

Organización de Estados Iberoamericanos

Centro de Altos Estudios Universitarios

Observatorio Iberoamericano de la Ciencia, la Tecnología y la  
Sociedad

Desarrollo de un portal de visualización de las temáticas de  
investigación de las universidades iberoamericanas a  
través de mapas conceptuales

**Informe N° 1:**  
**Criterios para la selección de documentos y definición del  
primer conjunto de universidades**



## 1. Las Web universitarias como fuente de información

La difusión de Internet ha hecho que, por su accesibilidad temporal y espacial, se convierta en una de las primeras fuentes ante cualquier búsqueda de información cotidiana. Por ese motivo, Internet es una vidriera obligada para las instituciones de todos los ámbitos, incluyendo el espacio universitario.

En los sitios Web de las instituciones de este sector se encuentra información institucional, de su oferta académica (carreras, cursos de posgrado, etc) y noticias de diferente índole. Sin embargo, las Web universitarias se han convertido también en un espacio privilegiado de intercambio de información en el ámbito académico. Más allá del avance en la edición digital de revistas científicas, por las ventajas de edición y llegada al público del medio digital, las universidades están ocupando un lugar cada vez importante como difusores directos de los resultados de investigación a través de sus sitios web.

La difusión de los resultados de la I+D por parte de las propias universidades está en gran medida asociado al impulso del movimiento *open access* y como reacción a la paradoja editorial por la cual las universidades, principales instituciones productoras de conocimiento en Iberoamérica, dedican un presupuesto significativo en adquirir las revistas científicas que contienen los resultados de la I+D ejecutada y financiada por ellas mismas. Esto toma la forma, no sólo de revistas de acceso abierto, sino también por la puesta a disposición de trabajos generados por los propios docentes e investigadores, acompañando su ficha personal y CV.

Por otra parte, los sitios web universitarios contienen información utilizada en el proceso de la actividad docente, especialmente bibliografía en formato digital, que ofrece un panorama de los enfoques y temáticas centrales que adoptan las distintas cátedras. Asimismo, es cada vez más extendida la creación de repositorios de documentos como *pre-prints* de artículos científicos, tesis y documentos de trabajo, de acceso libre.<sup>1</sup>

Todo este conjunto de documentos, que en algunas universidades supera las decenas de miles, ofrecen un conjunto de datos representativo de los temas abordados por cada universidad. Distintos trabajos<sup>2</sup> han abordado este tema, demostrado la presencia cada vez mayor de archivos ricos en contenido (doc, pdf, xls) en las web universitarias. Incluso se han desarrollado metodologías para la cuantificación de la presencia de los resultados de la investigación universitaria en la web, cuyos resultados son en gran medida consistentes con los indicadores bibliométricos tradicionales aplicados a nivel institucional y que son de acceso público y actualización regular.<sup>3</sup>

---

<sup>1</sup> HERNANDEZ PEREZ, Tony; RODRIGUEZ MATEOS, David y BUENO DE LA FUENTE, Gema. Open Access: El papel de las bibliotecas en los repositorios institucionales de acceso abierto. Anales de Documentación. 2007.

<sup>2</sup> AGUILLO, Isidro F; GRANADINO, Begoña y LLAMAS, Germán. Posicionamiento en el web del sector académico iberoamericano. INCI, 2005

<sup>3</sup> <http://www.webometrics.info>

## **2. La obtención de información cualitativa. Una aproximación a los contenidos.**

El objetivo de este proyecto es centrarse en los contenidos de los documentos que las universidades publican en sus portales web, brindando una forma de visualización de los conceptos principales que son tratados en los textos. De esta manera, se trata de un enfoque cualitativo, distinto y complementario a los abordajes cuantitativos existentes y ya mencionados, que más allá de su indiscutida utilidad para el monitoreo y análisis de la presencia de las universidades en la web, no ofrecen información sobre el contenido de los documentos que identifican.

Para ello, se dispone actualmente de un importante desarrollo en técnicas de análisis de lenguaje natural, a partir de técnicas de minado de textos. La aplicación de estas técnicas permite la extracción automatizada de conceptos relevantes a partir de textos sin necesidad de procesamientos previos, reconociendo estructuras sintácticas y construcciones semánticas. Los conceptos obtenidos del análisis de un volumen significativo de textos son representativos de los temas abordados en ellos y la aparición conjunta de conceptos en un mismo documento ofrece pautas de las relaciones entre ellos. De esta forma, de la visión conjunta de los documentos emerge un “mapa” que no puede apreciarse en la lectura aislada de los textos que componen el corpus a estudiar. Estas técnicas informáticas además permiten procesar grandes volúmenes de información, inabarcables sin su aplicación.

Para la representación gráfica de la información obtenida mediante técnicas de minado de textos en mapas conceptuales, es posible recurrir a diferentes técnicas que permiten disponer gráficamente en un plano los conceptos obtenidos y sus relaciones. Estas técnicas, basadas en teoría de grafos y análisis de redes permiten al observador obtener un panorama, de fácil e intuitiva interpretación, del contenido de grandes dominios de información. Algunos ejemplos de visualización de grandes volúmenes de información son los proyectos Atlas de la Ciencia (<http://www.atlasofscience.net>) y la plataforma Stanalyst, del INIST-CNRS . Por otra parte, las posibilidades gráficas de la tecnología web ofrecen características de gran interactividad y flexibilidad para los usuarios.

La integración de estas fuentes y tecnologías, cuya validez y funcionalidad está comprobada en sus ámbitos respectivos, puede ofrecer una valiosa herramienta para los distintos agentes del sistema científico y tecnológico iberoamericano (investigadores, gestores, empresarios, etc), mostrando de manera eficaz y dinámica la oferta de capacidades de cada universidad, ofreciendo posibilidades de colaboración y vinculación.

El producto final de este trabajo será un portal que contenga los mapas conceptuales de un conjunto de universidades, junto con la metodología para su expansión y actualización. Esos mapas serán navegables y permitirán la búsqueda de términos específicos, e incluso el acceso a los documentos que han generado las relaciones que resulten de interés a los usuarios.

Los siguientes gráficos presentan, a modo de ejemplo estático, una imagen general y dos recortes de un mapa elaborado como parte del proceso de desarrollo informático de las herramientas necesarias para alcanzar los objetivos de este proyecto. Se han utilizado para ello tan sólo seiscientos documentos en idioma español, extraídos de la página web de la Universidad Complutense de Madrid.

El tamaño de las esferas y cuerpo tipográfico da cuenta de la frecuencia de los conceptos, mientras que los colores representan agrupaciones temáticas conseguidas mediante técnicas de *clustering*. Los lazos entre las palabras están dados por su aparición conjunta en un mismo documento, aunque para facilitar la visualización, se han podado los enlaces hasta dejar tan sólo los más relevantes para la estructura general de la red.

Gráfico 1. Visión general de un mapa conceptual a partir de documentos de la Universidad Complutense de Madrid

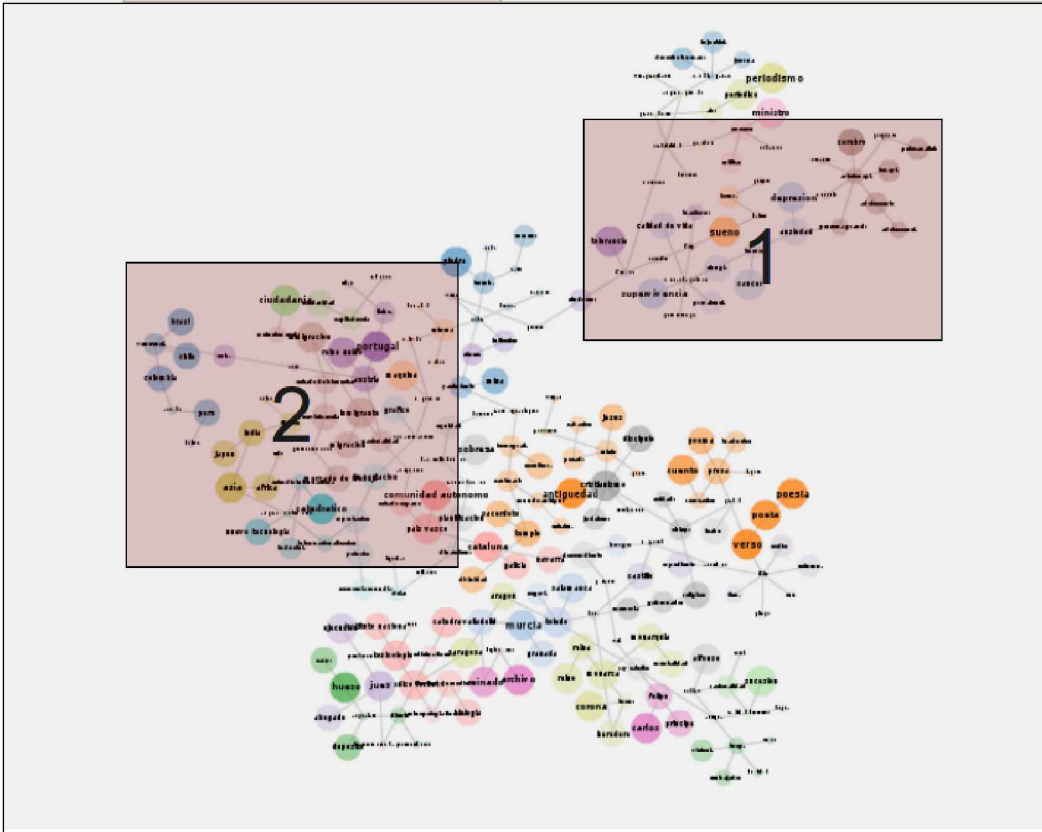
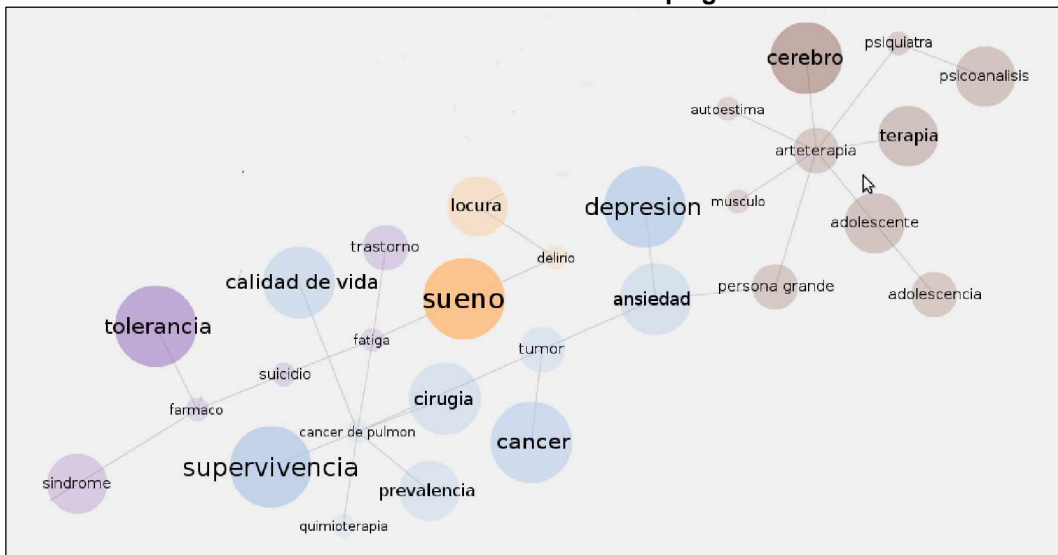
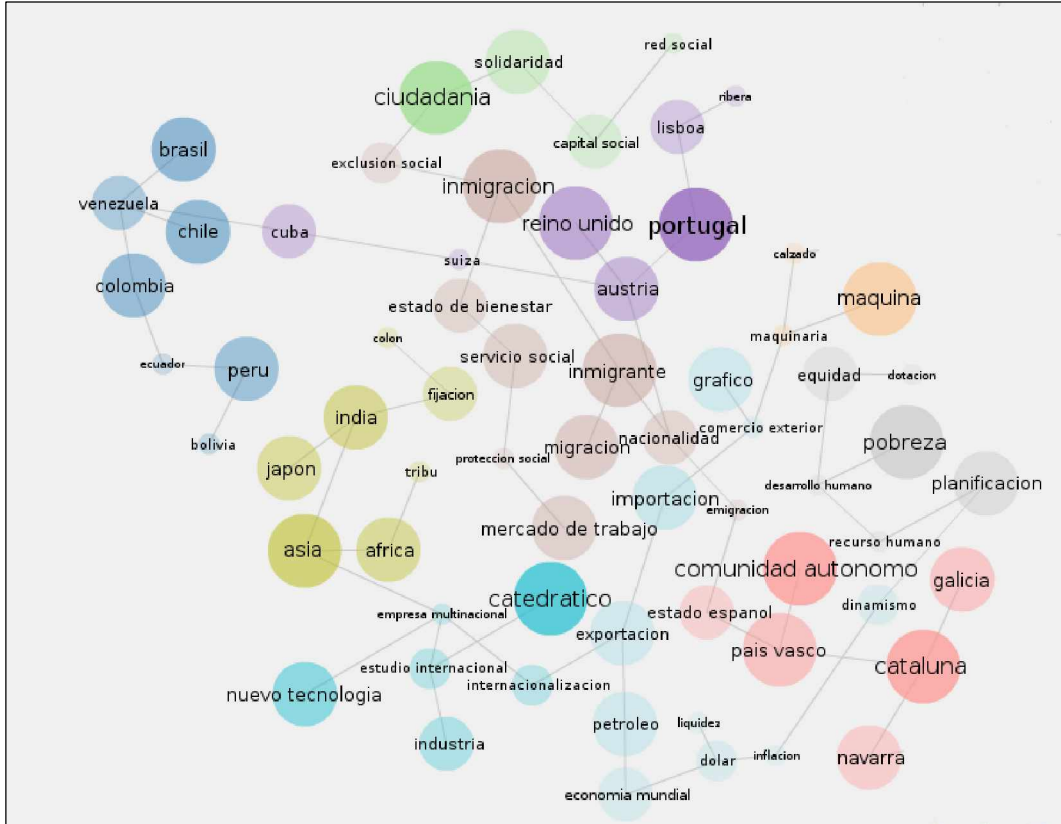


Gráfico 2. Acercamiento 1 al mapa general



**Gráfico 3. Acercamiento 2 al mapa general**



Los gráficos preliminares aquí presentados resultan efectivos como muestra de los avances que se van alcanzando. Sin embargo, es importante destacar que se trata del resultado de las pruebas iniciales del proceso de desarrollo. El perfeccionamiento de las técnicas de extracción de conceptos, *clustering* y visualización, junto con el procesamiento de un mayor número de documentos, darán como resultado conceptos más específicos y mapas más ricos en términos y relaciones.

Los mapas que se realicen como producto final de este trabajo serán completamente navegables e interactivos a través de la web. Incluirán también documentos en inglés y portugués y permitirán funciones como la búsqueda de términos, su comparación entre varias universidades y el acceso directo a los documentos a texto completo que han generado los nodos y sus relaciones.

### **3. La selección de documentos representativos**

Una parte vital de esta tarea es la selección de los documentos adecuados para la extracción de conceptos, que deben cumplir una doble condición. En primer lugar, ser de carácter académico, ya que una buena proporción de los documentos (incluso en formatos ricos en contenido) tienen que ver con la gestión y la divulgación más que con la producción de conocimiento. Por otra parte, es necesario que los documentos cuenten con el texto completo para poder realizar correctamente la extracción de

conceptos y su posterior mapeo, ya que en muchos casos se trata sólo de citas bibliográficas o breves resúmenes.

Luego de diversas pruebas y ensayos, para satisfacer la primera condición se ha decidido aprovechar las posibilidades que ofrece *Google Scholar* para la identificación del contenido de carácter académico en Internet. El éxito y difusión de esta herramienta ofrece posibilidades muy útiles para distinguir automáticamente los documentos académicos del resto de la información publicada en los sitios universitarios, permitiendo además hacer búsquedas específicas dentro de los sitios o sub-sitios de cada institución. Estas bondades son cada vez más explotadas para estudios bibliométricos y análisis de la producción científica de investigadores e instituciones, con resultados fuertemente alentadores.<sup>4</sup>

Sin embargo, los documentos presentados por *Google Scholar* no son siempre aptos para el tipo de procesamiento aquí propuesto, dado que se requiere de una cantidad de texto suficiente para que los resultados obtenidos sean plenamente representativos. Esto no es posible, por ejemplo, con las citas bibliográficas o las presentaciones PowerPoint. Por otra parte, incluso dentro de *Google Scholar* aparecen documentos no útiles para este trabajo, como por ejemplo los programas de las materias. Todos estos documentos, más allá del formato en que estén no son etiquetados por el sistema como "pdf", por lo que la selección final se ajusta perfectamente a los requerimientos.

El volumen de información disponible a través de los archivos pdf en *Google Scholar*, alojados dentro de las páginas web universitarias, es muy significativo, ascendiendo a casi medio millón de documentos para el total de las universidades iberoamericanas, en una consulta realizada el 30 de abril de 2009. Al igual que en la producción científica en general, la distribución de esos documentos presenta una marcada concentración en los países de mayor desarrollo relativo en ciencia y tecnología de la región.

En esta etapa inicial se desarrolló *software* capaz de automatizar las consultas, descargas, clasificaciones y post-procesamiento de los documentos indexados por *Google Scholar* y alojados en los sitios web de las universidades. A partir de estas herramientas, que poseen además la capacidad de generar las estadísticas básicas de los documentos descargados, ha sido posible generar una descripción general de los contenidos, que se presenta a continuación.

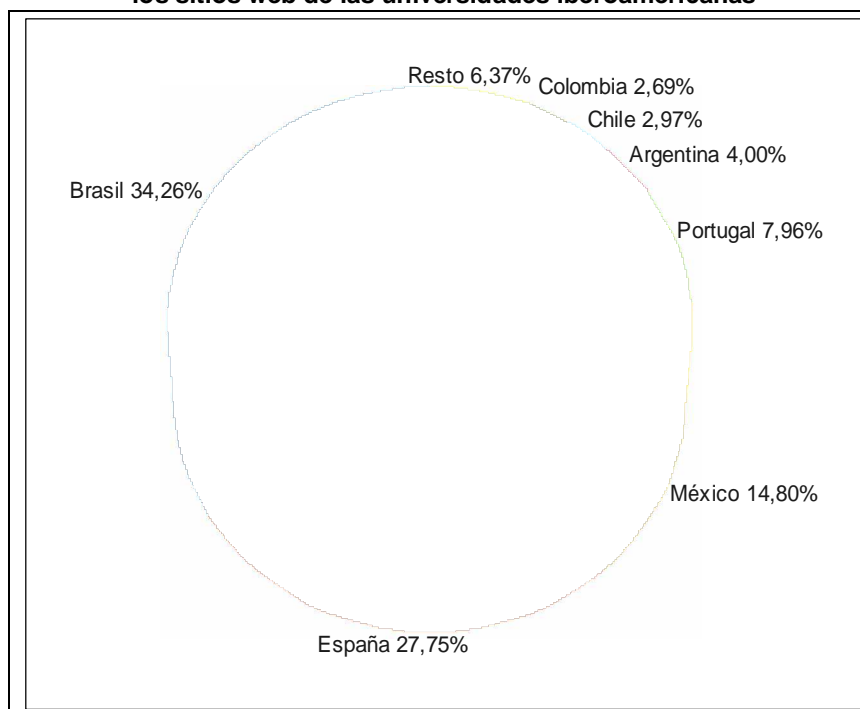
El gráfico 4 presenta la distribución de los archivos pdf identificados por *Google Scholar* en los sitios web de las universidades iberoamericanas. Siete países acumulan casi el 95% del total de documentos. Las universidades brasileñas acumulan el 34%, seguidas por las españolas con el 27% y las mexicanas con el 14%.

Con una participación menor aparecen Portugal y Argentina, con el 7% y 4% respectivamente, seguidos de Chile y Colombia con participaciones cercanas al 2%. Las universidades del conjunto restante de países iberoamericanos acumulan tan sólo el 6% de los documentos totales identificados.

---

<sup>4</sup> HARZINGL, Anne-Wil; VAN DER WAL, Ron . Google Scholar as a new source for citation analysis. *Ethics In Science And Environmental Politics*, 2008.

**Gráfico 4. Distribución por país de los archivos pdf identificados por *Google Scholar* en los sitios web de las universidades iberoamericanas**



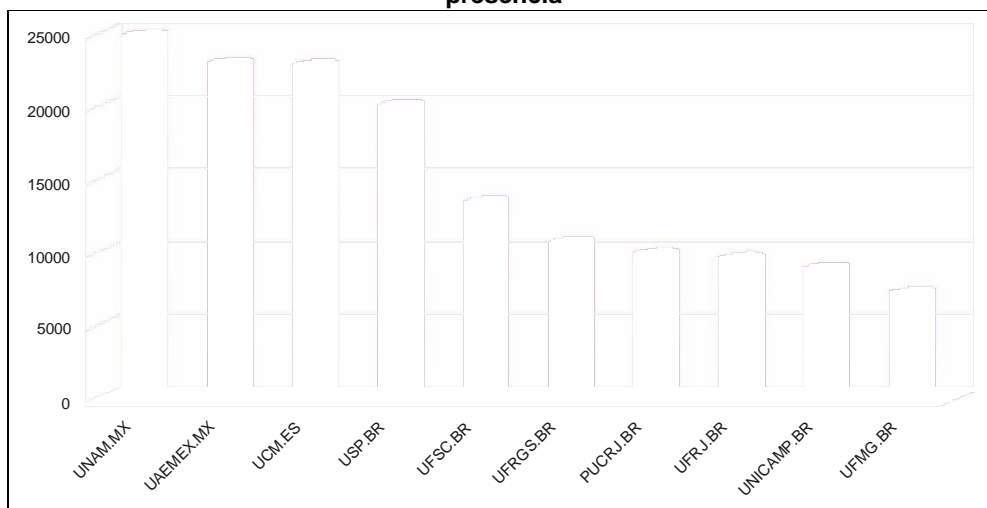
Por otra parte, la distribución de documentos por universidad resulta fuertemente polarizada, de la misma manera que sucede en muchos aspectos del análisis bibliométrico, como en la productividad de los autores. Las primeras seis universidades -de un total de 321 identificadas en este estudio- con mayor número de documentos pdf indexados por *Google Scholar* en sus sitios web suman el 25% del total iberoamericano, mientras que las primeras veinticinco superan el 50%.

El gráfico 5 presenta las diez universidades iberoamericanas con mayor cantidad de documentos pdf publicados en sus sitios web e indexados por *Google Scholar*. Las dos primeras son mexicanas, la Universidad Autónoma Nacional de México con 24.800 documentos y la Universidad Autónoma del Estado de México con 22.900. Con un volumen muy similar aparece en tercer lugar la española Universidad Complutense de Madrid.

El resto de la lista se completa con universidades brasileñas. Sin embargo, la única con una cantidad de documentos similar al conjunto anterior es la Universidad de San Pablo, con 20.000 documentos. La siguiente en este ordenamiento es la Universidad de Santa Catarina, con 13.400.



**Gráfico 5. Cantidad de documentos detectados en las diez universidades de mayor presencia**



#### **4. Selección del conjunto inicial de universidades**

Una de las características de este proyecto es su naturaleza escalable. En principio, a fin de llevar adelante la etapa inicial del desarrollo metodológico e informático, se tomará un conjunto de cuatro universidades, seleccionadas en base a la representatividad de las características de su presencia en la web con respecto a los rasgos del conjunto total de las universidades iberoamericanas.

Posteriormente, se realizará el mapeo de los documentos publicados en los sitios web de una decena de universidades iberoamericanas y se creará el portal donde serán alojados los gráficos interactivos de cada una, junto con las herramientas de navegación y búsqueda, así como los documentos metodológicos que le dan respaldo.

Para la identificación de las cuatro universidades del conjunto inicial, se realizó un relevamiento de los documentos indexados en las 320 universidades iberoamericanas presentes en el Ranking Web de Universidades del Mundo (<http://www.webometrics.info/>). Un resumen de los resultados obtenidos se presenta en la tabla 1.

**Tabla 1. Resumen del resultado del relevamiento de documentos en sitios web de universidades iberoamericanas**

	<b>UNIVERSIDAD</b>	<b>PAÍS</b>	<b>URL</b>	<b>CANTIDAD</b>
1	Universidad Nacional Autónoma de México	MX	<a href="http://www.unam.mx/">http://www.unam.mx/</a>	24800
2	Universidad Autónoma del Estado de México	MX	<a href="http://www.uaemex.mx/">http://www.uaemex.mx/</a>	22900
3	Universidad Complutense de Madrid	ES	<a href="http://www.ucm.es/">http://www.ucm.es/</a>	22800
4	Universidade de São Paulo	BR	<a href="http://www.usp.br/">http://www.usp.br/</a>	20000
5	Universidade Federal de Santa Catarina Brasil	BR	<a href="http://www.ufsc.br/">http://www.ufsc.br/</a>	13400
6	Universidade Federal do Rio Grande do Sul	BR	<a href="http://www.ufrgs.br/">http://www.ufrgs.br/</a>	10600
7	Pontificia Universidad Católica do Rio de Janeiro	BR	<a href="http://www.puc-rio.br/">http://www.puc-rio.br/</a>	9850
8	Universidade Federal do Rio de Janeiro	BR	<a href="http://www.ufrj.br/">http://www.ufrj.br/</a>	9550
9	Universidade Estadual de Campinas	BR	<a href="http://www.unicamp.br/">http://www.unicamp.br/</a>	8840
10	Universidade Federal de Minas Gerais	BR	<a href="http://www.ufmg.br/">http://www.ufmg.br/</a>	7160
11	Universidade Estadual Paulista	BR	<a href="http://www.unesp.br/">http://www.unesp.br/</a>	6780
12	Universidade do Porto	PT	<a href="http://www.up.pt/">http://www.up.pt/</a>	6610
13	Universidad de Granada	ES	<a href="http://www.ugr.es/">http://www.ugr.es/</a>	5980
14	Universidad Nacional Mayor de San Marcos	PE	<a href="http://www.unmsm.edu.pe/">http://www.unmsm.edu.pe/</a>	5810
15	Universidad de Sevilla	ES	<a href="http://portal.us.es/">http://portal.us.es/</a>	5470
16	Universidad del País Vasco	ES	<a href="http://www.ehu.es/">http://www.ehu.es/</a>	5260
17	Universidade Técnica de Lisboa	PT	<a href="http://www.utl.pt/">http://www.utl.pt/</a>	5000
18	Universitat Autònoma de Barcelona	ES	<a href="http://www.uab.es/">http://www.uab.es/</a>	4960
19	Universidad de Murcia	ES	<a href="http://www.um.es/">http://www.um.es/</a>	4250
20	Universidad Nacional del Nordeste	AR	<a href="http://www.unne.edu.ar/">http://www.unne.edu.ar/</a>	4060
21	Universitat de Valencia	ES	<a href="http://www.uv.es/">http://www.uv.es/</a>	3950
22	Universitat Politècnica de Catalunya	ES	<a href="http://www.upc.es/">http://www.upc.es/</a>	3830
23	Universidade de Lisboa	PT	<a href="http://www.ul.pt/">http://www.ul.pt/</a>	3540
24	Universidade de Coimbra	PT	<a href="http://www.uc.pt/">http://www.uc.pt/</a>	3540
25	Universitat de Barcelona	ES	<a href="http://www.ub.edu/">http://www.ub.edu/</a>	3510
26	Universidad de Chile	CL	<a href="http://www.uchile.cl/">http://www.uchile.cl/</a>	3500
27	Universidade de Brasília	BR	<a href="http://www.unb.br/">http://www.unb.br/</a>	3470
28	Pontificia Universidad Católica de Chile	CL	<a href="http://www.puc.cl/">http://www.puc.cl/</a>	3400
29	Universidade Federal Fluminense	BR	<a href="http://www.uff.br/">http://www.uff.br/</a>	3350
30	Universitat d'Alacant	ES	<a href="http://www.ua.es/">http://www.ua.es/</a>	3300
31	Universidade Federal da Bahia	BR	<a href="http://www.ufba.br/">http://www.ufba.br/</a>	3280
32	Universidad de Buenos Aires	AR	<a href="http://www.uba.ar/">http://www.uba.ar/</a>	3170
33	Universidad de Guadalajara	MX	<a href="http://www.udg.mx/">http://www.udg.mx/</a>	3170
34	Universidad de Costa Rica	CR	<a href="http://www.ucr.ac.cr/">http://www.ucr.ac.cr/</a>	3120
35	Universidad Politécnica de Madrid	ES	<a href="http://www.upm.es/">http://www.upm.es/</a>	3050
36	Universidad Politécnica de Valencia	ES	<a href="http://www.upv.es/">http://www.upv.es/</a>	2910
37	Universitat de Girona	ES	<a href="http://www.udg.es/">http://www.udg.es/</a>	2860
38	Universidade Federal de Santa Maria	BR	<a href="http://www.ufsm.br/">http://www.ufsm.br/</a>	2840
39	Universidad de Zaragoza	ES	<a href="http://www.unizar.es/">http://www.unizar.es/</a>	2820
40	Universidade Federal do Paraná	BR	<a href="http://www.ufpr.br/">http://www.ufpr.br/</a>	2690
41	Universidad Autónoma de Madrid	ES	<a href="http://www.uam.es/">http://www.uam.es/</a>	2670
42	Universidad de los Andes Mérida	VE	<a href="http://www.ula.ve/">http://www.ula.ve/</a>	2650

43	Universidad de Castilla la Mancha	ES	<a href="http://www.uclm.es/">http://www.uclm.es/</a>	2570
44	Universidade Federal de Lavras	BR	<a href="http://www.ufla.br/">http://www.ufla.br/</a>	2560
45	Universidade de Vigo	ES	<a href="http://www.uvigo.es/">http://www.uvigo.es/</a>	2430
46	Universidad Nacional de Colombia	CO	<a href="http://www.unal.edu.co/">http://www.unal.edu.co/</a>	2430
47	Universidade Federal de São Carlos	BR	<a href="http://www.ufscar.br/">http://www.ufscar.br/</a>	2410
48	Universidad Carlos III de Madrid	ES	<a href="http://www.uc3m.es/">http://www.uc3m.es/</a>	2240
49	Universidade do Minho	PT	<a href="http://www.uminho.pt/">http://www.uminho.pt/</a>	2160
50	Universidade Estadual de Maringá	BR	<a href="http://www.uem.br/">http://www.uem.br/</a>	2110
100	Universidad de Huelva	ES	<a href="http://www.uhu.es/">http://www.uhu.es/</a>	1060
150	Universidad de Sonora	MX	<a href="http://www.uson.mx/">http://www.uson.mx/</a>	476
200	Universidad del Salvador	AR	<a href="http://www.salvador.edu.ar">http://www.salvador.edu.ar</a>	301
250	Universidad de San Andrés Buenos Aires	AR	<a href="http://www.udesa.edu.ar/">http://www.udesa.edu.ar/</a>	129
300	Universidad Autónoma Chapingo	MX	<a href="http://www.chapingo.mx/">http://www.chapingo.mx/</a>	29

Nota: Consulta realizada el 30 de abril de 2009.

El objetivo de esta aproximación cuantitativa fue el de obtener un panorama general que brinde pistas para la selección de las primeras universidades, aunque dado que el objetivo de este trabajo es ofrecer una aproximación a los contenidos, si bien se requiere una cierta masa crítica de documentos, una vez que esta es garantizada el volumen total resulta de una importancia relativa.

Un primer criterio fue el de seleccionar universidades de países hispano parlantes. En esta primera etapa se están perfeccionando las herramientas de procesamiento de textos en español e inglés, para luego incorporar el idioma portugués en el segundo conjunto de instituciones.

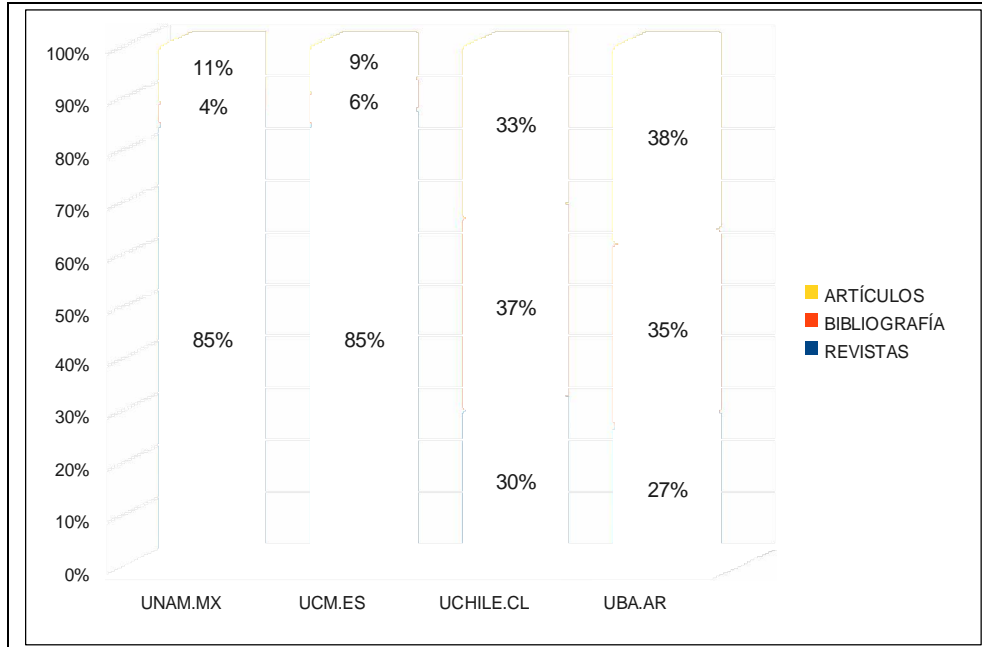
Dentro de éstas se seleccionaron dos con un gran volumen de documentos y dos de tamaño intermedio. Estas fueron la Universidad Autónoma Nacional de México, con 24.800 documentos, la Universidad Complutense de Madrid, con 22.800, la Universidad de Chile, con 3.500, y la Universidad de Buenos Aires, con 3.170. Se obtuvo también así una cierta representatividad geográfica de los países de mayor presencia en el ámbito de las web universitarias.

Paralelamente, se ha realizado un análisis de los tipos de documentos que se identificaron dentro de las web de cada universidad. Estos documentos se han clasificado en tres categorías, cuyas características e impactos sobre los resultados serán evaluadas en las próximas etapas del proyecto. Las categorías son:

- Revistas editadas por la universidad, de acceso libre y disponibles a texto completo.
- Bibliografía, generalmente publicada en otras revistas y utilizada como parte de la actividad docente.
- Artículos firmados por los docentes o investigadores de la universidad, presentados como *preprints* o incluso, en muchos casos, como separatas electrónicas de las revistas en que fueron publicados.

El gráfico 6 presenta la distribución por tipo de documento, según las categorías anteriormente descritas, de los archivos pdf detectados dentro de los sitios web de cada una de las cuatro universidades seleccionadas para el grupo inicial.

**Gráfico 6. Distribución por tipo de documento dentro de los sitios web de las universidades seleccionadas**



La Universidad Nacional Autónoma de México y la Universidad Complutense de Madrid presentan patrones similares. En ambas, el 85% de los documentos pertenecen a las revistas editadas por las universidades, mientras que el resto de los tipos de documento tienen una presencia mucho menor.

La Universidad de Chile y la Universidad de Buenos Aires, en cambio, tienen una distribución en partes prácticamente iguales de revistas, bibliografía y artículos. Por esta variedad de patrones es que se han elegido a estas universidades como sujetos para las primeras etapas del desarrollo de este proyecto.