Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

INVESTIGACIÓN EDUCATIVA A PARTIR DE LA INFORMACIÓN LATENTE EN INTERNET

INVESTIGAÇÕES EDUCACIONAIS REALIZADAS A PARTIR DO CORPUS LATENTE NA INTERNET

EDUCATIONAL RESEARCH FROM INTERNET LATENT DATA

Antonio Ramón Bartolomé Pina[1]
Universitat de Barcelona (UB), Espanha

Francislê Neri de Souza[2]
Universidade de Aveiro (UA), Portugal

Marcelo Carneiro Leão[3]
Universidade Federal Rural de Pernambuco (UFRPE), Brasil

Resumo
O objetivo desse estudo é analisar o uso de informação latente contida na Internet, em investigações no campo das Ciências Sociais e da Educação. Não se trata de recolher dados "através da Internet", por exemplo, por meio de questionários. Tampouco são utilizados dados pré-existentes na Internet. São estudos que analisam documentos contidos na Internet, extraindo os dados a partir de análises desses documentos. Pode-se dizer que se trata de um modelo de investigação de "traços", "cursos" ou "restos" deixados pelos usuários da Internet. Para tanto, foram consultados artigos de pesquisas realizadas a partir desse *corpus* latente, centrando-se em quatro questões: o que se pode investigar, qual é o contexto em que se encontram estes dados, como selecionar as amostras, que considerações éticas devem ser consideradas? A análise do contexto distingue entre estudos sobre um conteúdo e estudos sobre interação. Também analisa as diferenças entre os sítios "voltados à notícia" e os "voltados ao documento". São analisadas, em continuidade, os três tipos de populações e os processos de extração das amostras, incluindo a necessidade de correções a *posteriori*. As conclusões do estudo ressaltam de modo sucinto as possibilidades e oportunidades, assim como os limites desse tipo de investigação.
Palavras-chave: Pesquisa; Corpus latente; Internet; Educação.

Resumen
Este estudio pretende analizar las posibilidades y la problemática relacionada con el uso de información latente contenida en Internet como base para investigaciones en el campo de las Ciencias Sociales y de la Educación. Este tipo de trabajos se distingue de aquellos que recogen los datos "a través de Internet", por ejemplo, mediante cuestionarios. Igualmente se distinguen de aquellos que utilizan datos recogidos en Internet. Son estudios que analizan documentos contenidos en Internet, extrayendo los datos a partir del análisis de esos documentos. Esos documentos fueron depositados en su momento en la Red ajenos al futuro uso que se haría de ellos como objeto de investigación. Se puede decir que se trata de un modelo de investigación de "trazos" o "restos dejados por los usuarios de Internet". Para ello se han revisado artículos que recogen investigaciones realizadas a partir de este corpus latente, centrándose en 4 cuestiones: Qué se puede investigar, Cuál es el contexto en el que se encuentran estos datos, Cómo seleccionar las muestras, y Qué consideraciones ética habría que tener en cuenta. El tipo de contenidos que pueden ser objeto de investigación en Internet es muy amplio, teniendo en cuenta que en ocasiones se hace necesario trabajar con información indirecta. El análisis del contexto distingue entre estudios sobre el contenido y estudios sobre la interacción. También analiza las diferencias entre los sitios "orientados a la noticia" y los "orientados al documento". Entrando ya en el problema de la selección de la muestra se procede a analizar los tres tipos de poblaciones a

[1] Doctor en Filosofía y Ciencias de la Educación. Profesor titular de la Universitat de Barcelona, Facultad de Pedagogía. abartolome@ub.edu
[2] Doutor em Educação em Ciência. Pesquisador da Universidade de Aveiro, Departamento de Educação. fns@ua.pt
[3] Doutor em Química Computacional, Docente da Universidade Federal Rural de Pernambuco, Departamento de Química. mbcleao@terra.com.br

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

considerar y los procesos de extracción de muestras aplicables en cada caso, incluyendo la necesidad de correcciones a posteriori. Las conclusiones del estudio resaltan de modo sucinto las posibilidades y oportunidades tanto como los límites de este tipo de investigaciones.
Palabras clave: Investigación, Corpus latente, Internet, Educación.

Abstract
This study aims to analyse the use of latent information in the Internet, as basis for the research in Educational and Social Sciences. Such work is distinguished from those that collect the data "through Internet", e.g. with questionnaires or forms. It also differs from those using data stored in the Internet. We work with studies that analyse documents in the Internet, extracting data from the analysis of these documents. These documents were once distributed through the Web not knowing this future use as research objects. We can say that it is a research model based on trails left by Internet users. For this work we have revised the literature with research work based on this latent corpus, addressing four questions: What can be investigated with this data? In which context this data was found? How to select samples? And what ethics aspects should be considered? We have found very different kind of contents. In fact we have distinguished between studies on content and studies on interaction. Also we have found differences between websites oriented to news and oriented to knowledge. About samples, we have considered three kinds of population: closed-finite, open-infinite and The Web. We have also studies sample correction techniques. Conclusions of this work detail possibilities and opportunities as well as the limits of this research method.
Keywords: Research; latent Corpus; Internet; Education.

INVESTIGACIÓN EDUCATIVA A PARTIR DE LA INFORMACIÓN LATENTE EN INTERNET

Introducción

La investigación educativa basada en datos observados y recogidos de la realidad se ha enfrentado tradicionalmente a dos procesos que consumen tiempo y recursos: la obtención de los datos y el análisis de esos datos. A mayor volumen de la muestra, más tiempo y recursos debían ser invertidos en esas dos etapas. Un colectivo que se veía especialmente perjudicado era el de los estudiantes que debían realizar sus prácticas investigadoras sobre datos ya existentes y en una dimensión manejable.

Desde hace años el segundo problema se ha ido solucionando gracias a la introducción de las Tecnologías de la información, tanto en el manejo de datos cuantitativos como cualitativos. Lejos quedan los años en que una parte importante del tiempo de formación del investigador se dedicaba a la aplicación manual de complejas fórmulas estadísticas. Y aunque el camino por recorrer en el análisis de la información cualitativa es aún largo, el análisis de textos o registros audiovisuales se ve facilitado mediante nuevas herramientas que en mayor o menor medida facilitan esta tarea.

El primer problema también está encontrando fórmulas de solución. Por un lado, existen grandes bases de datos sobre las que podemos trabajar. Y por otro, día a día Internet (la Web) ha ido acumulando más y más información en forma de textos, imágenes, vídeos, etc. Es lo que podríamos llamar un corpus latente de contenidos, disponibles en Internet para quien desee y posea las habilidades necesarias para extraerlo.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

Ya hace más de 10 años, Robinson (2001) señalaba que Internet se había convertido en un fórum de comunicación informal, recogiendo un retrato de las convergencias y divergencias de las personas en diversas áreas de la sociedad humana. En Internet podemos encontrar historias, blogs, redes sociales, foros, agregadores, sitios web y otras muchas herramientas abiertas a las personas comunes. En Internet están presentes datos no numéricos y datos no estructurados como textos, sonidos, vídeos, imágenes, y también existe una gran cantidad de datos numéricos que reflejan el número de accesos, localizaciones, frecuencias, valoraciones rápidas, etc. Además de los datos ya disponibles, que Robinson (2001) llama "narrativas no solicitadas" una gran cantidad de datos pueden ser producidos a través de cuestionarios, encuestas y páginas dinámicas.

Diversos investigadores (BIRNBAUM, 2004; NERI DE SOUZA; ALMEIDA, 2009) afirman que es posible obtener fácilmente una muestra de gran dimensión a partir de los datos contenidos en Internet. Esta muestra es heterogénea por lo que se refiere a la edad, educación, raza, nacionalidad y clase social. Así, a pesar de que Internet tenía una tasa de penetración global escasa, el 15,4% según Neri de Souza (2010) su rápido crecimiento garantiza la relevancia y representatividad en todas las sociedades y niveles sociales. Como dice Birnbaum (2004), Internet es un medio conveniente para la investigación internacional o intercultural.

El concepto de "*crowdforcasting*" está cada vez más difundido y tiene su base en la investigación en plataformas como Twitter o Facebook o Google+ (*crowdsourcing*). Con esta técnica es posible estudiar el comportamiento de millones de usuarios de estos medios de interacción social (BARBOSA; O'REILLY, 2011). El análisis de estos datos (*datamining*) se ha convertido en objeto de trabajo para diversas empresas de prospección de mercado y previsión de comportamiento de los consumidores.

La investigación educativa no ha entendido todavía completamente la riqueza de información y el desafío que supone este contexto. Necesitamos de mejores modelos, técnicas y una reflexión metodológica que construya una epistemología de la investigación sobre datos contenidos en Internet. Por ejemplo, no está claro bajo qué condiciones las encuestas en Internet pueden ser efectivas, qué factores pueden influir en su validez, cómo la implementación de algunas técnicas puede mejorar la ratio de respuestas y la calidad de los datos, y cómo reaccionan los entrevistados a las entrevistas en línea (ZHANG, 1999).

En este texto realizaremos una aproximación a los temas planteados, dentro de una perspectiva amplia referida al campo de las Ciencias Humanas y Sociales, aunque con referencias directas al campo educativo, y atendiendo principalmente a los problemas derivados de la obtención de la muestra.

Primera aproximación al problema metodológico

Diversos trabajos han apuntado a los problemas y desafíos metodológicos planteados (BIRNBAUM, 2004; NERI DE SOUZA; ALMEIDA, 2009; ZHANG, 1999). Según Birnbaum (2004) los dos principales problemas metodológicos en estudios con base en la red son:

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

- Alta tasa de abandono
- Repetición de la participación.

Zhang (1999) apuntaba varios problemas para la validez de la muestra: la falta de acceso paritario a Internet por todos los ciudadanos, la falta de software de algunos usuarios y la imposibilidad de que toda la población conozca que se están realizando estos estudios.

Para resolver los problemas de la recogida de datos en línea, Granello y Wheaton (2004) discutieron procedimientos detallados, incluidas estrategias para lidiar con las limitaciones de estos procesos: "*Los beneficios de precisión, bajo costo, velocidad y tamaño de los datos pierden importancia si estas limitaciones no son adecuadamente tratadas. En el caso de los datos en línea debe prestarse una especial atención a la representatividad de la muestra*" (GRANELLO; WHEATON, 2004, p. 392). Es importante darse cuenta que estos problemas y sus soluciones son tratados generalmente en base a una metodología cuantitativa. En este texto por el contrario vamos a fijarnos en datos que ya existen en Internet, y que fueron elaborados en un contexto natural, muchas veces independiente del investigador. Estos datos esconden en sí un potencial latente para un gran número de cuestiones de investigación que interesan a las ciencias humanas y sociales (NERI DE SOUZA, 2010; NERI DE SOUZA; ALMEIDA, 2009).

La idea de estudiar sobre los datos ya disponibles en Internet no es nueva (BROWNLOW; O'DELL, 2002; NERI DE SOUZA; ALMEIDA, 2009; ROBINSON, 2001), y se remonta al comienzo mismo de la popularización de Internet, aunque, aparte de las cuestiones éticas, poco se ha avanzados en los últimos años para profundizar y sistematizar una metodología con base en datos latentes. Una recién creada revista, la Internet Latent Corpus Journal (ILCJ), ha tratado de profundizar en esta temática de modo específico. En este texto se estudiaran en estas cuatro cuestiones:

- Qué se puede investigar en Internet
- Cual es el contexto de los datos en Internet
- Cómo seleccionar muestras en Internet
- Cuáles son las consideraciones éticas a considerar.

Qué se puede investigar en Internet

Como ya se ha indicado, a pesar del creciente nivel de penetración de Internet, no todos los grupos y clases sociales están representados. Por ejemplo, no podemos hacer un estudio con indios de una tribu del Amazonas con datos exclusivamente contenidos en Internet. Pero sí podremos hacerlo sobre la opinión de antropólogos que han visitado estas tribus y que exponen sus trabajos académicos en fórums en línea.

En este ejemplo no consideramos los artículos que estos antropólogos han publicado en revistas científicas, lo que sería una revisión de textos sino que trabajamos sobre los datos producidos por etnógrafos, expertos

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

ambientales y otros agentes sociales que construyen opinión sobre los indios a través de sus blogs, foros, redes sociales y otros medios informales en Internet. Es decir, datos que tienen un potencial latente de investigación en ciencias sociales y que no fueron construidos para un propósito de investigación específico.

En nuestro ejemplo se ve que aunque no tengamos una representación directa de todos los niveles sociales, podemos tener una representación indirecta.

Otro ejemplo de contenidos analizables modificando los procesos nos lo proporcionan los estudios económicos para calcular la inflación. Ésta se estima mediante llamadas telefónicas, visitas a tiendas y mercados, recogiendo los precios de productos y servicios. Después, estos datos son confirmados y analizados y a partir de ellos se calcula la inflación. Pues bien, dos investigadores del MIT, Alberto Cavallo y Rigoberto Rigobon, pusieron en marcha el proyecto Billion Prices (bpp.mit.edu). Con este software en la red es posible seguir los precios de más de cinco millones de mercancías en más de 70 países, obteniendo tasas de inflación en tiempo real (EASTER, 2011). Evidentemente este proyecto no sustituye a las estadísticas oficiales ya que hay precios de prestación de servicios, como peluquería o dentista, que no son fácilmente accesibles a través de Internet. Pero proporcionan un valor añadido que antes no existía.

De modo similar centenares de empresas prestan un servicio de estudios de mercado a través del comportamiento de sus consumidores en la Red. Empresas de venta por Internet toman nota de los patrones de conducta de sus clientes, por ejemplo Amazon, para poder ofrecerles ofertas ajustadas a sus necesidades.

Veamos ahora cómo estos ejemplos pueden trasladarse al ámbito educativo. Si deseamos conocer los hábitos lingüísticos de los jóvenes de Cataluña podríamos hacer una encuesta sobre el uso de la lengua a partir de los centros educativos. Pero también podríamos analizar la lengua utilizada en los mensajes en una red social (o también en foros, etc.) en los que estén activos la localización geográfica y la identificación de edad. Más adelante trataremos el tema de la validez y fiabilidad de estos datos. Lo que nos interesa aquí es señalar que podemos analizar no sólo qué lengua utilizan sino el nivel de vocabulario empleado, la construcción sintáctica, y otros elementos de la redacción de los mensajes. Otro ejemplo, esta vez real, al que haremos referencia más adelante: la imagen que tienen de sí mismas las adolescentes españolas a partir del análisis de las imágenes contenidas en fotologs.

El segundo ejemplo nos muestra que no estamos hablando únicamente de investigaciones de carácter cuantitativo: imágenes y vídeos resultantes de innumerables interacciones sociales en el mundo virtual nos permitirán recoger datos para nuestros estudios.

Algunos investigadores (BATTELLE, 2005; NERI DE SOUZA, 2010) coinciden en afirmar que en el interior de las bases de datos de Google encontramos un potencial campo de investigación para miles de tesis doctorales en Antropología Cultural, Psicología, Historia, Sociología, Economía o Educación entre otros.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

Contexto de los datos en Internet

En función del tipo de estudio

Podemos considerar que los estudios sobre el corpus latente en Internet pueden consistir básicamente en estudios sobre el contenido y estudios sobre la interacción.

Los estudios sobre el contenido son aquellos que buscan datos en los documentos localizados en páginas y sitios web públicos en Internet. Son ejemplos de fuentes de datos los repositorios de documentos textuales, vídeos o música, periódicos, sitios web institucionales, páginas home personales, blogs, wikis, etc.

Los estudios sobre la interacción son aquellos que recogen los datos a partir de las interacciones de los usuarios a través de esos sitios web. Son ejemplos de fuentes de datos los foros, los mensajes en las redes sociales, los comentarios en blogs y en servicios de noticias, los mensajes de correo y las listas de distribución, etc.

Estos estudios sobre la interacción tienen una larga historia en Internet. En el ámbito educativo son numerosas las investigaciones sobre la actividad en los foros (SILVERMAN, 1995; ALAVI; LEIDNER, 2001; BELDARRAIN, 2006). Un objeto característico de estudio dentro de la Educación y el uso de TICs ha sido el trabajo colaborativo en cursos y programas formativos. Actualmente ha cobrado importancia el análisis de redes (HILTZ, 2005), habiéndose convertido en un campo prometedor y sugerente cara a la evaluación de los procesos de aprendizaje, desde una perspectiva Conectivista. (SIEMENS, 2008). Los análisis de la interacción en las redes también se están utilizando como herramienta de evaluación, como ya se hizo en su momento con los foros.

En este trabajo no consideramos los estudios sobre la interacción, centrándonos en los estudios sobre el contenido.

En sentido estricto también tendríamos que considerar la búsqueda de información a partir de bases de datos en la red. Pero este tema no es el objeto de nuestro estudio. Las bases de datos presentan características propias por cuanto son repositorios cerrados y estructurados, es decir, de población conocida y con sistemas de selección de datos propios y definidos.

En función del carácter del sitio

Dentro de los sitios y páginas web con contenido a analizar podemos realizar otra distinción importante entre las Noticias (*News oriented*) y los Documentos (*Content oriented*).

En el primer caso encontramos todas los sitios de diarios, sitios institucionales, sitios con noticias en la primera página, y, notablemente, todos los blogs. En el segundo caso todos los documentos, archivos de documentos, y en particular los sitios wiki.

Es importante notar cómo se altera el resultado en el tiempo en función de si seleccionamos “sitios web” o “páginas “. Los sitios web (por ejemplo, un blog) se compone de páginas (en este caso “entradas o noticias”). Lo mismo

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

pasa con un diario, con un documento de texto, con los apuntes de un curso, etc. Las páginas en la Web, a diferencia de lo que sucede con las páginas en papel, no son meras divisiones mecánicas por razones técnicas, sino que suponen la organización de contenidos diferentes. Una única página web puede equivaler a numerosas páginas en papel o incluso a un documento completo, por ejemplo, un artículo científico. Este artículo sería una única "página web" dentro de un "sitio web" que podría ser una revista o una biblioteca.

Veamos que sucede si nuestra selección se refiere a "sitios web". Si seleccionamos la página principal de un sitio web de actualidad (News oriented), el contenido de la página cambiará de un día para otro: recogerá otra noticia, otro hecho u otro comentario. En el caso de un blog puede ser que no tenga nada que ver con el que se había seleccionado la semana anterior.

Si escogemos la página principal de un documento (Content oriented), ésta siempre hará referencia al mismo tema y contendrá básicamente el mismo contenido.

La situación se invierte, curiosamente, si la selección de datos se realiza en base a "páginas web". En un sitio web de actualidad, cada página corresponde a una noticia, y esta no cambia en el tiempo (aunque ocasionalmente se hagan correcciones para subsanar errores de bulto). En una noticia o una entrada de un blog, los únicos contenidos que cambian (por adición) son los comentarios de los lectores, los cuales son trabajados en los "estudios de la interacción" ya señalados. Es decir, la página correspondiente a una noticia determinada no cambia en el tiempo.

En cambio en muchos sitios web que no recogen noticias (content oriented) los documentos que corresponden a las "páginas" del sitio están en continuo proceso de actualización y cambio. Por supuesto, en el caso de revistas científicas, repositorios tipo bibliotecas, etc. estos cambios no se producen o están limitados, aunque tampoco es una norma general (recordar el caso del artículo de Nature del año 2005 sobre la Wikipedia).

No necesitamos decir que estas consideraciones afectan directamente al "fiabilidad" de los datos. Así, las primeras decisiones que debe tomar el investigador se refieren a:

- ¿Voy a hacer un estudio sobre el contenido o sobre la interacción?
- En el primer caso, ¿estoy ante sitios informativos o sitios documentales?
- Y ¿mi muestra se refiere a "sitios web" o a "páginas web"?

Poblaciones objeto de estudio

Repositorios cerrados/finitos

Los repositorios cerrados, finitos, son colecciones de documentos que poseen un número determinado de objetos, aunque éste pueda ser muy grande. Un ejemplo sería el repositorio de recursos de aprendizaje MERLOT. Casi todos los repositorios de Objetos de aprendizaje, recursos docentes, etc. asociados a instituciones educativas poseen este carácter.

Un ejemplo característico son los diferentes estudios realizados sobre contenidos de un sitio determinado para establecer su validez y fiabilidad.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

El año 2005 Giles (2005) realizó un estudio comparativo entre los contenidos de la Wikipedia y de la Enciclopedia Británica encontrando niveles similares de error. Greenemeier (2007) comparó dentro de la misma Wikipedia las aportaciones anónimas con las realizadas por usuarios registrados sin encontrar diferencias significativas. Otro modelo es aquel que contrasta la validez de un sitio a partir del juicio de expertos sobre los contenidos del mismo. También aquí la Wikipedia es un objeto de estudio continuo en diferentes áreas como la Medicina al haberse convertido en la principal fuente de información en la red sobre salud (LEITHNER et al., 2010).

No debe extrañar que este sea también un procedimiento habitual para evaluar la calidad de los sitios educativos y de los cursos en línea. Aunque existe una larga tradición de valoración de la calidad de estos cursos a través de su efectividad y no de los contenidos (STROTHER, 2002), los estudios sobre la calidad de contenidos en los cursos son muchas veces publicados en informes internos.

El carácter finito del repositorio no implica que el número de documentos contenido sea pequeño ya que pueden ser miles o cientos de miles. Tampoco implica que el número de documentos sea fijo o estable: continuamente se van añadiendo nuevos documentos, y, en muchos casos, el paso del tiempo lleva a eliminar algunos de los documentos por razones de obsolescencia, inadecuación, contenido inapropiado, y otras. Como consecuencia tampoco permite suponer que en un momento dado podamos necesariamente decir con precisión el tamaño exacto de la muestra. Sin embargo, el aspecto más interesante de estos repositorios, desde el punto de vista que nos ocupa, es que, en el momento de extraer la muestra utilizando mecanismos propios del repositorio (e.g. “buscar materiales de evaluación sobre ciencias para el nivel de secundaria”), ésta se obtiene a partir del total de elementos contenidos en el repositorio.

Un ejemplo de este tipo de repositorios lo encontramos en el trabajo de Moedas et al. (2010) en el que analizan los confesionarios en línea en Internet.

Repositorios abiertos/infinitos

Los repositorios abiertos, infinitos parecen en sí mismos una contradicción: todo repositorio de documentos contiene un número limitado de objetos en un momento dado, sea este número conocido o no. Quizás sería más adecuado hablar de repositorios de fronteras no precisas, o limites que cambian muy rápidamente. Pensemos en el caso de Youtube. El procedimiento de recogida de documentos en Youtube (como en Flickr, etc.) hace que el número de elementos contenidos esté cambiando continuamente, y su número es tan grande que cada selección de una muestra generará un resultado diferente.

Quizás una diferencia conceptual importante sería considerar que las muestras de los repositorios cerrados son en realidad, subpoblaciones definidas, sobre las que a su vez podemos establecer procedimientos aleatorios de selección, en tanto que las muestras de los repositorios abiertos son siempre muestras o subconjuntos de la subpoblación que habríamos definido. La tabla del siguiente ejemplo puede aclararlo. En el primer caso la extracción repetida en el

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

tiempo daría razonablemente conjuntos similares. En el segundo caso, cada nueva extracción prácticamente nos proporcionaría una nueva muestra.

Cuadro 1: Comparación del proceso en repositorios finitos e infinitos

Repositorio	Merlot	Youtube
Proceso de extracción de datos	En el sistema de búsqueda introducimos la expresión "Teorema de Pitágoras"	En el sistema de búsqueda introducimos la expresión "Teorema de Pitágoras"
Resultado obtenido	*Todos* los recursos de aprendizaje sobre el Teorema de Pitágoras	Una serie de vídeos asociados a esos términos.
Resultado una semana después.	Prácticamente el mismo.	Alta probabilidad de encontrar elementos diferentes
Selección de la muestra	Puede escogerse una muestra representativa – escogida aleatoriamente- a partir del resultado obtenido.	Podemos trabajar con el resultado o escoger una muestra aleatoria.
Valor inferencial	Esta muestra permite inferir resultados para el total de elementos del repositorio referidos al Teorema de Pitágoras.	Esta muestra no posee en sentido estricto el carácter inferencial al producirse cambios rápidos en la población, con lo que deja de ser representativa (en el sentido estadístico del término).

Estos repositorios proporcionan herramientas que facilitan la extracción de muestras, tanto aleatorias como arbitrarias. Es el caso del estudio sobre los vídeos promocionales de la Universidades en YouTube de Silva et al. (2010). Frigola (en preparación) construye una propuesta de géneros audiovisuales en Internet a partir de muestras de vídeos en repositorios específicos. Tortajada et al. (2012) analizan la imagen que los adolescentes ofrecen de sí mismos también repositorios específicos de fotologs.

La Web

La tercera opción para la obtención de los datos es realizar una selección a partir del conjunto de la Web. El carácter no estructurado, no controlado, distribuido y heterogéneo de la Web hace imposible escoger datos a partir de toda la Web. Es decir, siempre escogeremos muestras a partir de un subconjunto de datos, es decir, de una subpoblación. El caso más notorio es la división idiomática: generalmente nuestras búsquedas, se limitarán a uno o varios idiomas. Así pues, deberíamos hablar de "la Web de habla inglesa", etc.

Para la selección de los datos requerimos los servicios de un indexador, por ejemplo, un buscador como Google. Esta es otra fuente de reducción poblacional, ya que automáticamente nos estamos limitando a las páginas sobre las que trabaja el buscador.

Nuestro estudio raramente pretenderá indagar sobre todos los documentos contenidos en la Web. En general buscaremos documentos con algunas características como por ejemplo, "páginas web que tratan el tema de la violencia doméstica". De nuevo aquí, el proceso pasa por generar o definir una subpoblación de la que luego extraer la muestra a estudiar.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

Más adelante vamos a discutir las soluciones técnicas para la selección de muestras en el conjunto de la Web. En esas soluciones encontraremos repetido este aspecto.

Como resultado, el cuadro 2 muestra las opciones a las que nos enfrentamos no importa si nuestra fuente de datos es un repositorio abierto o cerrado o la propia web, pero con mecanismos diferentes en cada caso. La tabla expresa la viabilidad del procedimiento.

Cuadro 2: Opciones según tipo de repositorio

Procedimiento de selección de los datos	**REPOSITORIOS**		
	Cerrados	**Abiertos**	**La Web**
Se analizan todos los elementos de la población	Sí	No	No
Se extrae una muestra aleatoria de toda la población. Se analiza la muestra.	Sí	Sí	Sí
Se genera un subconjunto de la población con características definidas. Se analiza el subconjunto.	Sí	Sí	No
Se extrae una muestra aleatoria del subconjunto generado. Se analiza la muestra	Sí	Sí	Sí

Selección de la muestra

Definición de subconjuntos poblacionales

Hemos visto que los estudios sobre el corpus latente de datos en la Web comienzan por, una vez establecida la fuente de datos, definir un subconjunto poblacional acorde a nuestros intereses investigadores así como a nuestras posibilidades prácticas. Los repositorios cerrados no suelen presentar excesiva dificultad al estar perfectamente indexados todos los elementos de la población. No es el mismo caso cuando nos enfrentamos a repositorios abiertos o a la Web en su conjunto.

Para establecer el subconjunto poblacional en estos dos casos recurriremos a sistemas de indexación sobre los que aplicaremos mecanismos de búsqueda mediante cadenas de caracteres que definan nuestra población de estudio. Pero esa población puede quedar definida por elementos claramente reconocidos o no. El sistema de indexación utiliza para la búsqueda la información almacenada en el propio documento (título, texto contenido…) tanto como en los metadatos asociados. Algunas búsquedas se ajustarán con bastante exactitud a nuestra intención, por ejemplo, si buscamos en Youtube vídeos que expliquen algún aspecto relacionado con el Teorema de Pitágoras o que nos muestren el uso de una Pizarra digital, bastará introducir una de esas expresiones para obtener un subconjunto razonablemente ajustado a nuestros intereses.

Pero supongamos ahora que deseamos buscar en Youtube vídeos con intencionalidad educativa o vídeos hechos por profesores/as. ¿Cómo identificar la intencionalidad con la que se realizó un vídeo? ¿Cómo identificar que el que realizó (¿o “subió”?) un vídeo era un profesional de la enseñanza? En ocasiones el dato está indexado pero no para todos los elementos (por ejemplo, el dato sexo del autor puede estar parcialmente indexado). En otras simplemente no

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

se puede acceder directamente a esa información (se sabe algo sobre quién "subió" el vídeo a Youtube, pero no sobre quién lo "produjo").

En estos casos se suele realizar una "reducción" poblacional o seleccionar una "muestra previa" mediante procedimientos automáticos, para luego pasar a escoger los elementos que identificamos que cumplen nuestros requerimientos mediante una intervención humana. A continuación se explican en detalle estos procedimientos.

Extracción de muestras aleatorias a partir de la Web

El uso de muestras aleatorias o representativas nos permitirá posteriormente aplicar técnicas de análisis procedentes de la estadística inferencial, con las ventajas tradicionalmente atribuidas a estas técnicas, como por ejemplo, el control de variables extrañas, el control de la subjetividad en la selección de la muestra, etc. Naturalmente veremos otros estudios en los que nos interesa más la relevancia de los elementos analizados, para lo que será preferible una muestra arbitraria pero con significado.

La extracción de una muestra aleatoria en un proceso tradicional y estricto implica la asignación de un identificador numérico a cada elemento de la población. A continuación se generan números al azar de entre el conjunto de números asignados y esos elementos pasan a formar la muestra. Hace años esto se realizaba mediante tablas de números aleatorios u otros procedimientos manuales, en tanto que hoy es una tarea que realizan los ordenadores que, directamente, pueden extraer muestras aleatorias de repositorios como directorios telefónicos, etc.

Los procedimientos aleatorios para la extracción de muestras requieren repositorios cerrados. Es posible que en ciertos casos sea posible aplicarlos en repositorios abiertos, aunque no hemos identificado ninguno. Pero resulta de todo punto imposible utilizarlos cuando trabajamos a partir de la Web. Y este es el punto que vamos a analizar a continuación.

Podríamos pensar que, puesto que en el protocolo IP v.4 de Internet, todo ordenador tiene asignada una dirección compuesta por cuatro números, cada uno en el intervalo de 0 a 255, bastaría generar una muestra aleatoria de direcciones numéricas de este tipo para obtener una muestra de páginas o sitios web. Sin embargo este procedimiento no funciona.

La mayoría de esos números corresponderán a ordenadores que no realizan tareas de servidor web y que no contendrán páginas web. Otras direcciones serán compartidas por grupos de ordenadores con subdirecciones IP locales asignadas por ejemplo por un servidor DHCP. Finalmente, dentro de una misma IP podemos encontrarnos numerosos sitios o páginas web diferentes y sin ninguna relación entre sí, bien porque se encuentran en directorios diferentes (indicados en el camino en la URL), bien porque una misma IP gestiona diferentes nombres de Internet asignados a diferentes sitios ubicados en equipos o carpetas diferentes. Esta es una rápida explicación de carácter técnico pero lo que debe hacer comprender es que no existe una relación biunívoca entre esas direcciones numéricas y los sitios y páginas web. Finalmente incluso si un servidor web con una dirección IP fija contiene páginas, es posible que la simple indicación de su IP

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

no proporcionara ningún resultado o lo diera en forma de página de "error" o de "acceso prohibido".

Evidentemente, el conjunto de páginas Web no están indexadas en ningún sitio. O mejor dicho, casi en ningún sitio. Los buscadores genéricos (Google, Yahoo…) llevan años buscando páginas de todo tipo (informativas, foros, documentos…) en la web, indexándolas y facilitando su localización. Por tanto estos buscadores constituyen una buena fuente a la que recurrir. ¿Cómo realizar una selección aleatoria en Google?

Una opción podría consistir en introducir combinaciones aleatorias de caracteres en el buscador de Google (e.g. "axigh", "wkess", etc.). A partir de los resultados de la búsqueda se deben seleccionar las páginas que conformarán la muestra. Sin embargo este método introduce un elemento no controlado: el conjunto de caracteres que se obtienen pueden ser significativos en un idioma pero no en otro. Además proporcionaría resultados con contenidos en lenguas quizás desconocidas para el investigador, lo que dificultaría el análisis.

El método escogido por Bartolomé y Willem (2008), y adoptado posteriormente por otros investigadores, fue construir una lista de 50.000 palabras de uso más corriente en un idioma a partir de un diccionario de la lengua. Esto reduce la población (con matices) a un único ámbito lingüístico. En el estudio se trabajó con dos vocabularios: uno español y otro inglés.

Algunas características de este vocabulario es la eliminación de artículos, preposiciones, conjunciones, etc.

A partir de ese léxico, el ordenador puede escoger muestras aleatorias de por ejemplo 10, 20, 30 términos… del diccionario, e introducir sucesivamente esos términos en Google. De las páginas que ofrece Google, se escogen por ejemplo los 20 primeros resultados. Así, si escogiéramos una muestra de 100 palabras y obtuviéramos 20 resultados por búsqueda obtendríamos una muestra de 2000 páginas.

A la restricción antes indicada de ámbito lingüístico, debemos añadir algunas reflexiones que pueden obligar a eliminar, manual o automáticamente, algunos resultados.

Correcciones a la muestra obtenida

En primer lugar, una palabra puede existir en más de un idioma, lo que introduciría páginas procedentes de otro ámbito lingüístico y por tanto de otra población. Además pueden tener significados muy diferentes, por ejemplo, "Sex" significa "sexo" en Inglés, pero es el número "seis" en Sueco.

Un importante problema: el mismo sitio puede aparecer varias veces en la muestra a través de diferentes páginas lo que lleva a la discusión ya presentada si la población la componen los sitios Web o las páginas web. En el primer caso será necesario reducir todas las páginas procedentes de un mismo sitio (misma dirección de servidor) a la primera aparición. Pero también es posible que diferentes sitios utilicen los servicios de un mismo proveedor de Internet, diferenciándose las direcciones únicamente en los directorios que marcan el camino.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

Un aspecto a tener en cuenta, en particular si se automatiza el proceso, es si se deben considerar o eliminar los resultados “esponsorizados” (es decir, los incluidos como publicidad). Estos resultados aparecen claramente diferenciados visualmente en Google, pero un ordenador podría no identificarlos si no se programa adecuadamente.

La búsqueda en Google presenta una importante fuente de error en el carácter aleatorio de la selección. Google presenta los resultados de acuerdo con la información que posee sobre el que realiza la búsqueda. Esa “adaptación” la realiza en base al nombre del usuario de Google que está registrado como activo en el ordenador. Incluso si desactivamos las “cookies” o cerramos el usuario Google, la búsqueda se realiza condicionada a la localización geográfica del equipo, la que deduce de la IP. Si también escondiéramos la dirección IP, seguramente la búsqueda respondería a una búsqueda en Estados Unidos. Según el objeto de nuestra investigación esto no representará un problema, pero si lo fuera, habrá que considerar la posibilidad de realizar búsquedas combinadas en diferentes países (por ejemplo de Latinoamérica).

Aunque ya se ha comentado no podemos dejar de señalar que las búsquedas en Google raramente responde a “toda la web”. Un parte muy importante de la Web es inaccesible para nosotros (escrita en Chino, Ruso, Árabe, entre otras lenguas). Otra parte de la web se encuentra en servidores de acceso restringido. Esto significa que en ocasiones Google no puede ni siquiera indexar esa información mientras que en otras, aunque la haya indexado (y de hecho disponga una copia en memoria caché) nosotros no podemos ya entrar y analizarla.

Consideraciones éticas

Las cuestiones éticas relacionadas con la recogida y utilización de datos obtenidos en Internet son objeto hoy de gran preocupación (BASSETT; O’RIORDAN, 2002). Robinson (2001) propone un modelo que describe el proceso para decidir cuando los datos son directamente publicables o necesitan de una autorización.

Además de la autorización de las personas afectadas por los datos, debemos considerar la posible autorización de quien colocó esos datos en Internet. De alguna manera nos estamos beneficiando de su trabajo y esto puede ser resultado de una actitud generosa y activa del “productor” de los datos o simplemente puede ser fruto de nuestra habilidad para encontrar los datos.

A pesar de la escasez de trabajos específicamente sobre la ética en el uso de datos del cuerpo latente en Internet, podemos encontrar algunas propuestas como las de Nosek, Banaji, y Greenwald (2002). Estos investigadores señalan las posibles consecuencias de ausencia de un investigador durante la creación de los datos, la publicidad de datos confidenciales o referidos a la identidad de los autores, la seguridad del almacenamiento, transmisión o monitorización de esos datos. De estos y otros autores hemos extraído estas consideraciones:

- Citación clara de las fuentes

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

- Garantía de anonimato (considerando las posibles contradicciones con el criterio anterior)
- Respeto a la privacidad.
- Respeto a los valores humanos y morales más ampliamente aceptados.

Conclusiones

Como conclusión de este estudio podemos señalar las enormes potencialidades para la investigación sobre temas educativos y sociales que ofrece los contenidos latentes, tanto por lo que se refiere a contenidos de difícil acceso como a amplitud de muestras.

También las posibilidades para la formación de investigadores noveles que ven la posibilidad de realizar estudios sobre muestras difíciles de generar a partir de observaciones directas por sus limitaciones de medios.

Por otro lado hemos visto las precauciones que imponen esta metodología en relación a la validez y a la fiabilidad de la muestra.

Y hemos hecho mención a una todavía incipiente reflexión sobre los aspectos éticos, en muchos casos condicionados por los cambios sociales generados desde Internet en relación a aspectos como privacidad, transparencia y ubiquidad.

Quizás un aspecto que merece recogerse aquí es la posibilidad de generar procesos híbridos (datos latentes en Internet y datos recogidos en muestras observadas).

Referencias

ALAVI, M.; LEIDNER, D. E. Research Commentary: Technology-mediated Learning - A call for greater depth and breadth of research. **Information Systems Research,** 12 (1), 1-10. 2001.

BARBOSA, P.; O'REILLY, A. S. **Harvard Trends:** Tendências de Gestão (1ª ed.). Porto: Vida Económica, 2011.

BARTOLOMÉ, A.; WILLEM, C. **Integración y desarrollos de nuevos elementos de la sintaxis audiovisual en los clips de vídeo digital distribuidos por Internet.** Paper presented at the Congreso Investigar la comunicación, Santiago, España, 2008.

BASSETT, E. H.; O'RIORDAN, K. Ethics of Internet research: Contesting the human subjects research model. **Ethics and Information Technology,** 4, 233-247, 2002.

BATTELLE, J. **The Search:** Como o Google Mudou as Regras do Negócio e Revolucionou a Cultura (1ª ed.). Lisboa: Casa das Letras, 2005.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

BELDARRAIN, Y. Distance Education Trends: Integrating new technologies to foster student interaction and collaboration. **Distance Education**, 27 (2), 139-153, 2006.

BIRNBAUM, M. H. Human Research and Data Collection Via the Internet. **Annu. Rev. Psychol.,** 55, 803–832, 2004.
doi: 10.1146/annurev.psych.55.090902.141601

BROWNLOW, C.; O'DELL, L. Ethical Issues for Qualitative Research in Online Communities. Disability and Society. **Disability and Society,** 17 (6), 685–694, 2002.

EASTER, M. The Prices Are Right: Economists Find a Faster, Cheaper Way to Measure Inflation. **Scientific American**, 305 (4), 13, 2011.

GILES, J. Internet Encyclopaedias Go Head to Head. **Nature**, 438 (7070), 900-901, 2005. doi: 10.1038/438900a

GRANELLO, D. H.; WHEATON, J. E. Online Data Collection: Strategies for Research. **Journal of Counseling y Development,** 82, 387-393, 2004.

GREENEMEIER, L. Wikipedia "Good Samaritans" are on the Money. Scientific American. **Scientific American,** 19 Octubre, 2007 http://www.sciam.com/article.cfm?id=good-samaritans-are-on-the-money, 2007.

HILTZ, S. R.; GOLDMAN, R. **Learning Together Online:** Research on Asynchronous Learning Networks. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2005.

LEITHNER, A.; MAURER-ERTL, W.; GLEHR, M.; FRIESENBICHLER, J.; LEITHNER, K.; WINDHAGER, R. Wikipedia and osteosarcoma: a trustworthy patients' information? **Journal of the American Medical Informatics Association***,* 17 (4), 373–374, 2010.

MOEDAS, A.; GEITOSO, A.; GRAÇA, D. Liberdade na Internet: Que utilização é feita dos confessionários on-line? **Internet Latent Corpus Journal,** 1 (1), 19-33, 2010.

NERI DE SOUZA, F. Internet: Florestas de Dados ainda por Explorar. **Internet Latent Corpus Journal,** 1 (1), 2-4, 2010.

NERI DE SOUZA, F.; ALMEIDA, P. **Investigação em Educação em Ciência baseada em dados provenientes da internet.** Paper presented at the XIII Encontro Nacional de Educação em Ciências., Castelo Branco, 2009.

NOSEK, B. A.; BANAJI, M. R.; GREENWALD, A. G. E-Research: Ethics, Security, Design, and Control in Psychological Research on the Internet. **Journal of Social Issues,** 58 (1), 161-176, 2002.

Revista Eletrônica de Educação, v. 7, n. 2, 2013, p. 301-316. *Artigos*. ISSN 1982-7199.
Programa de Pós-Graduação em Educação, Universidade Federal de São Carlos, Brasil.

ROBINSON, K. M. Unsolicited Narratives from the Internet: A Rich Source of Qualitative Data. **Qualitative Health Research,** 11 (5), 706-714, 2001. doi: 10.1177/104973201129119398

SIEMENS, G. **Learning and knowing in networks:** Changing roles for educators and designers. Paper presented at the ITFORUM for Discussion, University of Georgia. (2008, 27 January).

SILVA, I.; MARTINS, S.; OLIVEIRA, T. Vídeos promocionais das Universidades no YouTube. **Internet Latent Corpus Journal,** 1 (1), 34-46, 2010.

SILVERMAN, B. G. Computer Supported Collaborative Learning (CSCL). **Computers & Education**, 25 (3), 81-91, 1995)

STROTHER, J. An Assessment of the Effectiveness of e-learning in Corporate Training Programs. **International Review of Research in Open and Distance Learning**, 3 (1), 1-16, 2002.

TORTAJADA, I.; WILLEM, C.; CRESCENZI, L.; ARAÜNA, N.; TELLADO, I. **Fotologs and Love Socialisation processes. A conventional or a transformative model of sexuality and relationships?.eYouth : Balancing between opportunities and risks**.pp. 215 - 232.(Bélgica): Peter Lang, 2012.

ZHANG, Y. Using the Internet for Survey Research: A Case Study. **Journal of the American Society for Information Science***,* 51 (1), 57-68, 1999.

Enviado em: 31/01/2013
Aceito em: 03/11/2013