permitidos	401	No autorizado
		403	Prohibido
		406	No aceptable

Anexo 5 – Tiempos de Respuestas según Nivel de Conectividad

Conectividad	Tiempo Máximo Recomendable
28.8 kb MODEM	12.00 seg.
512 kb cable	4.34 seg.
2 Mb ADSL	2.60 seg.

Anexo 6 – Códigos Socsibot [20]

El archivo de estructura especifica los links de una página seguidos por el URL de la página y el código resultante de la operación:

1- "página HTML valida",
2- "error al intentar obtener la página",
5- "página no-HTML" (eje: PDF, Word).

Anexo 7 – Restricciones Socsibot [21]

Restricciones: Socscibot no registra las páginas si el sitio requiere que sean excluidas por medio d el protocolo "robots.txt". Tampoco registra URLs con las siguientes extensiones comúnmente encontradas en mirrors o páginas dinámicas:

/cgi-bin/	hypermail	mirror	timetable
.cgi	javadoc	/parser.pl/	twiki
.dll	java/doc	pipermail	unixhelp
archive	/JDK1.	/record=	wwwstats
/calendar/	/JDK/	/roombooking/	webstats
/ftp/	/JDK2.	sashtml	bbs.
ftp.	/manual/	/search/	wwwboard
/handbook/	/manuals/	sessionid	

Anexo 8 –Datos Requeridos

Aspectos	Datos requeridos
Tecnoología	Las URL de las páginas Web de los portales educativos
	La extensión que indique la tecnología del sitio (php, asp, html, etc.),
	La extensión que indique el tipo de documento no-html (pdf, doc, wav, avi, etc.),
Funcionamiento	El código de respuesta HTTP del servidor a la petición de las páginas (campo Status CLF),
	El tiempo que toma la acción de petición en milisegundos (campo Time Taken ECLF).
Contenido del Sitio	El dato "meta-name [DC:identifier]" existente en el código de la página si se ha catalogado con los criterios RELPE.
	El tamaño del documento recibido medido en bytes para cada una las URL de las páginas de los portales (campo Bytes Recv ECLF). Se excluye el peso de las imágenes.
	El dato "meta-name [DC:date]" existente en el código de la página si se ha catalogado con los criterios RELPE.

[20] http://socscibot.wlv.ac.uk/
[21] http://socscibot.wlv.ac.uk/

	El dato "meta-name [RELPE:Type]" existente en el código de la página si se ha catalogado con los criterios RELPE.	
Estructura Interna	URLs de las páginas Web de los portales educativos.	
	Links que conectan las URLs.	
Uso	Host	Maquina Cliente que realiza el requerimiento al servidor (DNS name o IP)
	Ident	Identificador RFC931 (logname remoto) del usuario. (no siempre este datos está disponible)
	Authuser	User name tal como el usuario se ha autenticado a si mismo. (no siempre este datos está disponible)
	Date	Fecha de realización del requerimiento
	Time	Hora de realización del requerimiento
	Request	La línea de requerimiento exactamente como viene del cliente -(URL Destino. En CLF también contiene la operación?)
	Status	El código de estado HTTP de la transacción que retorna al cliente
	Bytes	Es el tamaño del documento transferido en bytes.
	Browser - User Agent	Tipo de Browser del cliente
	Referrer	URL desde la cual se accedió a la URL actual - The site that the user last visited. This site provided a link to the current site.
Estructura Externa	URLs de las páginas Web correspondientes al listado S.	
	Links que conectan las URLs del Listado S a las URLs de los portales.	

CLF = Formato Comun (Common Log Format) ECLF= Formato Extendido (Extended Common Log Format)

El análisis de contenido de un sitio Web presenta mayores dificultades cuando se trata de páginas dinámicas. Si es posible que un sitio genere una vista distinta dependiendo del usuario entonces la pregunta que se desprende es si las diferentes versiones deberían ser tenidas en cuenta (Cooley, 2000). A un mismo URI le pueden corresponder distinto grupo de contenidos. Para resolver esto se utiliza un "session ID" que torna cada string en un URI único.

Los datos de estructura interna de un sitio Web son aquellos que describen la organización del contenido del mismo (Poblete, 2004). En este trabajo se incluye solo el análisis de la organización entre páginas (inter-page structure) a través de los hyperlinks o enlaces que le da una organización a todo el sitio. Los objetos de estudio son las páginas mismas y los links que las unen que pueden ser de tres tipos out-links (links que salen de una página hacia otra), in-links (que llegan a la página desde otra) y co-citation links (links que conectan dos páginas a través de una tercera). Estos datos no se encuentran disponibles en una base de datos como ocurre con los datos de uso, por lo que debe definirse un proceso para obtenerlos.

VI. GLOSARIO:

Accesibilidad: El contenido debe estar accesible a todos los usuarios. La W3C formulo una serie de estándares para garantizar este punto[22] (La accesibilidad se ve garantizada cuando se utilizan herramientas de autoría[23]).

Acceso, **hit** o **impresión**: Es una petición individual de carga de un objeto (páginas, documentos, imágenes, vídeos, etc.) que recibe el servidor por parte del cliente. Así, el tráfico de un determinado sitio Web es la carga del servidor, el número de objetos servidos en un período de tiempo (accesos/tiempo) (Villena, Gonzalez et al., 2002). Un acceso es una acción única en el servidor Web tal como aparece en el archivo de registro. Un visitante que descargue un solo archivo se registra como un solo acceso en el servidor, mientras que un visitante que solicite una página Web que incluya dos imágenes se registra como tres accesos, uno por solicitar la página .html, y los otros dos por solicitar los archivos de imagen descargados. El volumen de accesos es sin duda un indicador del tráfico en el servidor Web, pero no refleja con exactitud el número de páginas que se están viendo (Web Trend).

Centralidad de los nodos (páginas): es el grado de conexión de un nodo respecto a los demás nodos en un grafo. Se calcula la suma de las distancias de cada nodo respecto los demás. Estas distancias se normalizan por la "converted distance" que es un factor normalizador de la centralidad. (El nodo central corresponde al que muestra el valor más alto de centralidad relativa y se convierte en la raíz del grafo).

Click-streams: Es la serie secuencial de Vistas de Página requeridas. De nuevo, el dato disponible del servidor no siempre provee información suficiente para reconstruir el click-stream completo del sitio. Cualquier vista de página accedida a través del cliente o el nivel Proxy del caché no estará visible desde el lado del servidor (Srivastava, Cooley et al., 2000).

Confiabilidad: Conjunto de atributos que tienen que ver con la capacidad del software de mantener su nivel de funcionamiento en condiciones indicadas durante un período indicado de tiempo (estándar ISO/IEC 9126-1).

Contenido Web: Es lo dicho a los usuarios a través del lenguaje natural, imágenes, sonidos, películas, animaciones, etc. En HTML existen elementos estructurales que especifican la forma de presentar los contenidos. Por ejemplo: el contenido de un "header" es lo que este dice, por ejemplo "Web Content Accessibility Guidelines". El "header" es un elemento estructural (W3C).

ECLF (Extended Common Log Format): Es un formato estandar de Logs de Acceso.

Eficiencia: Conjunto de atributos que tienen que ver con la relación entre el nivel de funcionamiento del software y la cantidad de recursos usados, en condiciones indicadas (estándar ISO/IEC 9126-1).

[22] *http://www.w3.org/TR/WCAG10/checkpoint-list.html*
[23] *http://www.w3.org/TR/WAI-AUTOOLS/*

Episodios: cualquier sub-conjunto de sesiones de servidor o de usuario semánticamente significativo (Srivastava, Cooley et al., 2000). *"Esta técnica trata de encontrar patrones entre transacciones tales que la presencia de un conjunto de ítems es seguido por otro ítem en el conjunto de transacciones ordenadas por estampillas de tiempo. En general, en los Server Web, la visita de un cliente es registrada por un período de tiempo. El descubrimiento de patrones secuenciales en bitácoras de accesos a un serverWeb permite a organizaciones predecir patrones de navegación de usuarios y ayudar en futuras estrategias de marketing, por ejemplo, a que grupos dirigir las diversas ofertas y promociones. Por el análisis de esta información, Web Mining puede determinar relaciones temporales entre ítems de datos" (Filocamo & Chesñevar, 2003).*

Funcionalidad*:* Conjunto de atributos que tienen que ver con la existencia de un grupo de funciones y sus propiedades especificadas. Las funciones son aquellas que satisfacen el uso indicado o implícito de los usuarios (estándar ISO/IEC 9126-1).

Lenguajes Script: Son lenguajes de programación que permiten generar páginas dinámicas puesto que poseen componentes que interactúan con bases de datos. Los lenguajes mas conocidos son: CGI Perl, PHP, ASP, JSP y Cold Fusion

Log File: es el archivo que almacena los requerimientos hechos a un servidor por sus clientes. Existen dos tipos de Log Files: Access Log y Error Log. Algunos servidores registran además: Referrer Log y Agent Log (Referrer Information) los cuales almacenan en información sobre los links entre las páginas Web y los documentos locales, las keywords usadas por los motores de búsqueda, el browser y el sistema operativo utilizados en la máquina cliente.

Mantenibilidad: Conjunto de atributos que tienen que ver con el esfuerzo realizado para hacer modificaciones especificadas (estándar ISO/IEC 9126-1).

NAT (Network Address Translation): es el mapeo de una direccion IP usda en una red por otra direccion IP de otra red. Generalmente, una organización mapea las direcciones *IP* de su red interna en unas pocas direcciones *IP* para salir a *Internet* y realiza el mapeo inverso cuando recibe información desde *Internet,* de esta manera la taxonomía de la red interna no queda expuesta.

Outlinks: links hipertextuales que apuntan a páginas externas al sitio.

Página Web: Documento de una dirección que puede contener texto, imágenes u otros elementos. Cuando la página está formada por varios marcos, el conjunto de los mismos tendrá, a efectos de cómputo, la consideración de página unitaria. Tienen la consideración de página los cgi's que realicen llamadas a páginas de hipertexto, como consecuencia de la acción de un usuario.

Página Dinámica: Son páginas en código HTML generadas a partir de lenguajes de programación (scripts) que son ejecutados en el servidor Web a diferencia del JavaScript que se ejecuta en el navegador del usuario. El código HTML puede ser modificado en función de una petición realizada por el usuario a una base de datos haciendo que los contenidos que se generen sean diferentes. Estas páginas son indexadas por Google

siempre que no tengan demasiados parámetros especificados en la URL. Si los parámetros son mayores a uno Google no lo indexa puesto que teme que el contenido no sea estable[24].

Página Estática: son aquellas páginas cuyo código de programación permite que sean recuperadas por un navegador desde un servidor remoto y no sufren modificaciones en el proceso.

Portabilidad: Conjunto de atributos que tienen que ver con la capacidad del software para ser transferido de un ambiente a otro (estándar ISO/IEC 9126-1).

Requerimiento a un servidor: es la petición realizada a un servidor por parte de un cliente cuando establece una conexión TCP con dicho servidor. Cada archivo de una página Web accedida por un cliente es considerado un requerimiento distinto.

Sesión de servidor: Se lo conoce comúnmente como ***visita***. Es la colección de los click del usuario en un servidor Web particular durante una sesión de usuario (Lavoie & Frystyk Nielsen, 1999). El conjunto de sesiones de servidor es necesariamente un input para cualquier análisis de Uso de la Web o herramienta de data mining. El final de una sesión de servidor es definido como el punto en el que la sesión del usuario de un navegador en el sitio ha terminado (Srivastava, Cooley et al., 2000).

Sesión de usuario: Es el click-stream de vistas de páginas de un usuario simple a lo largo de toda la Web. En general, solo la porción de cada sesión de usuario que esta accediendo un sitio especifico puede ser usada para el análisis, a partir de que la información de acceso no esta públicamente disponible en la mayoría de los servidores Web. El conjunto de vistas de página en una sesión de usuario para un sitio Web particular es referido como ***sesión de servidor*** – comúnmente referido como ***visita***[25] - o la colección de los click del usuario en un servidor Web particular durante una sesión de usuario (Lavoie & Frystyk Nielsen, 1999). El conjunto de sesiones de servidor es necesariamente un input para cualquier análisis de Uso de la Web o herramienta de data mining. El final de una sesión de servidor es definido como el punto en el que la sesión del usuario de un navegador en el sitio ha terminado. Nuevamente, este es un concepto simple que es muy difícil de rastrear de manera confiable. Cualquier sub-conjunto de sesiones de servidor o de usuario semánticamente significativo es referido como un "episodio" por la W3C WCA (Srivastava, Cooley et al., 2000). Se considera que una sesión ha concluido cuando se produce un período de inactividad superior a un valor (30 min. para WebTend y 10 min. para Lawerinto), y no existe una sucesión lógica de las páginas visitadas cuando se supera ese período de inactividad (Villena, Gonzalez et al., 2002).

Permite representar el Uso de un Sitio Web (usuarios/tiempo) en contraposición de la carga del servidor que se mide a través del número de accesos durante un periodo de tiempo determinado. Se considera que una sesión ha concluido cuando se produce un

[24] http://google.dirson.com/posicionamiento.net/google-páginas-dinamicas/
[25] *Otros autores definen a esto como Sesión de Usuario como ser: Villena, J., J. Gonzalez, et al. (2002), quienes definen a la Sesión de Usuario como la agrupación de todas las páginas que visita un mismo usuario. Es el conjunto de páginas consultadas por un usuario durante una sola visita al sitio Web.*

período de inactividad superior a un valor (30 min. para WebTend y 10 min. para Lawerinto), y no existe una sucesión lógica de las páginas visitadas cuando se supera ese período de inactividad. Es decir que si se ha superado este umbral y la nueva página visitada por el presuntamente mismo visitante es accedida desde la última página registrada entonces no se rompe la sesión.
Se recomienda la utilización de metodologías de huellas dinámicas para la medición de las sesiones puesto que proporciona una visión más acertada del uso de la Web. Esto se justifica por la mayor perdida de registros de accesos sufrida por el efecto caché en el caso de huellas estáticas. El efecto caché en este caso afecta al número de páginas vistas en cada sesión y por tanto a la duración de la sesión. También el efecto caché sufrido en el caso de huellas estáticas no registra adecuadamente los periodos de inactividad puesto que hace que las sesiones se rompan (Villena, Gonzalez et al., 2002). Desde el lado del servidor (es decir con los datos disponibles en el servidor) la única forma de identificar sesiones de usuario/servidor es utilizando los datos de la dirección de IP, el agente (un programa actuando en nombre de una persona u organización) y el Click stream. La otra opciones utilizar una metodología del lado del Cliente.

Usabilidad: Conjunto de atributos que tienen que ver con el esfuerzo necesario para el empleo, y la evaluación individual de tal uso, por un grupo indicado o implícito de usuarios (estándar ISO/IEC 9126-1).

Usuario único: Representa el número de usuarios no duplicados (contabilizados una sola vez) que han accedido a su sitio Web durante el transcurso del período de tiempo especificado. Se determinan mediante un proceso de autenticación de la IP por medio del uso de nombre de dominio o por las "cookies" (Google, ; WebTrends, 2006). También hay referencias a que se puede determinar mediante la dirección IP, el agente y el clic stream (Srivastava, Cooley et al., 2000). Es lo mismo que: **Número de visitantes únicos**: El recuento de los números de IP únicos para el período del informe autenticados por medio del uso de nombres de dominio o "cookies". Los visitantes únicos se cuentan utilizando la dirección IP del visitante, el nombre de dominio o el "cookie". Los "cookies" persistentes se definen en la ficha Cookies en la ventana Opciones. Los "cookies" ofrecen los resultados más exactos.
Número de visitantes únicos = Número de visitantes que visitaron una vez + Número de visitantes que visitaron más de una vez.

Usuario: es una persona individual que accede a un archivo a través de uno o más servidores utilizando un navegador de Internet (Srivastava, Cooley et al., 2000).

Vista de Página: se compone de todos los archivos que contribuye a visualizar en el navegador de un usuario en un mismo momento. Las vistas de páginas están asociadas usualmente con acción de un usuario individual (como por ejemplo un clic del mouse) y pueden estar compuestos por varios archivos de frames, gráficos y scripts. Cuando se discute y se analiza el comportamiento del usuario es importante tener en cuenta el conjunto de páginas vistas. El usuario no explicita el pedido por tal frame o por tal grafico sino que solicita una página web completa. Toda la información necesaria para determinar cuales archivos constituyen una vista de página es accesible desde el servidor Web (Srivastava, Cooley et al., 2000).