7.
Indicadores de producción científica

Sistemas científicos complejos y su abordaje metodológico

Rodolfo Barrere, Martín Bageneta y Lautaro Matas*

1. Introducción

Se entiende por "sistemas científicos complejos" al conjunto de organismos e instituciones que en un país determinado cuentan con un universo de investigadores significativo y dentro del cual un conjunto importante de los recursos humanos se desempeña en más de una institución. Este fenómeno, muy común en la dinámica de investigación actual, tiene un fuerte impacto en la medición de la producción científica. Las múltiples afiliaciones institucionales de los autores conforman conjuntos de instituciones solapadas en su producción y, consecuentemente, generan problemas metodológicos al momento de identificar la pertenencia institucional de los trabajos en las bases de datos disponibles, a la hora de construir indicadores bibliométricos.

Este inconveniente está relacionado con dos cuestiones vinculadas con la información primaria disponible en las bases de datos bibliográficas comúnmente utilizadas. Por un lado, los autores no siempre incluyen referencias a todas sus instituciones de pertenencia. Por otro lado, las bases de datos no cuentan con un trabajo de normalización en el campo que recoge la firma institucional de los autores. Este inconveniente puede afectar fuertemente los resultados obtenidos en los indicadores de producción científica desagregados hasta el nivel de las instituciones de investigación, que resultan una importante herramienta de evaluación de las actividades científicas y tecnológicas, así como un excelente insumo para la toma de decisiones.

En Argentina, la investigación científica y tecnológica es llevada a cabo, principalmente, por instituciones del sector público. Entre ellas se destacan las unidades de I+D del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y del sistema universitario.

El CONICET es un ente autárquico que tiene como misión el fomento y la ejecución de actividades científicas y tecnológicas en todo el territorio nacional y en las distintas áreas del conocimiento. El Consejo cuenta con alrededor de cien unidades ejecutoras, entre institutos de investigación, centros regionales y de servicios. Su personal supera las diez mil personas, incluyendo alrededor de cinco mil investigadores, tres mil becarios y dos mil profesionales en el rol de personal de apoyo. Sin embargo, cerca del 50% de sus investigadores y becarios tienen por lugar de trabajo centros de investigación radicados en alguna de las 38 universidades nacionales existentes.

Este solapamiento en la dependencia institucional tiene implicancias a la hora de medir la producción científica, tanto del CONICET como de las universidades nacionales, dado que los resultados de la investigación que se publican deben ser computados para las dos instituciones que la hicieron posible.

Por los motivos expuestos, la identificación completa y normalizada de la afiliación de

* Los autores son miembros del Centro Argentino de Información Científica y Tecnológica (CAICYT-CONICET) (correos electrónicos: rbarrere@gmail.com, mbageneta@caicyt.gov.ar, lmatas@gmail.com).

los autores que firman un artículo resulta una tarea compleja, que implica el tratamiento de los datos originales y la utilización de otras fuentes de información. A continuación se describe la metodología desarrollada por el Centro Argentino de Información Científica y Tecnológica (CAICYT-CONICET) para solucionar este problema en lo que hace a la producción de los investigadores del CONICET en el *Science Citation Index* (SCI).

Este trabajo presenta los resultados del análisis de la producción científica argentina registrada en esa fuente. Se detalla comparativamente el volumen de la producción de las principales instituciones, las redes de colaboración que se han establecido entre ellas y los distintos niveles de colaboración con otros países, en base a la observación de las coautorías.

2. Los indicadores bibliométricos

La medición de los resultados de la ciencia a partir de las publicaciones requiere una reflexión inicial sobre el objeto de análisis de los indicadores bibliométricos: el sistema de publicación de la ciencia y su papel en los procesos de producción del conocimiento. Las revistas científicas, junto con las pautas y reglas que regulan su funcionamiento, son el canal por el cual los investigadores hacen público de manera "oficial" el resultado de su trabajo. El conjunto de las publicaciones científicas encarna, entonces, el acervo de conocimiento disponible y, a la vez, demarca el campo y da escenario a los debates científicos.

La validez de los indicadores bibliométricos depende de que su materia prima, las publicaciones científicas, sean representativas de los resultados de la investigación. Esto implica que existan mecanismos que garanticen la estabilidad de sus contenidos y la calidad de los mismos. Esos mecanismos están dados por estrictas restricciones de acceso a la publicación, basados en la evaluación de pares, de manera que los miembros de la propia comunidad científica, desempeñan alternativamente el papel de evaluador o de evaluado. De esta manera se establece un mecanismo de revisión rigurosa que permite confiar en la calidad de la obra de otros que se encuentra publicada. Estos mecanismos de control de calidad y adecuación a los cánones científicos están orientados a mantener el crédito y la reputación de los miembros de la comunidad.

Según Maltrás (2003), la revisión que sirve de filtro al ingreso del sistema de publicación se caracteriza por tres aspectos: paridad (es realizada por colegas de la misma condición), pluralidad (se demanda el dictamen de varios árbitros) y anonimato (la identidad del autor y de los árbitros se mantiene oculta en todo el proceso de evaluación).

En base a lo considerado anteriormente, es posible dar cuenta de la base teórica de los indicadores bibliométricos, como expresiones de la producción del conocimiento. Por un lado, la necesidad de reconocimiento (reflejada también en los mecanismos burocráticos de evaluación de los investigadores en base a sus publicaciones) impulsa a los científicos a publicar todos sus resultados, mientras que el sistema de revisión por pares de las revistas científicas garantiza que los documentos que se contabilizan tengan calidad y originalidad científica.

La fuente más difundida para los indicadores bibliométricos consiste en la extracción de información estadística de bases bibliográficas. Estas fuentes de información cuentan con datos acumulados durante muchos años, de los documentos publicados en revistas científicas seleccionadas. Contienen referencias bibliográficas que incluyen el título del artículo, sus autores, la pertenencia institucional de los mismos, la revista de publicación y el abstract del documento, entre otros datos. Existen bases

multidisciplinarias, como el *Science Citation Index* (utilizada en este estudio) y *Pascal*, y otras de disciplinas específicas, como *Medline* o *Chemical Abstracts*.

La selección de las revistas que son indexadas en esas bases de datos se realiza con fuertes criterios de calidad editorial (reconocimiento del comité editor, calidad académica de los encargados del referato, etc.), opiniones de expertos y análisis de las citas recibidas por las revistas como una muestra de su visibilidad. Esa selección también debe garantizar una correcta cobertura de los temas que la base de datos pretende cubrir. En el caso de las bases internacionales se busca cubrir la corriente principal (*mainstream*) de la ciencia internacional.

Sin embargo, los indicadores bibliométricos tienen ciertas limitaciones como método de medición de la producción científica. En primer lugar, la investigación tiene diversos aspectos que no pueden ser captados por los estudios bibliométricos. Siguiendo a María Bordons (2001), la investigación incluye, además de la científica básica expresada en la publicación de resultados, tareas de carácter tecnológico, docente y social. La bibliometría sólo puede abordar la faceta científica, mientras que el resto de las actividades deben ser estudiadas por otro tipo de indicadores.

Por otra parte, no existen bases bibliográficas capaces de cubrir la totalidad de la producción científica de un país. En general estas fuentes intentan representar la corriente principal (*mainstream*) internacional de la ciencia, mediante la selección de las revistas más representativas de cada tema para la comunidad de los investigadores, principalmente pertenecientes a los países centrales. Esto implica que los temas que interesan a esa corriente principal se verán representados, mientras que otros casi no aparecerán. Este fenómeno afecta fuertemente a la los países latinoamericanos, cuyos temas de investigación, en algunas disciplinas más que en otras, pueden divergir de aquellos estudiados en los países más desarrollados.

A pesar de estos inconvenientes, los indicadores bibliométricos se han convertido en las herramientas más difundidas para la medición de la producción científica en todo el mundo, ya que brindan un enfoque objetivo y comparable. La madurez de estas metodologías puede verse en la amplia bibliografía existente sobre el tema, incluyendo documentos de la OCDE, el organismo rector de las estadísticas de ciencia, tecnología e innovación a nivel mundial (Okubo, 1997).

3. Metodología para la normalización de afiliaciones institucionales

Si bien los indicadores bibliométricos resultan una importante herramienta para la gestión institucional y la evaluación de las actividades científicas y tecnológicas, la identificación de la afiliación de los autores que firman un artículo resulta una tarea compleja.

La fuente utilizada en este estudio fue el *Science Citation Index*, una base interdisciplinaria, que recopila las citas bibliográficas de todos los documentos publicados en alrededor de ocho mil revistas científicas internacionales de primer nivel. En el SCI, como en la mayoría de las bases de datos bibliográficas, se cuenta con un campo que incluye la afiliación institucional de los autores, pero éste carece de normalización. De esta manera, una misma institución puede figurar de maneras muy diversas, a causa de la utilización de siglas distintas o de abreviaturas diferentes. Por este motivo, resulta imposible hacer una clasificación automática de la afiliación institucional de los registros sin un tratamiento previo de los datos.

La metodología desarrollada consta de dos etapas sucesivas. La primera de ellas, totalmente automatizada, identifica las principales instituciones del sistema científico

argentino en las firmas incluidas en los registros SCI. La segunda etapa, que aprovecha los resultados de la anterior, soluciona el problema de la dependencia institucional múltiple de los investigadores del CONICET argentino e incluye un componente automático y otro asistido por operadores para los casos de mayor complejidad. A continuación se detallan las técnicas y características de cada una de las etapas.

3.1. Etapa I: clasificación automática de instituciones con uso de expresiones regulares

Teniendo en cuenta la problemática que representa la identificación de la filiación en los grandes sistemas complejos se desarrolló como primer paso un método de clasificación automática. Como punto de partida se generó una tabla con las instituciones que se deseaba identificar, en este caso las universidades nacionales, institutos nacionales de investigación y el total de las unidades ejecutoras del CONICET. Para cada una de ellas se incluyó el nombre completo en español, el nombre completo en inglés y una o varias abreviaturas normalmente utilizadas.

Esta tabla sirve como base de conocimiento sobre la que se contrastan los registros de afiliación institucional disponibles en cada uno de los artículos descargados del SCI. Sin embargo, dada la falta de normalización de los datos, ya mencionada anteriormente, esta tarea no puede realizarse de forma directa.

Fue necesario para ello recurrir a un procedimiento capaz de identificar patrones flexibles en las firmas institucionales. Se optó entonces por la aplicación de técnicas de expresiones regulares que subsanan las variaciones en las abreviaturas, entre otros casos de desnormalización. Para ello se parte de la base de conocimiento anteriormente generada, tomando las tres primeras letras de cada una de las palabras de más de tres letras que componen el nombre de la institución se genera un patrón. Por ejemplo, para el caso de la Universidad de Buenos Aires el patrón resultante es “Uni ->_Bue ->_Air”.

A continuación, se extraen de la base de datos todas las afiliaciones de instituciones argentinas y se intenta parear con ellas el patrón generado, con arbitraria cantidad de caracteres entre los trigramas que los componen, pero siempre manteniendo el orden entre ellos.

De esta manera, distintas formas de nombrar a la Universidad de Buenos Aires, como por ejemplo Univ Buenos Aires, Univ de Buenos Aires o Univers Buenos Aires son automáticamente detectadas y normalizadas, sin la intervención manual de operadores.

Esta técnica acelera mucho el proceso de clasificación en relación con la alternativa del trabajo manual, permitiendo analizar importantes volúmenes de datos en pocos minutos, aunque en su etapa de desarrollo requirió una fuerte tarea de análisis de los datos y posterior chequeo de los resultados. Asimismo, al ser independiente de la toma de decisiones de operadores, los criterios a lo largo del tiempo se mantienen y los resultados, por lo tanto, son totalmente comparables.

Con las instituciones incluidas en la base de conocimiento, mencionadas anteriormente, se obtuvo al menos una referencia institucional para el 95% de los registros argentinos de cada año. Por otra parte, se obtuvieron márgenes de error inferiores al 2% de la publicación total de cada institución.

3.2. Etapa II: clasificación de autores a partir de grupos de confianza

La identificación de los artículos del CONICET, sin embargo, no puede ser abordada con esta misma técnica, ya que muchos investigadores tienen una doble dependen-

cia institucional, manteniendo como lugar de trabajo universidades, institutos o centros que no pertenecen a la esfera del Consejo. Si bien en ocasiones los autores mencionan su afiliación institucional al CONICET junto con su otra dependencia, esto no sucede en todos los casos.

Por ese motivo, como segundo paso se decidió identificar la producción de cada uno de los investigadores en la base de datos de 2005, utilizando como insumo el listado completo de personal del CONICET, incluyendo investigadores, becarios y personal de apoyo de todas las áreas disciplinares.

Al igual que la identificación de las afiliaciones institucionales, la detección de los artículos de cada investigador presenta ciertas dificultades. Por un lado, los nombres tampoco están normalizados, de manera que un mismo investigador puede aparecer de formas distintas, sobre todo en el caso de los apellidos compuestos. Además, el SCI sólo consigna el apellido y las iniciales de los autores, por lo que resulta aún más complejo identificar de forma unívoca a cada autor (Javier Fernández y Josefina Fernández resultan indistinguibles: "J. Fernández"). Para garantizar la correcta asignación de registros se utilizaron variables accesorias que permiten verificar la relación entre autores y publicaciones. Particularmente, se verificó el lugar de trabajo declarado en las bases de personal del CONICET y la disciplina científica consignada.

La metodología para la asignación de las publicaciones de cada investigador se basa en la construcción de distintos grupos de confianza. Mediante un cruce inicial de bases de datos entre el SCI y la base de personal de CONICET se identificó un conjunto de publicaciones en las que los apellidos de autores e investigadores eran coincidentes. En 2006, este conjunto ronda los 10.000 autores.

Luego de diversos ensayos, que permitieron constatar la confiabilidad del método, se asignaron automáticamente aquellos artículos en los que coincidían el apellido, las iniciales completas y el lugar de trabajo, que había sido previamente codificado, con lo cual se logró identificar el 40% de los autores.

Posteriormente el 60% restante es asignado, con la supervisión de técnicos, mediante un sistema informático centralizado, con interfase web, que asigna a cada publicación su autor más probable dentro de la base de personal de CONICET. Para facilitar la decisión se presentan las citas bibliográficas junto con los datos de lugar de trabajo y disciplina de los investigadores.

A partir de la metodología aplicada se obtuvieron significativos resultados en la asignación de la producción del CONICET. Los registros que firman como esta institución y que son reconocidos sin esta metodología rondan el 40% del total de la producción identificada luego de su aplicación.

A continuación se presentan indicadores descriptivos de la distribución institucional de los artículos argentinos registrados en el SCI. Para los indicadores de colaboración entre instituciones se utilizó la metodología de contabilización por enteros, es decir que un artículo firmado por investigadores de más de una institución se contabiliza como uno para cada una de las entidades participantes.

4. Distribución institucional de la producción

Dentro de la producción argentina, la institución que presenta una mayor producción es el CONICET, cuyos investigadores tienen participación en el 71,7% de los artículos registrados en 2006. Sin embargo, si se agrupa el conjunto de las universidades nacionales (UU.NN.), éstas también alcanzan una presencia muy importante, con el 70,5%. La tabla 1 presenta los valores para cada conjunto y para los artículos que tienen en común.

Tabla 1. Producción científica del CONICET y las universidades nacionales en SCI 2006

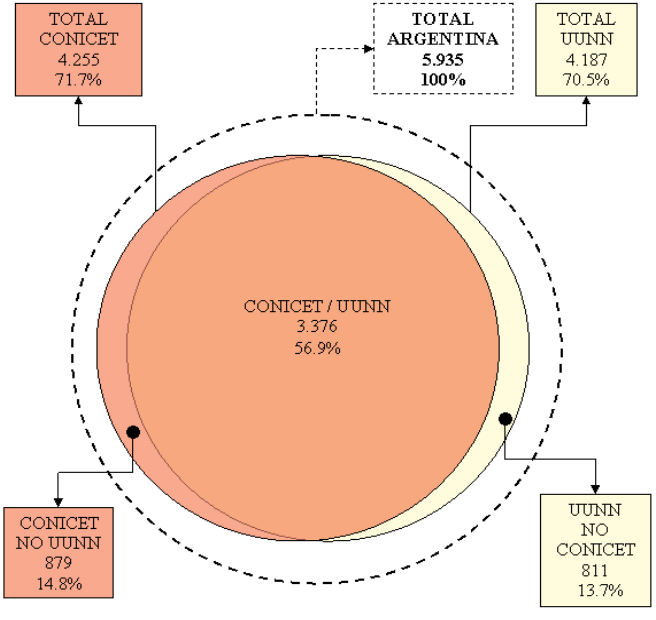
CONICET	4 255	71,7%
UU.NN.	4 187	70,5%
CONICET y UU.NN.	3 376	56,9%
TOTAL ARGENTINA	5 835	100,0%

Nota: En los casos de artículos firmados en conjunto por el CONICET y las universidades nacionales, se ha contabilizado como una publicación completa para cada uno. Por ese motivo la suma de las publicaciones de cada institución da un resultado mayor que el total.

Como se observa en el gráfico 1, existe una importante superposición de ambos conjuntos, que alcanza el 56,9% de la producción total. En esta intersección se incluyen tanto los artículos publicados en colaboración por ambas instituciones, como los publicados por autores de doble dependencia institucional, es decir investigadores miembros del CONICET con lugar de trabajo en alguna de las universidades nacionales.

Es también interesante observar el volumen de artículos que tanto el CONICET como las universidades nacionales aportan de manera independiente al conjunto total de la producción argentina. En ese sentido, el gráfico 1 muestra que el CONICET aporta un 14,8% de las publicaciones sin la participación de las universidades, mientras que estas generan un conjunto separado del 13,7%.

Gráfico 1. Producción del CONICET y las universidades nacionales en SCI 2006

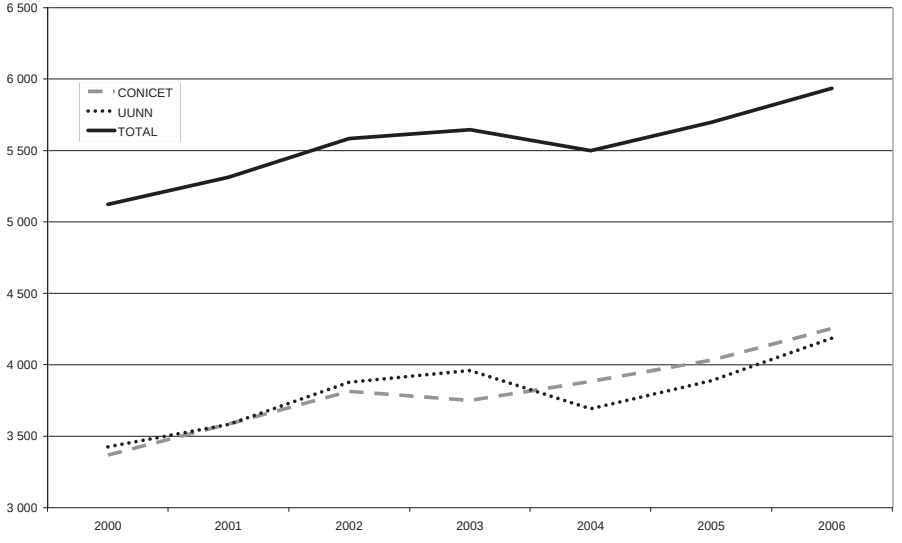


Como se puede ver en el gráfico 2, la producción científica del CONICET presenta un crecimiento sostenido desde 2000, sólo interrumpido por una leve caída en 2003. El conjunto de las universidades nacionales, en cambio, presenta un crecimiento moderado en 2003 y un descenso marcado en 2004. Estos fenómenos, que por supuesto se reflejan también en la producción total de Argentina, pueden estar asociados con la crisis económica de los años previos, cuyo impacto en términos de los resultados de la investigación son apreciables posteriormente.

Desde entonces, y hasta 2006, se ha dado una fuerte recuperación de la producción total argentina en SCI, siendo este último el año de mayor volumen de artículos de la serie disponible. En ese marco, el CONICET presenta un crecimiento sostenido desde 2004, cercano al 6% anual. El conjunto de las universidades, por su parte, inicia su recuperación con posterioridad pero con mayor fuerza, alcanzando una tasa de crecimiento del 8% en 2006.

De esta forma, y luego de los vaivenes ocasionados por la crisis, la producción del CONICET y del conjunto de las universidades vuelve a ser prácticamente equivalente, como lo había sido hasta 2002. Es importante también señalar que ha crecido la superposición de ambos conjuntos, dada por la firma conjunta de artículos o por la autoría de investigadores con una doble pertenencia institucional. En el gráfico 2 se puede ver cómo la superposición de estos conjuntos en 2006 es equivalente al 57% de la producción total, mientras que en 2004 era del 53%.

Gráfico 2. Evolución de la producción argentina en SCI



La tabla 2 presenta la evolución de la cantidad de artículos registrados para las instituciones relevadas en el periodo 2000-2006, con excepción del CONICET, que ya fue descripta. El primer lugar lo ocupa la Universidad de Buenos Aires (UBA), con participación en 1.510 artículos en 2006. El segundo lugar corresponde a la Universidad Nacional de La Plata (UNLP), registrando 678 publicaciones. La participación de estas instituciones en el total nacional se ha mantenido relativamente estable en el periodo analizado.

Por otra parte, se observa un alto nivel de copublicaciones entre las distintas instituciones nacionales analizadas. En el 68% de los registros argentinos del SCI participa más de una institución nacional.

Tabla 2. Producción científica por institución

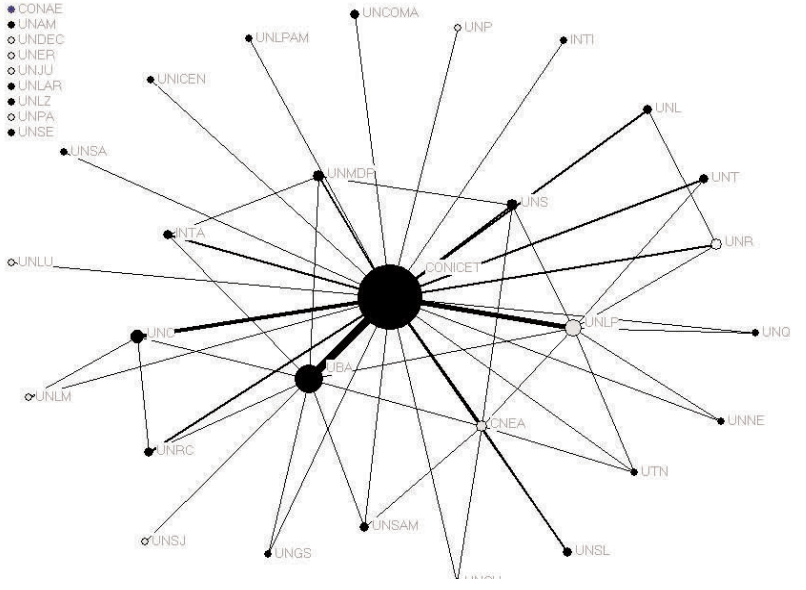
	2000		2001		2002		2003		2004		2005		2006	
	CANT	%	CANT	%	CANT	%	CANT	%	CANT	%	CANT	%	CANT	%
UBA	1 324	25,8%	1 318	24,8%	1 474	26,4%	1 514	26,8%	1 450	26,4%	1 403	24,6%	1 510	25,4%
UNLP	562	11,0%	647	12,2%	676	12,1%	643	11,4%	608	11,1%	654	11,5%	678	11,4%
UNC	356	6,9%	403	7,6%	460	8,2%	468	8,3%	381	6,9%	419	7,4%	466	7,9%
CNEA	361	7,0%	346	6,5%	393	7,0%	311	5,5%	421	7,7%	334	5,9%	331	5,6%
UNR	197	3,8%	193	3,6%	218	3,9%	220	3,9%	207	3,8%	220	3,9%	256	4,3%
UNMDP	146	2,8%	174	3,3%	180	3,2%	216	3,8%	209	3,8%	212	3,7%	248	4,2%
UNS	207	4,0%	215	4,0%	210	3,8%	202	3,6%	211	3,8%	215	3,8%	221	3,7%
UNL	120	2,3%	130	2,4%	141	2,5%	160	2,8%	163	3,0%	185	3,2%	178	3,0%
INTA	105	2,0%	127	2,4%	133	2,4%	141	2,5%	147	2,7%	148	2,6%	160	2,7%
UNT	123	2,4%	155	2,9%	181	3,2%	166	2,9%	118	2,1%	138	2,4%	144	2,4%
UNSL	124	2,4%	128	2,4%	106	1,9%	123	2,2%	97	1,8%	121	2,1%	116	2,0%
UNRC	79	1,5%	73	1,4%	86	1,5%	86	1,5%	61	1,1%	87	1,5%	115	1,9%
UNCOMA	52	1,0%	59	1,1%	72	1,3%	73	1,3%	73	1,3%	74	1,3%	99	1,7%
UNCU	60	1,2%	62	1,2%	78	1,4%	72	1,3%	69	1,3%	75	1,3%	83	1,4%
UNSAM	32	0,6%	44	0,8%	39	0,7%	45	0,8%	53	1,0%	58	1,0%	79	1,3%
UNICEN	47	0,9%	38	0,7%	46	0,8%	45	0,8%	46	0,8%	61	1,1%	61	1,0%
UNNE	37	0,7%	38	0,7%	41	0,7%	43	0,8%	46	0,8%	47	0,8%	49	0,8%
UNQ	30	0,6%	28	0,5%	43	0,8%	30	0,5%	32	0,6%	29	0,5%	42	0,7%
UNP	10	0,2%	7	0,1%	15	0,3%	13	0,2%	16	0,3%	32	0,6%	39	0,7%
UTN	22	0,4%	23	0,4%	26	0,5%	29	0,5%	32	0,6%	30	0,5%	38	0,6%
UNLPAM	23	0,4%	14	0,3%	13	0,2%	20	0,4%	24	0,4%	35	0,6%	37	0,6%
UNSJ	21	0,4%	14	0,3%	25	0,4%	29	0,5%	25	0,5%	20	0,4%	34	0,6%
UNSA	27	0,5%	34	0,6%	32	0,6%	35	0,6%	20	0,4%	37	0,6%	33	0,6%
UNGS	8	0,2%	7	0,1%	16	0,3%	9	0,2%	9	0,2%	15	0,3%	26	0,4%
UNLU	25	0,5%	21	0,4%	22	0,4%	16	0,3%	21	0,4%	24	0,4%	22	0,4%
INTI	9	0,2%	18	0,3%	13	0,2%	11	0,2%	11	0,2%	13	0,2%	19	0,3%
UNLM	8	0,2%	6	0,1%	7	0,1%	10	0,2%	4	0,1%	19	0,3%	15	0,3%
UNER	7	0,1%	7	0,1%	8	0,1%	7	0,1%	10	0,2%	11	0,2%	14	0,2%
UNAM	16	0,3%	12	0,2%	14	0,3%	12	0,2%	15	0,3%	17	0,3%	11	0,2%
UNLZ	6	0,1%	7	0,1%	6	0,1%	6	0,1%	11	0,2%	9	0,2%	8	0,1%
UNPA	7	0,1%	8	0,2%	10	0,2%	7	0,1%	8	0,1%	12	0,2%	8	0,1%
UNJU	11	0,2%	7	0,1%	3	0,1%	10	0,2%	9	0,2%	8	0,1%	5	0,1%
UNSE	3	0,1%	8	0,2%	9	0,2%	8	0,1%	8	0,1%	12	0,2%	5	0,1%
UNDEC	0	0,0%	1	0,0%	0	0,0%	0	0,0%	0	0,0%	2	0,0%	4	0,1%
UNLAR	3	0,1%	4	0,1%	2	0,0%	1	0,0%	0	0,0%	2	0,0%	2	0,0%
CONAE	1	0,0%	3	0,1%	3	0,1%	3	0,1%	3	0,1%	6	0,1%	1	0,0%
UNF	1	0,0%	0	0,0%	0	0,0%	0	0,0%	0	0,0%	1	0,0%	0	0,0%
UNCA	1	0,0%	1	0,0%	0	0,0%	0	0,0%	1	0,0%	0	0,0%	0	0,0%
UNVM	1	0,0%		0,0%	0	0,0%	0	0,0%	1	0,0%	0	0,0%	0	0,0%
UNTREF	0	0,0%		0,0%	0	0,0%	1	0,0%	0	0,0%	0	0,0%	0	0,0%

Nota: En los casos de artículos firmados en conjunto por varias instituciones se ha contabilizado como una publicación completa para cada una. Por ese motivo la suma de las publicaciones de cada institución da un resultado mayor que el total.

El gráfico 3 muestra la red construida a partir de la copublicación de artículos en la base de datos en 2006. El diámetro de los círculos representa la cantidad de registros por institución, mientras que las líneas dan cuenta de los artículos con autores pertenecientes a las dos instituciones que vinculan. El grosor de estas líneas señala la cantidad de artículos compartidos.

Otro fenómeno interesante es la copublicación de artículos en conjunto con instituciones extranjeras. En los últimos quince años la colaboración internacional dentro de la producción argentina ha crecido fuertemente, triplicando su valor porcentual en relación con 1990. Para dar cuenta de esto, se han coloreado las instituciones de acuerdo a si se encuentran por arriba o por debajo de la media (43%). Las esferas de color gris tienen un porcentaje de publicaciones en colaboración internacional superior al 43%, mientras que las de color negro están por debajo de ese valor.

Gráfico 3. Red de colaboración entre instituciones nacionales (SCI 2006)



Nota: Para facilitar la visualización, se presentan los lazos mayores a siete copublicaciones. El color gris señala las instituciones por encima de la media en cuanto a colaboración internacional, el negro a las que se encuentran por debajo.

Para facilitar la visualización, se presentan solamente los lazos mayores a siete copublicaciones. De esta forma se pueden ver los lazos más fuertes que posee cada institución, es decir aquellos que presentan una mayor constancia e intensidad. Esto favorece la visualización de la red, que de otra manera resultaría muy difícil de comprender a simple vista. Si se disminuyera la cantidad mínima de copublicaciones, de manera que los lazos aparecieran en el gráfico, la red se presentaría mucho más conectada, aunque la gran mayoría de las relaciones tienen valores muy bajos. Del total de las relaciones entre instituciones, el 16% se da a través de una sola copublicación y el 50% por menos de cinco copublicaciones.

En el gráfico se puede observar el papel central del CONICET, que más allá de su

dimensión en cuanto a cantidad de publicaciones, cuenta con vínculos con todas las demás instituciones. La Comisión Nacional de Energía Atómica (CNEA), por su parte, muestra una gran integración a la red, ocho enlaces superiores a siete publicaciones, sólo superada por la UBA, con nueve, a pesar de contar con una cantidad de publicaciones mucho menor y una mayor especialización disciplinaria que muchas universidades.

En cuanto a la colaboración internacional, el CONICET tiene un porcentaje de copublicación con autores extranjeros levemente inferior al promedio, al igual que la UBA. Por el contrario, la UNLP, que ocupan el tercer lugar en cantidad de publicaciones, está un poco por encima de la media, con un 44%. Sin embargo, observando el total de las instituciones, no se aprecia una relación lineal entre la cantidad total de publicaciones registradas y la tendencia a colaborar con instituciones de otros países.

5. Comentarios finales

El desarrollo de la metodología para la normalización de la afiliación institucional en bases de datos bibliográficas aquí presentada resulta de particular interés por el nivel de automatización que se alcanzó en el proceso. Ello tiene dos implicancias principales. Por un lado, se ha conseguido acelerar fuertemente la tarea de normalización, lo que permite mantener el proceso, disminuyendo el impacto de la cantidad de artículos sobre los que es necesario trabajar. En el caso de países con un volumen de producción relativamente menor y con sistemas institucionales de I+D integrados por una cantidad de instituciones no muy amplia, la normalización manual de las afiliaciones puede resultar accesible, pero en la medida en que aumenta la cantidad de artículos publicados e instituciones firmantes, la tarea se vuelve virtualmente imposible.

Por otro lado, la identificación de las instituciones mediante técnicas informáticas permite mantener una consistencia de criterios sostenida a lo largo del tiempo, lo que resulta muy difícil en trabajos manuales en los que muchas veces participan más de un operador. Esto permite mejorar la consistencia de los resultados que se obtienen, principalmente en series de tiempo amplias y de actualización regular.

La aplicación de esta metodología en años sucesivos ha agregado un gran valor a la base de datos bibliográfica disponible en el CAICYT, con respecto a los datos originalmente descargados del SCI. Esto garantiza una mayor precisión en los indicadores publicados a nivel institucional y ofrece un gran potencial para estudios en profundidad del sistema científico local, ya sea de manera global o por áreas temáticas de interés, permitiendo la identificación de instituciones y grupos de investigación líderes en cada temática, la evolución de su productividad y sus relaciones con el resto de los agentes de la comunidad científica.

Bibliografía

BORDONS, M. (2001): “Aspectos metodológicos en la obtención de indicadores bibliométricos”, *Cuadernos de Indicios*, N° 1, RICYT.

CALLON, M., COURTIAL, J.-P. y PENAN, H. (1995): *Cienciometría. El estudio cuantitativo de la actividad científica: de la bibliometría a la vigilancia tecnológica*, Madrid, Trea.

GLÄNZEL, W. (2003): *Bibliometrics as a Research Field. A course on theory and application of bibliometric indicators*.

MALTRÁS, B. (2003): *Los indicadores bibliométricos*, Madrid, Trea.

OKUBO, Y. (1997): “Bibliometric indicators and analysis of research systems: Methods and examples”, *STI Working Paper*, Paris, OCDE.

Comparación entre rankings de universidades e instituciones de investigación de Iberoamérica

Mario Fernández, Isidro F. Aguillo,
José Luís Ortega y Begoña Granadino*

1. Introducción

Uno de los objetivos principales a la hora de desarrollar un ranking es el de poder usarlo como una herramienta que permita la valoración objetiva del sujeto estudiado, ayudando así a tomar decisiones relacionadas con la materia analizada. En el caso de los rankings sobre la ciencia y la tecnología, existen en la actualidad diferentes trabajos a nivel internacional que desde distintas aproximaciones intentan llegar al mismo objetivo. En este marco, la RICYT está realizando un claro esfuerzo para la obtención y normalización de indicadores que permitan la medición y el análisis de la ciencia en Iberoamérica. En esta línea, el grupo SCImago,[1] formado por investigadores de las universidades de Granada, Extremadura, Carlos III y Alcalá de Henares, presentó a finales del año 2006 su "Ranking de Instituciones de Investigación Iberoamericanas" (RI3).[2] Este ranking está elaborado a partir del número de publicaciones y el factor de impacto asociados disponibles en las bases de datos Thomson-ISI (1990-2005) para diez países de la región; se pretende que el ranking cubra el total de la base en el mediano plazo.

Por otro lado, el Laboratorio de Cibermetría[3] perteneciente al Centro de Información y Documentación Científica (CINDOC-CSIC) del Consejo Superior de Investigaciones Científicas de España elabora el "Ranking Mundial de Universidades en la Web" (RMUW) (Marcos, 2006). Éste es construido a partir de indicadores relacionados con el contenido web y la presencia en línea de material de carácter científico obtenidos por medio de diferentes motores de búsqueda (Aguillo et al., 2005 y 2006). El ranking comenzó su andadura en 2004 y es actualizado completamente cada seis meses.

La ventaja de los rankings es que pueden combinar datos procedentes de diferentes fuentes y características, los cuales, tratados correctamente, minimizan los errores sistemáticos. Además, resultan fáciles de interpretar y tienen un gran impacto tanto mediático como académico. En el caso del presente trabajo, la disponibilidad de dos clasificaciones tan similares en lo que se refiere a la cobertura institucional y geográfica, pero tan dispares en cuanto a los indicadores utilizados, permite la realización de este estudio de comparación con el objetivo de poder ilustrar las ventajas y limitaciones de ambos y de la metodología empleada en cada caso.

---

* Los autores son miembros del CINDOC-CSIC de España (correos electrónicos: mariofdez@cindoc.csic.es, isidro@cindoc.csic.es, jortega@cindoc.csic.es, bgranadino@cindoc.csic.es).

1 Véase http://www.scimago.es

2 Véase http://www.universia.es/portada/actualidad/noticia_actualidad.jsp?noticia=90459

3 Véase http://internetlab.cindoc.csic.es/

2. Metodología

Se ha procedido a la extracción de los datos publicados por el RI3 actualizados a fecha del 26 de abril de 2007 en el Portal Universia.[4] Para ello se seleccionó el indicador de "producción total" y se escogió un valor mínimo de cien documentos. De esta forma se obtuvieron los datos de un total de 763 instituciones de los siguientes países englobados dentro del parámetro "Iberoamérica": Argentina, Brasil, Chile, Colombia, Cuba, España, México, Perú, Portugal y Venezuela. Tomando los datos correspondientes al periodo 2000-2005 se procedió al reordenamiento del ranking sumando el valor de publicaciones parcial para cada año.

Para el caso del RMUW se seleccionaron los valores absolutos correspondientes a la última actualización de enero de 2007[5] para los diez países anteriormente indicados y se procedió de igual forma a su reordenamiento, obteniéndose así un total de 1.757 instituciones.

Posteriormente se realizó una comprobación manual de las primeras quinientas instituciones presentes en el RI3, con el fin de tratar de eliminar posibles errores de trascripción que impidieran la correcta comparación de las mismas entre ambos rankings. Salvo que se mencione lo contrario, se han utilizado para las diferentes comparaciones las primeras quinientas entidades de cada ranking.

3. Resultados

El análisis preliminar del RI3 permitió descubrir que existían algunas erratas e inconsistencias que se resolvieron adecuando las posiciones y normalizando los nombres para facilitar la comparación entre los dos listados. Así, entre las entradas encontradas en el campo "Institución" del RI3 apareció una bajo el título "Dirección particular". En otros casos, se identificaron diferentes entradas pertenecientes a entidades que financian estudios de investigación pero que no los desarrollan propiamente, tales como ministerios y consejerías de salud. Entre los errores comunes a ambos rankings se dan casos como el de la Fundación Augusto Pi y Sunyer que en 2004 pasó a ser el Instituto de Investigación Biomédica de Bellvitge, el cual, a su vez, se fusionó con la Fundación Instituto de Investigación Oncológica (IRO). La afiliación de estas entidades ha cambiado durante el periodo de tiempo incluido en ambos rankings y da problemas de asignación.

Si se compara el RI3 con el RMUW, se observa que de las quinientas primeras instituciones presentes en el RI3, 148 (29,6%) no se encuentran en el total de las 1.757 del RMUW. Al hacer la comparación contraria, son 217 entidades de las quinientas primeras presentes en el RMUW (43,4%) las que no se encuentran en el total de las 763 del RI3.

En cuanto a la distribución por países de las primeras quinientas entidades, se aprecia que éstas difieren (tabla 1 y figura 1) aunque no de forma significativa. Los países con un mayor tamaño se encuentran en las primeras posiciones de ambos rankings, aunque cabe destacar la mayor presencia en la web de las instituciones de Chile, Colombia y Perú.

4 Véase http://investigacion.universia.es/isi/isi.html

5 Véase http://www.webometrics.info/methodology_es.html

Figura 1. Número de instituciones por país

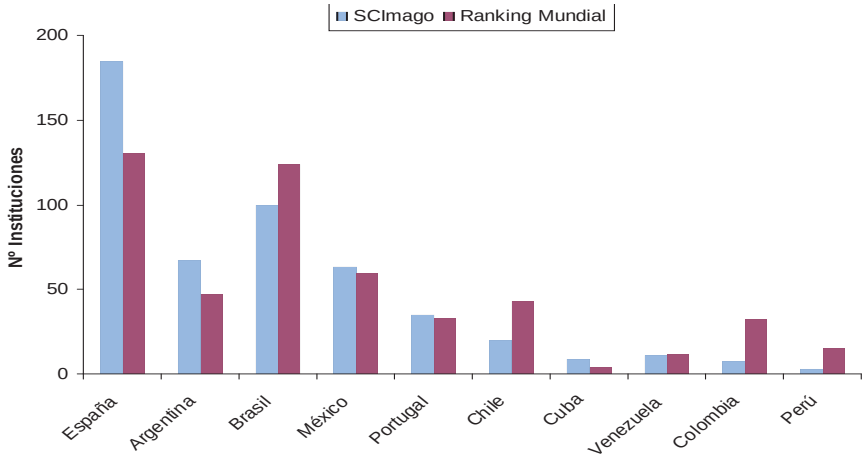


Tabla 1. Orden según nº instituciones/país

	RI3	RMUW
España	1	1
Brasil	2	2
Argentina	3	4
México	4	3
Portugal	5	6
Chile	6	5
Venezuela	7	9
Cuba	8	10
Colombia	9	7
Perú	10	8

Si se toma como referencia el total de instituciones de cada país según el RMUW, se le asigna un valor normalizado al 100% y se lo compara con el mismo valor observado según el RI3 (figura 2), se puede comprobar más fácilmente cómo las instituciones de los tres países mencionados anteriormente han potenciado claramente su presencia web pero necesitan mejorar en lo referente a la calidad de sus publicaciones. También se puede observar que el caso de Cuba es exactamente el contrario: su producción científica es alta pero su presencia web es muy baja.

En cuanto a la composición por tipo de institución de investigación (tabla 2), en el RI3 hay un menor número de universidades y, sin embargo, un gran número de entidades de tipo hospitalario y empresas privadas en comparación al RMUW. Según los datos del RI3, los centros en donde se debería concentrar en mayor medida la investigación científica ven disminuido su número en detrimento de otras entidades, como ser los hospitales.

Figura 2. Producción científica vs. presencia web

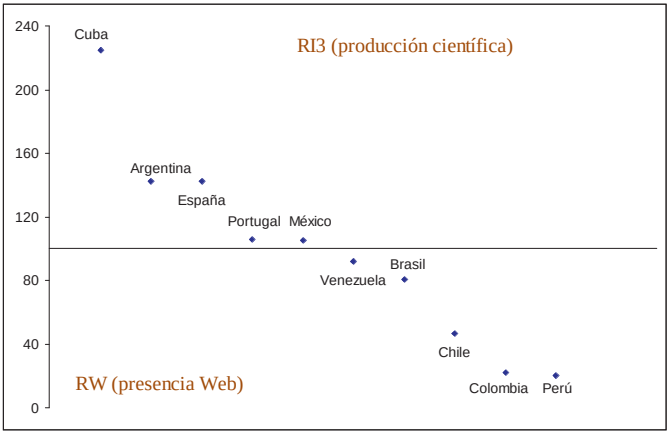


Tabla 2. Tipos de institución

	RI3	RMUW
Universidades	211 (42%)	344 (69%)
Centros de investigación y otros	177 (35%)	153 (31%)
Hospitales	104 (21%)	2 (0%)
Empresas privadas	8 (2%)	1 (0%)

Cuando se utiliza como referencia el RI3 y se lo enfrenta al RMUW se halla que, aunque varias instituciones líderes coinciden en los primeros puestos de ambos rankings, hay otras cuya presencia en la web es muy discreta (tabla 3). De hecho, entre las quinientas primeras clasificadas según producción científica, 196 no logran situarse en el ranking web por debajo de la posición 1.000. La mayoría de estas instituciones corresponden a hospitales y centros de investigación biomédica, lo cual confirma el sesgo de las bases de datos de Thomson Scientific. Por otro lado, ello muestra la brecha que actualmente existe por parte de renombradas instituciones a la hora de darse a conocer a través de Internet, a pesar de que su producción científica sea elevada.

Cabe destacar la diferencia presente en cuanto a la posición del Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina) y la del Hospital Clínico y Provincial de Barcelona (España). Ambas, a pesar de tener una producción científica elevada, mantienen una posición en la web muy discreta. El caso del Consejo Superior de Investigaciones Científicas se explica por el hecho de que varios de sus centros más productivos son de carácter mixto con universidades y sus sedes web se hospedan en el dominio de éstas.

Tabla 3. RI3 frente a RMUW

Institución	RI3	RMUW
Consejo Superior de Investigaciones Científicas	1	7
Universidade de São Paulo	2	1
Universidad Nacional Autónoma de México	3	2
Universidad de Barcelona	4	5
Consejo Nacional de Investigaciones Científicas y Técnicas	5	221
Universidad Complutense de Madrid	6	3
Universidade Estadual de Campinas	7	4
Universidad de Buenos Aires	8	20
Universidade Federal do Rio de Janeiro	9	21
Universidad Autónoma de Barcelona	10	9
Universidad de Valencia	11	11
Universidade Estadual Paulista Julio de Mesquita Filho	12	48
Universidad Autónoma de Madrid	13	24
Universidad de Santiago de Compostela	14	43
Universidad de Chile	15	8
Universidad de Granada	16	12
Hospital Clínico y Provincial de Barcelona	17	1184
Universidade do Porto	18	6
Universidade Federal do Rio Grande do Sul	19	23
Universidade Técnica de Lisboa	20	16

La visión complementaria que se obtiene cuando se hace la comparación partiendo del RMUW (tabla 4) muestra una mayor correlación entre las instituciones en ambas clasificaciones. Las diferencias no son tan grandes y el listado es más homogéneo, ya que consta principalmente de universidades generalistas de tamaño medio o grande. Ello es consistente con datos obtenidos anteriormente por el Laboratorio de Cibermetría (Aguillo, 2005; Aguillo et al., 2006). En general, esta segunda lista muestra un mayor número de universidades tecnológicas en las primeras posiciones, lo que es consistente con su baja producción de artículos para revistas científicas, aunque claramente se preocupan por mantener una buena presencia en la web.

Tabla 4. RMUW frente a RI3

Institución	RMUW	RI3
Universidade de São Paulo	1	2
Universidad Nacional Autónoma de México	2	3
Universidad Complutense de Madrid	3	5
Universidade Estadual de Campinas	4	6
Universidad de Barcelona	5	4
Universidade do Porto	6	21
Consejo Superior de Investigaciones Científicas	7	1
Universidad de Chile	8	15
Universidad de Sevilla	9	23
Universidad Autónoma de Barcelona	9	10
Universidad de Valencia	11	11
Universidad de Granada	12	16
Universidade Federal de Santa Catarina	13	50
Universidad Politécnica de Madrid	14	33
Universidad Politécnica de Valencia	15	34
Universidade Técnica de Lisboa	16	19
Universidad Politécnica de Cataluña	17	24
Tecnológico de Monterrey	18	196
Universidad de Alicante	19	53
Universidad de Buenos Aires	20	7

4. Discusión

En la actualidad, los rankings se encuentran muy extendidos como herramienta de comparación y evaluación científica. Diferentes grupos y estamentos elaboran sus propios rankings[6] utilizando distintos indicadores, siendo el factor de impacto que proporciona Thomson-ISI uno de los más incluidos. El ranking realizado por el grupo SCImago se basa exclusivamente en este valor y en el número de publicaciones recogidos en las revistas indexadas en las bases de datos Thomson-ISI (Grupo SCImago, 2007).

Entre los rankings basados en datos web se destacan el TrafficRank de Alexa[7] y los que se elaboran a partir del PageRank de Google (Aguillo, Granadino y Llamas, 2005). El Laboratorio de Cibermetría mantiene el Ranking Mundial de Universidades en la Web, que está construido a partir de varios indicadores que no sólo miden el rendimiento de las instituciones, sino además su presencia en la web (Aguillo et al., 2005; Marcos, 2006).

El RI3 analiza el estatus científico de las instituciones de investigación de diez países pertenecientes a la región iberoamericana. Este hecho ha permitido hacer una com-

6 Véanse, por caso, los sitios de ARWU (http://ed.sjtu.edu.cn/ranking.htm), THES (http://www.thes.co.uk/worldrankings), el European Report on Science & Technology Indicators (http://cordis.europa.eu/indicators/third_report.htm), y los Essential Science Indicators (http://www.esi-topics.com), entre otros.

7 Véase http://www.alexa.com

paración entre rankings con el objetivo de estudiar las ventajas e inconvenientes de dos metodologías tan distintas aplicadas al mismo problema. La disparidad de indicadores utilizados en ambos rankings hacía esperar que hubiese grandes diferencias en las posiciones de las instituciones investigadoras. Se ha podido comprobar que, si bien la coincidencia entre ambas clasificaciones no es mucha, sí existe una buena concordancia, sobre todo en las primeras posiciones. Sin embargo, se demuestra un claro sesgo hacia las instituciones de investigación centradas en biomedicina, lo que hace que en el RI3 se encuentren más entidades de tipo hospitalario que en el RMUW. Otro detalle importante se encuentra en el hecho de que instituciones con una producción científica notable muestran una presencia web muy baja, como por ejemplo el Hospital Clínico y Provincial de Barcelona. La visibilidad de este tipo de centros es muy baja y no da una medida auténtica de la calidad de la institución investigadora.

Durante la recolección de datos del RI3 y su posterior análisis se ha podido comprobar cómo todavía es muy necesario recalcar la necesidad, a la hora de elaborar cualquier ranking que pretenda dar una justa valoración del objeto que evalúa, de comprobar la calidad de los datos que se obtienen de forma automática. Así, se ha podido encontrar en el RI3 una entrada de institución como “Dirección particular” o valores que designan entidades que financian los estudios realizados pero que no ejercen una actividad investigadora, tales como ministerios, secretarías de estado, consejerías, etc. La elaboración de un ranking exige que los datos obtenidos sean comprobados de manera exhaustiva para que identifiquen de forma inequívoca las instituciones que son analizadas (Van Raan, 2005; Zitt, 2006). En el caso de rankings basados en el factor de impacto, esto a veces se ve impedido por la falta de un estándar por parte de los propios autores de artículos a la hora de designar sus instituciones de investigación. En otros casos, la unión de centros de investigación dificulta la tarea de asignar correctamente la labor de investigación desarrollada al nombre actual de la entidad correspondiente.

En el caso del RMUW también existen diferentes problemas que es necesario solventar. Los cambios de dominio, la existencia de varios dominios idénticos pero con terminación diferente (.edu, .es) que designan a la misma institución o los sesgos propios de los motores de búsqueda suelen ser el problema más común.

El ejercicio de comparación de ambos rankings muestra que, a pesar de la diferencia de abordaje del desarrollo de los mismos, las diferencias no son tan amplias y las posiciones de las instituciones se encuentran más cerca de lo que a priori se podía esperar.

Existe la necesidad de mejorar y aumentar la calidad de los indicadores utilizados para que sean capaces de reflejar fielmente el potencial científico de las instituciones. Por otra parte, los científicos y las instituciones que los albergan deberían mejorar la estandarización en cuanto a lo que a la afiliación del trabajo se refiere. Esto permitiría que la labor de limpieza de los datos obtenidos para elaborar los rankings fuera más eficaz.

Por último, cabe destacar que el incremento de la presencia en la Web puede ser un medio eficaz para aumentar la exposición del conocimiento científico y mejorar la imagen de excelencia de las instituciones de investigación.

Bibliografía

AGUILLO, I. F. (2005): “Indicadores de contenidos para la web académica iberoamericana”, *BiD: textos universitaris de biblioteconomia i documentació*, desembre, núm. 15, disponible en formato electrónico en: http://www2.ub.edu/bid/consulta_articulos.php?fichero=15aguil2.htm

AGUILLO, I. F., GRANADINO, B., ORTEGA, J. L. y PRIETO, J. A. (2005): “What the Internet says about Science”, *The Scientist*, 19 (14)10, Jul. 18.

AGUILLO, I. F., GRANADINO, B. y LLAMAS, G. (2005): “Posicionamiento en el Web del sector académico iberoamericano”, *Interciencia*, 30(12), pp. 1-5.

AGUILLO, I. F., GRANADINO, B., ORTEGA, J. L. y PRIETO, J. A. (2006): “Scientific research activity and communication measured with cybermetric indicators”, *JASIST*, 57(10), pp. 1296-1302.

GRUPO SCIMAGO (2007): “Ranking de instituciones de investigación iberoamericanas (RI3)”, *El Profesional de la Información*, 16(3), pp. 258-260.

MARCOS, M. C. (2006): “Webometrics pone orden en las universidades”, *El Profesional de la Información*, 15(3), pp. 231-236.

VAN RAAN, A. F. J. (2005): “Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods”, *Scientometrics*, 62(1), pp. 133-143.

ZITT, M. (2006): “Scientometric indicators: a few challenges: data mine-clearing; knowledge flows measurements; diversity issues”, en *Proceedings International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting*, Nancy (France), http://eprints.rclis.org/archive/00006306/

STANALYST-SciELO: modelo y uso para la vigilancia científica

Xavier Polanco, Mario Albornoz, Abel Packer, Anna Maria Prat,
Dominique Besagni, Claire François, Ivana Roche,
Rodolfo Barrere, Lautaro Matas,
Fabio Batalha Cunha dos Santos y Jorge Walters*

1. Introducción

El objetivo de este trabajo es presentar las características del sistema STANALYST - SciELO[1] para el análisis de la información científica y técnica y mostrar por medio de un estudio de caso cómo se puede analizar con esta herramienta el contenido de las bases SciELO. STANALYST - SciELO se encuentra actualmente en la fase conocida como "alfa test", es decir, la etapa de prueba que precede su apertura experimental a los usuarios, conocida como "beta test".

STANALYST - SciELO es el resultado de un proyecto multilateral que se realizó entre 2005 y 2006 y cuya meta fue compatibilizar la plataforma STANALYST y las bases SciELO. El proyecto tuvo una duración de un año (entre agosto de 2005 y septiembre de 2006) y contó con el apoyo del Ministère des Affaires Etrangères de Francia. En él participaron el Institut de l'Information Scientifique et Technique del Centre National de la Recherche Scientifique de Francia (INIST/CNRS), el Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde de la Organización Panamericana de la Salud / Organización Mundial de la Salud (BIREME/OPS/OMS), el Centro Argentino de Información Científica y Tecnológica del Consejo Nacional de

* Xavier Polanco (xavier.polanco@inist.fr, xavier.polanco@lip6.fr) es investigador del INIST / CNRS de Francia y de la Université Pierre et Marie Curie de París. Mario Albornoz (albornoz@ricyt.org) es coordinador de la RICYT y director del CAICYT / CONICET de Argentina. Abel Packer (packerab@bireme.ops-oms.org) es director de BIREME / OPS / OMS, Brasil. Anna Maria Prat (amprat@conicyt.cl) es miembro del CONICYT de Chile. Dominique Besagni (dominique.besagni@inist.fr), Claire François (claire.francois@inist.fr) e Ivana Roche (ivana.roche@inist.fr) son investigadoras del INIST / CNRS. Rodolfo Barrere (rbarrere@ricyt.org) es miembro de la secretaría técnica de la RICYT. Lautaro Matas (lmatas@caicyt.gov.ar) se desempeña en el CAICYT / CONICET de Argentina. Fabio Batalha Cunha dos Santos (santosfa@bireme.ops-oms.org) es miembro de BIREME / OPS / OMS, Brasil. Jorge Walters (jwalters@pbct.cl) es miembro del CONICYT de Chile.

1 La primera referencia al proyecto STANALYST-SciELO se hizo en el Segundo Seminario Internacional sobre Indicadores de Ciencia, Tecnología e Innovación organizado por Kawax, el Observatorio Chileno de Ciencia, Tecnología e Innovación del Programa Bicentenario, entre el 16 y el 18 de enero de 2006 en Santiago de Chile; la presentación de Xavier Polanco al respecto fue "STANALYST. Una aplicación para nuevos estudios bibliométricos sobre bases de datos locales". Otra referencia más detallada del proyecto es Polanco (2006), en las Jornadas Internacionales El espacio público de las ciencias sociales y humanas, organizadas por el Centro Franco-Argentino de Altos Estudios de la Universidad de Buenos Aires, los días 20 y 21 de noviembre de 2006.

Investigaciones Científicas y Técnicas de Argentina (CAICYT-CONICET), la Comisión Nacional de Investigación Científica y Tecnológica de Chile (CONICYT) y la RICYT.

En este trabajo se presentarán los modelos respectivos de STANALYST y SciELO, el instrumento o tecnología que constituye STANALYST y las bases SciELO. Siendo al origen una tecnología de la información creada por el INIST/CNRS, STANALYST permitía exclusivamente el acceso a las base multidisciplinarias FRANCIS y PASCAL, de propiedad del INIST/CNRS. Pero lo que aquí interesa es subrayar la relación STANALYST - SciELO. Una aplicación ilustrará el tipo de análisis que se puede hacer con STANALYST a partir de las bases SciELO. El estudio se limita a las bases SciELO Argentina, Brasil y Chile de 2002 a 2006. De ellas se obtuvieron 724 publicaciones sobre el cáncer. Se escogió el cáncer por tratarse de un tema de investigación con producción suficiente para los análisis, transversal a todas las bases, y porque posee importancia biomédica y en salud pública. El presente se trata de un ejemplo de lo que se llama "vigilancia científica" o "inteligencia estratégica en la ciencia", que puede extenderse a otros temas y áreas científicas. La ambición es aquí ilustrar el uso que puede hacerse de STANALYST con una intención de vigilancia científica o inteligencia. A continuación se precisará el sentido de las nociones de "análisis" e "inteligencia" de la información científica y técnica.

1.1. Análisis de la información

El análisis de la información puede ser definido como el uso de (a) técnicas estadisticas, (b) procesamientos automáticos del lenguaje, es decir del languaje escrito en el que se expresan los conocimientos cientificos, (c) técnicas de clasificación automática y de cartografía, en otras palabras elaboracion de mapas (Polanco 1997a, 1997b). Es esta una definicion operacional en el sentido que corresponde al análisis asistido por computador. En general, se entiende por análisis de la información la fase de interpretación que el usuario realiza de una manera directa y manual de los datos colectados. Los límites de este tipo de análisis son obvios cuando la tarea consiste en tratar una masa significativa de datos. Cuando este es el caso, el analista necesitará apoyarse sobre tecnologías de la información especialmente concebidas para asistir la tarea de análisis. STANALYST es un ejemplo de este tipo de tecnología. El objetivo es poder detectar los centros de interés en un campo científico dado (en otras palabras los temas o los tópicos de conocimiento contenidos en los datos bibliograficos o textuales) y los diferentes actores y elementos que componen esta información (es decir los autores, los artículos, las revistas, los laboratorios, los países), así como disponer de una representacion visual como un mapa para apreciar la posición relativa de los conocimientos y actores.

1.2. Tecnologías de la inteligencia

Una cosa es producir tecnologías inteligentes y otra producir tecnologías para la inteligencia. A fin de esbozar una idea de lo que aquí se llama "inteligencia", es posible citar la siguiente definición: "La inteligencia social es la capacidad de reunir y aplicar la información para asegurar la viabilidad o el éxito en un contexto particular" (Cronin y Davenport, 1993). Ahora bien, si se llama inteligencia a las operaciones de análisis, evaluación y decisión relativas a la definición de estrategias, entonces se puede llamar tecnologías de la inteligencia a las tecnologías de la información al servicio de tales operaciones. Las tecnologías de la inteligencia son instrumentos informáticos complejos, en este caso particular dedicadas a ayudar a la inteligencia en el terreno de la información científica y técnica.

En la segunda sección de este trabajo se presenta el modelo y el tipo de instrumento que STANALYST representa y los indicadores que produce. La tercera sección presenta el modelo SciELO y las características de las bases que responden a este modelo. La sección cuarta está dedicada al análisis de la aplicación STANALYST - SciELO para el caso del cáncer. Finalmente, la quinta sección presenta conclusiones.

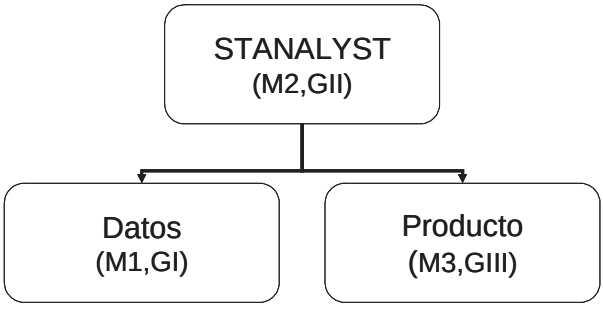
2. STANALYST: modelo, instrumento, indicadores

2.1. Modelo

Según Polanco (1997a), el analista debe ocuparse no del conocimiento en acción como competencia de los individuos (sujetos del conocimiento), sino del conocimiento producido por ellos y almacenado en las bases de datos, con el objetivo de extraer los conocimientos adaptados o útiles para la toma de decisiones, la definición de estrategias y la evaluación del estado de la ciencia y la tecnología en un momento dado. Esta proposición tiene su base epistemológica en las teorías de los "tres mundos" de Popper (1972) y de las "tres generalidades" de Althusser (1965). Las dos formulaciones, por lo demás, se asemejan fuertemente. La teoría de Popper distingue tres "mundos", a saber: el físico (mundo 1), el de la conciencia o de los estados psíquicos o de los conocimientos en el sujeto (mundo 2) y el del contenido intelectual de libros y documentos (mundo 3), llamado "conocimiento objetivo", es decir un conocimiento sin sujeto cognoscente. Un objeto físico, tal como un artículo científico, pertenece al mundo 1 y contiene conocimientos que pertenecen al mundo 3. Según el esquema de Althusser de las tres generalidades, se llama generalidades I (G I) los objetos (= materias primas) de estudio de una ciencia, generalidades II (G II) a los medios de trabajo teórico y generalidades III (G III) a los conocimientos producidos como resultado del trabajo de G II sobre G I. Por ello, en el caso de este trabajo, las fuentes de información serían las G I, que el analista transforma en conocimientos G III, mediante la aplicación de una tecnología de la información (G II), tal como STANALYST. Por su parte, Brooks (1980) retoma y aplica la teoría de los tres mundos de Popper en la ciencia de la información, y sobre esta base formula su "ecuación fundamental", en la cual una estructura dada del conocimiento es modificada por el aporte de nueva o más información, y al mismo tiempo establece claramente que "documento y conocimiento no son entidades idénticas".

En este marco, el objetivo al utilizar un instrumento como STANALYST es pasar del análisis de documentos (nivel 1 o bibliográfico), de autores o investigadores (nivel 2 o sociológico), al estudio del conocimiento que ellos producen y difunden a través de sus escritos (nivel 3 o del conocimiento objetivo). Se insiste, asimismo, en la asimetría entre documento y conocimiento como entidades distintas: una cosa es contar documentos (bibliometría) y otra extraer y representar conocimientos a partir de los documentos (datos textuales o bibliográficos), como se lo propone STANALYST.

Figura 1. Modelo STANALYST



En la figura 1 de ve que en el lugar del sujeto o agente cognoscente, al cual hacen referencias tanto el mundo 2 de Popper como las generalidades II de Althusser, aparece la tecnología de la inteligencia que STANALYST representa, la cual, aplicada sobre una materia prima dada (datos), produce un cierto número de resultados (información elaborada), como se verá en la cuarta sección de este trabajo. La aplicación de STANALYST sobre los datos produce resultados que corresponden al producto de cada uno de los módulos que constituyen STANALYST, como se expondrá en la sección siguiente.

2.2. Instrumento

En tanto que tecnología de la información, STANALYST constituye un instrumento que de acuerdo con lo que se ha dicho antes actúa sobre una "materia prima" (los datos) a fin de producir una información sobre el estado del "conocimiento objetivo". Es así que se tienen estos tres elementos: la materia prima, un instrumento y el producto del trabajo del instrumento. Aquí se detallan las características técnicas del instrumento -es decir, STANALYST- y luego los indicadores que produce.

2.2.1. Características técnicas

Las principales características técnicas de STANALYST pueden ser resumidas en las siguientes:

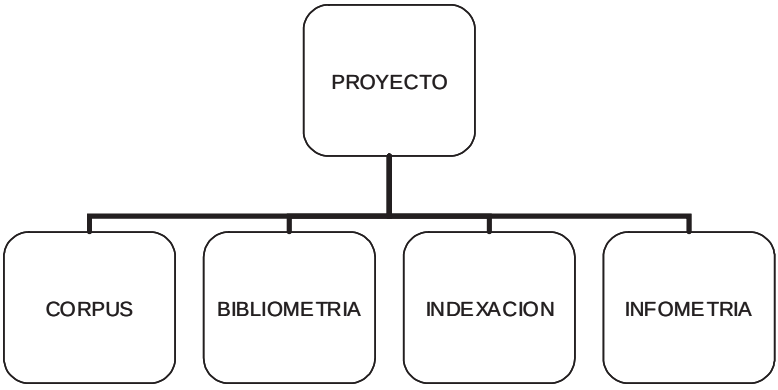
- uso de un servidor HTTP Apache,
- programas CGI desarrollados en lenguaje Perl y shell UNIX,
- visualización por medio de un navegador HTTP (Firefox o Internet Explorer),
- datos y ficheros en formato SGML manejados por una biblioteca de funciones (denominada ILIB por Inist Library) y almacenados en un sistema de ficheros llamado HFD (Hierarchical File Data). Por el momento, STANALYST funciona con el formato interno del INIST/CNRS, utilizado para las bases PASCAL y FRANCIS. Los datos de las bases SciELO son actualmente convertidos a este formato para que puedan ser procesados por STANALYST.

2.2.2. Arquitectura de STANALYST

La arquitectura de STANALYST es modular, y se compone de un conjunto de módulos autónomos: "Proyecto", "Corpus", "Bibliometría", "Indexación", "Infometría". Estos módulos existen en la forma de scripts (programas no compilados) en lenguaje shell Unix, y algunos de ellos son acompañados por paginas HTML generadas dinámicamente mediante scripts Perl. La integración de los módulos de acuerdo con una interfaz gráfica común, accesible desde un navegador HTTP, constituye el servidor STA-

NALYST. La interfaz gráfica está solamente en francés, aunque en un futuro próximo presentará opciones en español, portugués e inglés.

Figura 2. Los módulos de STANALYST



El acceso a STANALYST necesita la identificación del usuario (nombre y contraseña). Cada módulo ejecuta una tarea bien precisa, funcionando independientemente los unos de los otros. Los módulos utilizan un conjunto de repertorios de trabajo que contienen los programas, los scripts y los parámetros necesarios a su ejecución, así como el conjunto de los proyectos creados por los usuarios. Con el objetivo de manejar la transferencia de datos de un módulo al otro, pero también de ampliar el concepto del repertorio privado a todos los módulos, la organización de los repertorios en el servidor obedece a una arborescencia. El conjunto de los tratamientos se hace dentro de la estructura arborescente del mismo proyecto, que refleja así la secuencia lógica del proceso de análisis de la información:

- El módulo “Proyecto” realiza la gestión y creación de proyectos.
- El módulo “Corpus” maneja la creación del corpus a partir de las bases de datos, de acuerdo con la interrogación dictada por el usuario. Puede exportarse directamente a los módulos “Indexación” e “Infometría”.
- El módulo “Bibliometría” permite al usuario definir y producir los análisis estadísticos que corresponden a los indicadores bibliométricos tradicionales.
- El módulo “Indexación” permite revisar la indización preexistente o bien realizar una indización automática, denominada ILC (Royauté, 1999; Polanco et al., 1995), que toma en cuenta la variación en los textos de los términos de un vocabulario terminológico (recursos terminológicos integrados en el sistema). ILC trata solamente las lenguas inglesa y francesa. El módulo permite también la visualización y la validación de los resultados por parte del usuario.
- El módulo “Infometría” asegura la ejecución de métodos de clasificación automática no supervisada, llamados “Neurodoc” y “Sdoc” (Grivel y François, 1995), así como la visualización de los resultados como listas y bajo la forma de mapas.

Cada módulo posee una página de entrada que permite crear, suprimir o elegir un repertorio. Luego de elegir un repertorio el usuario entra en la página de acceso, que constituye la página principal del módulo. En ella se encuentran:

- las informaciones relativas al repertorio,

- las acciones de administración (habilitaciones, descarga de los corpus, de los resultados bibliométricos, de las listas de términos, de los resultados de las clasificaciones,…) y
- las acciones específicas del módulo (interrogaciones, análisis estadísticos, perfil de indización, perfil de clasificación,…), que han sido parametradas por el usuario.

En cada módulo, el usuario puede autorizar el acceso al repertorio a otros usuarios identificados por STANALYST. Son posibles dos niveles de autorización: a) la simple consulta, que permite solamente visualizar los resultados, y b) la colaboración, que ofrece a los usuarios los mismos derechos que los del propietario del repertorio, excepto la administración de las habilitaciones. Por otra parte, en cada módulo el usuario puede descargar los resultados en el disco duro de su computador.

2.2.3. Indicadores

Diferentes clases de indicadores se producen de acuerdo con cada módulo. En el módulo "Bibliometría", los indicadores de actividad que son producidos utilizan los distintos elementos bibliográficos, principalmente el tipo de documento, la fecha de publicación, la lengua del documento, el país editor, los autores y sus afiliaciones, las revistas y las palabras claves. Las distribuciones correspondientes son visualizadas bajo la forma de tablas y/o gráficos (curvas, histogramas u otras figuras). Sólo el gráfico de la ley de Bradford sobre las revistas puede visualizarse directamente. Los gráficos de las otras leyes bibliométricas (Lotka para los autores, Zipf para las palabras claves) necesitan que los resultados numéricos sean descargados hacia una herramienta de análisis estadístico para obtener el gráfico correspondiente.

En el módulo "Indexación", los indicadores se restringen a la frecuencia de las palabras-claves y sus variaciones. No todas la palabras claves comportan (sea por su frecuencia elevada, demasiado genérica, sea por su significado menor o nulo) un interés igual en tanto que indicadores de conceptos a retener para la clasificación; es así como el usuario puede excluirlos, o bien simplemente eliminarlos cuando a su juicio se trata de "ruido".

En el módulo "Infometría", las clases constituyen indicadores de tópicos o temas de investigación contenidos en los datos, mientras que los mapas indican la posición relativa de las clases en el conjunto de clases. Estos indicadores son comunes a los dos métodos de clasificación del módulo: NEURODOC y SDOC (Grivel y François, 1995). Actualmente, sólo uno de los métodos (SDOC) puede ser utilizado en la versión de STANALYST compatible con las bases SciELO. Se espera que un futuro próximo NEURODOC también lo sea (restan completar detalles de programación).

3. SciELO: modelo y bases de datos

SciELO - Scientific Electronic Library Online (Biblioteca Científica Electrónica en Línea) es un modelo para la publicación electrónica cooperativa de revistas científicas en Internet. Especialmente desarrollado para responder a las necesidades de la comunicación científica en los países en desarrollo y particularmente de América Latina y el Caribe, el modelo proporciona una solución eficiente para asegurar la visibilidad y el acceso universal a su literatura científica, contribuyendo para la superación del fenómeno conocido como "ciencia perdida". Además, el modelo SciELO contiene procedimientos integrados para la medida del uso y del impacto de las revistas científicas.

El modelo SciELO es el producto de la cooperación entre la Fundação de Apoio à Pesquisa do Estado de São Paulo (FAPESP), el Centro Latinoamericano y del Caribe

de Información en Ciencias de la Salud (BIREME) e instituciones nacionales e internacionales relacionadas con la comunicación científica y los editores científicos. A partir de 2002, el proyecto cuenta con el apoyo del Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) de Brasil.

SciELO contiene tres componentes principales:

- La metodología SciELO, que permite la publicación electrónica de ediciones completas de las revistas científicas, la organización de bases de datos bibliográficas y de textos completos, la recuperación de textos por su contenido, la preservación de archivos electrónicos y la producción de indicadores estadísticos de uso e impacto de la literatura científica. La metodología también incluye criterios de evaluación de revistas, basados en los estándares internacionales de comunicación científica. Los textos completos son enriquecidos dinámicamente con enlaces de hipertexto con bases de datos nacionales e internacionales, tales como LILACS y MEDLINE.
- La aplicación de la metodología SciELO en la operación de los sitios web de colecciones de revistas electrónicas. El modelo SciELO favorece la operación de sitios nacionales y también de sitios temáticos.
- El desarrollo de alianzas entre los actores nacionales e internacionales de la comunicación científica –autores, editores, instituciones científicas y tecnológicas, agencias de apoyo, universidades, bibliotecas, centros de información científica y tecnológica, etc.- con el objetivo de diseminar, perfeccionar y mantener el modelo SciELO. El funcionamiento de la red SciELO se basa fuertemente en las infraestructuras nacionales, lo que contribuye a garantizar su futura sostenibilidad.

Junto con el desarrollo de la biblioteca electrónica de acceso abierto, se ha establecido una base complementaria de datos cientométricos / bibliométricos, que permite recuperar datos de citas de más de 40.000 artículos. La solidez que esta base de datos ha alcanzado, como Meneghini et al. (2006) lo muestran, permite actualmente hacer estudios importantes que antes eran posibles solamente utilizando la base Science Citation Index (SCI) del Institute for Scientific Information (ISI). A dicha posibilidad viene ahora a agregarse el hecho de que se puede utilizar igualmente STANALYST como instrumento de análisis cientométrico. Por el momento sólo las bases de Argentina, Brasil y Chile están disponibles vía STANALYST, pero el propósito es que toda la base SciELO –es decir, conforme con el modelo SciELO- pueda ser integrada en STANALYST.

4. Un ejemplo de aplicación

Lo que aquí se presentará es sólo un ejemplo, el cual, como se ha dicho en la introducción, se puede extender a cualquier problema científico presente en las bases de datos. En el caso presentado no se trata tanto de analizar el contenido de la información acerca del cáncer desde un punto de vista oncológico, sino más bien de mostrar el uso de STANALYST sobre las bases SciELO y los resultados que se obtienen en la perspectiva del análisis de la información (conforme a lo que se ha dicho en las secciones 1.1 y 1.2).

4.1. Estudio de caso

El estudio se limita a las bases SciELO Argentina, Brasil y Chile de 2002 a 2006. Como se ha adelantado en la introducción, se obtuvieron de ellas 724 publicaciones sobre el cáncer. Las tablas 2 y 3 (ver sección 4.4.1) muestran las once clases generadas a partir de los 724 datos. Recordemos que las clases indican temas de investi-

gación, agrupando autores, publicaciones y revistas; la densidad de una clase y su centralidad son valores que permiten apreciar a la vez la cohesión interna de la clase y su importancia relativa en el conjunto. Estos dos valores (indicadores) constituyen las coordenadas de los mapas de la sección 5. Nótese que cada clase es un grafo que STANALYST permite también visualizar (como puede constatarse más abajo, en la figura 6).

Es posible distinguir dos niveles de interpretación de los resultados: el primero, sobre la base del conocimiento y la experiencia de los métodos aplicados, así como de los indicadores cuantitativos y cualitativos; el segundo, la interpretación del contenido científico de los resultados, la significación que ellos poseen desde el punto de vista de la ciencia o la tecnología a la cual los datos hacen referencia. Este segundo nivel de interpretación es competencia de los expertos e investigadores, es decir, de quienes poseen los conocimientos, la experiencia y la práctica del campo científico considerado. El primer nivel, en cambio, es responsabilidad del analista de la información, en base al conocimiento y la experiencia de los métodos aplicados y los indicadores empleados.

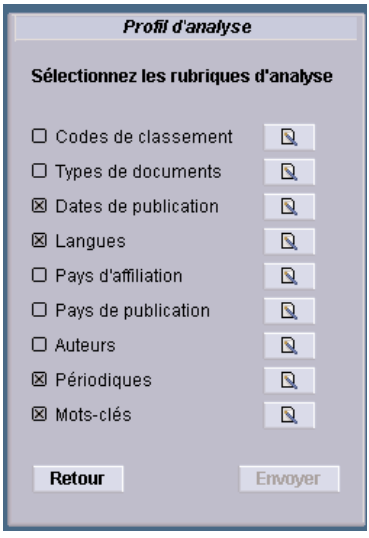
4.2. Bibliometría

Se llama bibliometria a la aplicación de técnicas estadísticas sobre datos bibliográficos. El módulo “Bibliometría” genera una información cuantitativa sobre la frecuencia y distribución de los datos (como se ha indicado en la sección 2.2.2), brindando así al analista información estadística descriptiva acerca de los datos del corpus. Para obtener estas informaciones, el usuario debe completar un formulario determinando lo que le interesa. La figura 3 muestra el formulario que STANALYST propone al usuario, compuesto por las siguientes rúbricas: tipo de documento (artículos, informes, tesis y libros, si ellos existen en las bases), fecha de publicación, país de publicación, lengua, autores, país de afiliación, revistas y palabras claves.

Este módulo produce la materia estadística de base como para observar la distribución hiperbólica que las leyes bibliométricas enuncian. Por otra parte, la información estadística generada por el módulo “Bibliometría” constituye la primera fase del análisis del corpus, análisis estadístico que se completa con el paso al análisis multidimensional, del módulo “Infometría”, mediante la aplicación de técnicas de clasificación automática bien conocidas en el análisis de datos.

El 2% del corpus, es decir trece datos, no presenta ninguna indexación. Los 711 datos restantes están indexados por 2.015 palabras claves, de las cuales 1.698 son de frecuencia 1, es decir el 84%, hecho que sugiere que la distribución hiperbólica de las palabras claves, de acuerdo con la ley de Zipf, es sumamente fuerte. Por otra parte, sólo las palabras claves con una frecuencia $\geq$ 2 se prestan para la clasificación en el módulo “Infometría”. El número medio de palabras claves por documento es 4, éste es un índice del número medio de conceptos significados por las palabras claves en los documentos –en este caso aparece como bastante bajo. Esta información importa porque las características de la indización tendrán consecuencias en los resultados de la clasificación.

Figura 3. Perfil de análisis bibliométrico



4.3. Indexación

Se puede considerar a esta fase como la del "análisis terminológico" de la lengua científica en cuestión –el cual puede constituir una finalidad en sí misma-, o bien como la fase previa y necesaria para la clasificación automática que supone la existencia de una matriz "datos × palabras claves o términos", ya que contar con datos indexados es una condición necesaria para su clasificación (en el módulo "Infometría"). Por otra parte, el módulo "Indexación" permite analizar la terminología científica caracterizando el contenido conceptual de los documentos. Las palabras claves juegan el rol de representar lingüísticamente conceptos científicos.

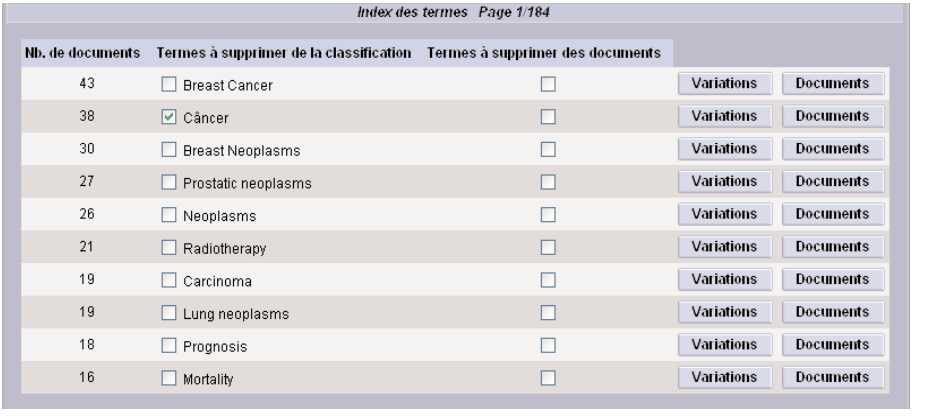
La figura 3 proporciona una muestra del vocabulario sobre cáncer presente en la indización. Se trata de los términos de frecuencia más elevada; el término "cáncer" fue excluido del vocabulario, dado que se utilizó en la indagación de las bases SciELO y, por lo tanto, no aporta ninguna información específica para el análisis temático que significa la clasificación automática. Como se ha dicho en la sección anterior, la distribución de las 2.015 palabras claves según su frecuencia obedece a la ley de Zipf y solamente un 16% servirá para representar los documentos en la clasificación que supone estadísticamente la frecuencia 2 como mínimo, puesto que (como se verá en la sección 4.4) ella está fundada en la co-ocurrencia de las palabras claves.

La indización forma parte del análisis documental y su objetivo es representar y describir el contenido de los documentos, mediante conceptos principales incluidos en ellos (palabras claves) o vocabularios controlados (descriptores, términos o encabezamientos de materia), con el fin de guiar al usuario en la recuperación de los documentos que necesita.[2] Ahora bien, cuando se trata del análisis de la información y no de la búsqueda de información documental, como es el caso, la indización cambia de rol, porque aquí su uso es distinto: la información ya ha sido encontrada y de lo que

2 Fuente (12/05/2007): Marta Godoy Velasco, Universidad Carlos III de Madrid, Sistemas Avanzados de Recuperación de Información: http://www.galeon.com /indizacion/indizacion.html. Véase también http://fr.wikipedia.org/wiki/Indexation

se trata es de analizarla; aquí importa entonces su capacidad de representación de conceptos en significación y extensión (específicos, genéricos). La indización de que se dispone en el módulo "Indexación" se designa en minería de textos "saco de palabras" (*word bag*), a causa de su falta de organización de acuerdo con un criterio lógico o semántico. Se trata simplemente de una lista de términos (o palabras claves) sin otro orden que la frecuencia de aparición en la colección de documentos.

Figura 4. Control y selección del vocabulario de indexación



Cuando se dispone de la distinción entre términos genéricos y específicos (como sucede con el vocabulario de indización PASCAL), el módulo "Indexación" ofrece al usuario la posibilidad de elegir al momento de fijar los parámetros genérico y/o específico de la indización. Además, el usuario puede elegir entre la indización ya existente en los datos y la indización automática llamada ILC, la cual opera según un criterio lingüístico fundado en la noción de variación (proceso expuesto en detalle en Jacquemin et al., 2002; Daille et al., 2000; Royauté, 1999; Polanco et al., 1995). ILC es una plataforma de ingeniería lingüística capaz de procesar (extraer e indexar) textos en inglés y francés. Su extensión al español y al portugués supone incorporar y ajustar técnicas de ingeniería lingüística capaces de procesar estas lenguas. En la aplicación se usó la indización en inglés de los datos SciELO, que presenta la legitimidad científica de ser el producto de los propios autores (investigadores).

En esta fase el usuario tiene la responsabilidad de seleccionar los términos que serán empleados para la clasificación, o puede también suprimirlos definitivamente por considerarlos "ruido".

4.4. Clasificación

La clasificación automática que es propuesta en el módulo "Infometría" se denomina "no supervisada", porque se realiza sin ninguna información previa acerca de las clases a obtener –en la clasificación conocida como "supervisada", en cambio, se clasifica en función de una taxonomía preexistente, y el problema consiste en asignar los datos a las clases previamente definidas. La clasificación no supervisada constituye en el análisis de datos un método exploratorio, es decir que busca descubrir en los datos mismos una estructura de ellos en clases. En principio, las clases agrupan los datos en función de su proximidad o similitud.

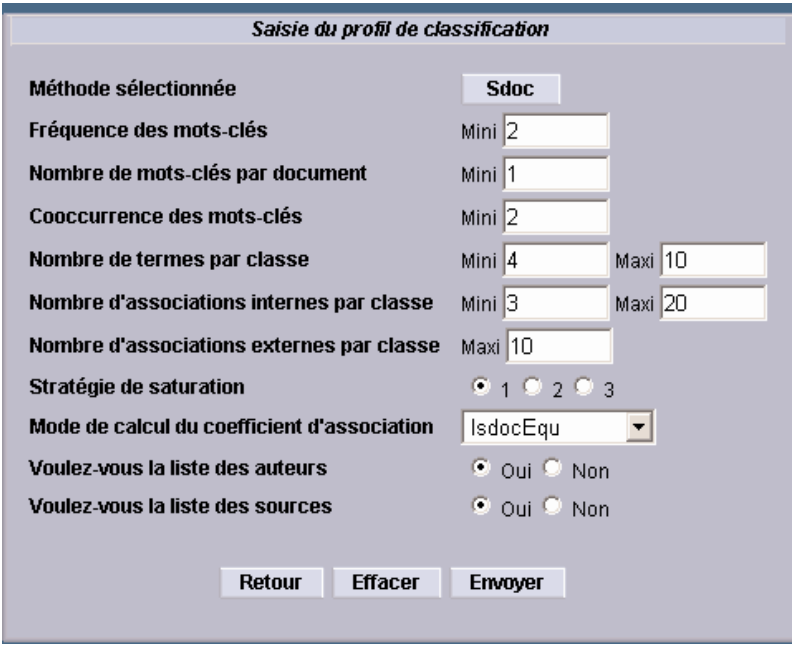
En el módulo "Infometría" se dispone de dos métodos de clasificación: NEURODOC

y SDOC (Grivel y François, 1995). Por el momento, en la versión de STANALYST compatibilizada con las bases SciELO solamente puede ser utilizado SDOC, razón por la cual aquí nos limitamos a considerar los resultados de este método.

Sin entrar en el detalle técnico del método y del algoritmo que SDOC ejecuta (ampliamente expuestos en Grivel et al., 1995; Polanco y Grivel, 1995), es posible decir que se trata de un método de clasificación jerárquica ascendente del simple enlace ("*single link*"), un método estándar descrito en los manuales consagrados al análisis multidimensional de datos. Con respecto a ello, el algoritmo SDOC presenta algunas particularidades que se encuentran descritas en las referencias aquí citadas. El método aplicado por SDOC es conocido, en el campo de la cienciometría, como el de las palabras asociadas, en inglés "*co-word analysis*" (Callon et al., 1993; Courtial, 1990).

Al escoger aplicar SDOC en el módulo "Infometría" como método de clasificación, el usuario tiene que definir los parámetros, o bien validar los que el sistema propone como parámetros predeterminados. La figura 5 presenta ese formulario.

Figura 5. Definición de los parámetros de clasificación



SDOC produce la clasificación de los datos en función de estos parámetros, cuyo resultado son las clases y los mapas. En la tabla 2 se verán los resultados de este proceso, que reúne los artículos sobre el cáncer (y con ellos autores y revistas) en once clases.

4.4.1. Clases

Cada clase señala un tema de investigación o, si se prefiere, un centro de interés científico. La información de la tabla 1 se refiere al tipo de estructura de la red de clases: los nombres o etiquetas de cada una de las once clases, luego su centralidad y densidad, así como el número de palabras claves internas (PC-int) y externas (PC-

ext) y de las asociaciones internas (As-int) o intra-clase y externas (As-ext) o inter-clases. Si se considera una clase determinada, sus asociaciones externas suponen la existencia de palabras claves internas (que constituyen la clase) y de palabras claves externas (pertenecientes a otras clases). Además de los dos tipos de asociación, se tiene lo que aquí se llama la "citación" entre clases para dar cuenta del sentido o dirección de la asociación inter-clase. Su número indica las relaciones que una clase recibe de las otras clases y que no es necesariamente igual al número de sus asociaciones externas (As-ext).

La centralidad y la densidad caracterizan las clases y definen sus propiedades estructurales. La centralidad es un indicador, como en el análisis de redes sociales (Degenne y Forsé, 2001; Wasserman y Faust, 1999), de la importancia o preeminencia de la clase en el conjunto, en la red. La centralidad es aquí definida por las asociaciones externas (As-ext) entre las clases: se trata de la llamada centralidad de grado. Por otra parte, la densidad de una clase depende de las asociaciones internas (As-int) entre las palabras claves que la constituyen (PC-int), e indica la cohesión interna de las clases. Dado que las asociaciones son valuadas entre 0 y 1 de acuerdo con el índice $E(i,j)$ que se exponen más abajo (ver fórmula 1), tanto la centralidad como la densidad exprimen los valores medios, respectivamente, de las (As-ext) y (As-int), presentados en las dos primeras columnas de la tabla 1.

Tabla 1. Las clases y sus propiedades estructurales

		Centralidad	Densidad	PC-int	PC-ext	As-int	As-ext	Citación
1	Breast Neoplasms	0.022	0.116	10	7	9	8	2
2	Carcinoma	0.041	0.134	7	3	6	4	2
3	Skin Neoplasms	0.000	0.074	4	0	3	0	0
4	Neoplasms	0.010	0.246	9	1	10	1	1
5	Screening	0.027	0.057	10	5	10	7	7
6	Cervical Intraepithelial Neoplas	0.000	0.119	7	0	7	0	0
7	Mortality	0.010	0.137	5	1	5	1	1
8	Women's Health	0.019	0.139	8	4	11	4	4
9	Prostatic neoplasms	0.081	0.183	8	2	15	2	10
10	Silicon dioxide	0.000	0.496	8	0	20	0	0
11	Cisplatin	0.000	0.429	5	0	5	0	0

Tabla 2. Los elementos clasificados por clase

		Autores	Revistas	Articulos	Específicos	
1	Breast Neoplasms	296	23	62	46	74%
2	Carcinoma	165	15	38	27	71%
3	Skin Neoplasms	120	15	26	23	88%
4	Neoplasms	130	13	32	26	81%
5	Screening	349	30	81	65	80%
6	Cervical Intraepithelial Neoplas	164	11	32	31	97%
7	Mortality	53	7	19	12	63%
8	Women's Health	84	10	25	13	52%
9	Prostatic neoplasms	149	7	34	21	62%
10	Silicon dioxide	77	8	24	20	83%
11	Cisplatin	20	4	7	7	100%

La citación (última columna de la tabla 1) es un efecto del algoritmo empleado por SDOC, que se asemeja a una relación orientada de *i* a *j* o de *j* a *i*. Se llama citación, aquí, al hecho de que la relación nace en una clase *i* (origen) y termina en otra clase *j* (citada). Este orden no es necesariamente recíproco, en el sentido que la clase *i* presenta un enlace con la clase *j* pero la clase *j* puede no presentar ningún enlace con la clase *i*. Si se hacen las sumas de las columnas "As-ext" y "Citación" el total es igual en las dos: 27. En cambio si se comparan las líneas de las dos columnas se ve que los números son desiguales. Tomemos por ejemplo la clase 9 "Prostatic neoplasms" con As-ext = 2 y Citación = 10. Esta diferencia indica que la clase atrae más que lo que ella emite. Esta es otra forma de evaluar las clases (como se argumenta en Polanco, 2002, en base a la teoría de grafos). Nótese, de paso, que en el análisis de enlaces entre los sitios de la web (conocido como "link analysis") se distingue entre "hubs" (emisores de enlaces) y "autorities" (receptores de enlaces) (Kleimberg, 1998).

En la tabla 1 se encuentran clases con un valor nulo (0.000) de centralidad, las clases 3, 10 y 11, lo cual indica que se trata de clases aisladas y, en el caso de las clases 10 y 11 (Silicon dioxide y Cisplatin), con una elevada densidad, es decir grado de cohesión interna.

Además de las propiedades estructurales, las clases contienen los datos clasificados por ellas: número de palabras claves ("*termes*") que cada una de ellas representa, de autores ("*auteurs*") y de artículos ("*titres*") que pertenecen a la clase, así como de las revistas ("*sources*") en donde los artículos han sido publicados por los autores. En el caso de la clase 5, "Screening", por ejemplo:

TERMES	AUTEURS	TITRES	SOURCES
10	349	81	30

Cada una de las cifras puede ser activada a fin de visualizar su contenido. La clasificación SDOC al nivel de las palabras claves es exclusiva: cada término pertenece a una sola y única clase. Los documentos, en cambio, pueden ser clasificados en más de una clase, dado que son indexados por más de una palabra clave.

La tabla 2 señala el número de autores, artículos y revistas agrupados por clase, permitiendo visualizar los centros de interés alrededor de los cuales se distribuye la investigación. Se distinguen los artículos específicos por clase, es decir pertenecientes exclusivamente a la clase, indicándose su número y porcentaje. La relación entre "artículos" y "específicos" permite apreciar el grado de especificidad de las clases, como puede apreciarse. En suma, la tabla 2 representa otro cuadro-indicador que refleja la situación de la investigación científica en el período considerado (2002-2006). Como se ha dicho, la evaluación científica de dicha situación es de competencia de los expertos en el campo de la oncología.

La información cuantitativa que se lee en la tabla 2 permite delinear una idea del volumen o talla de las clases. Si ellas representan "centros de interés", las cifras muestran en dónde se concentra o se distribuye el interés científico de los investigadores y en qué medida. Al igual que los "autores" y las "revistas", los "artículos" pueden pertenecer a más de una clase, excepto los "artículos específicos*".*

STANALYST presenta las clases exponiéndolas como texto HTML en cuatro listas:

- Lista de las palabras clave internas
- Lista de las palabras clave externas
- Lista de las asociaciones internas

- Lista de las asociaciones externas

La primera lista informa sobre los términos que constituyen la clase, y la segunda sobre los términos externos con los cuales los primeros tienen asociaciones, que aparecen en la lista de asociaciones externas. Luego vienen las listas de las asociaciones internas constitutivas de la clase y de las asociaciones externas, es decir, que ligan la clase con otras clases. Las asociaciones internas y externas son relaciones fundadas en la co-ocurrencia y valuadas entre 0 y 1 de acuerdo con el coeficiente de asociación *E*(*i*,*j*), llamado coeficiente de equivalencia, cuya fórmula de cálculo es:

$$E_{(i,j)} = \frac{\left|C_{(i,j)}\right|^2}{\left|F_{(i)}\right| \square \left|F_{(j)}\right|} \Rightarrow A_{(i,j)} \qquad (1)$$

La fórmula expresa la co-ocurrencia de los términos *i* y *j* al cuadrado dividida por el producto de la frecuencia (ocurrencia) respectiva de cada uno de los términos en la colección de documentos. El algoritmo de clasificación se ejerce sobre las asociaciones ponderadas por *E*(*i*,*j*). En STANALYST, bajo la etiqueta "*poids*", es decir "peso", figura el valor de *E*(*i*,*j*), indicándose además el número de documentos en los que el par de términos *i* y *j* aparecen juntos, es decir su coocurrencia, y que, con relación al valor de *E*(*i*,*j*), puede considerarse su extensión o soporte –la extensión de *E*(i,*j*) es entonces el número de documentos que soportan la coocurrencia (*i*,*j*).

El siguiente es un ejemplo de tres asociaciones internas de la clase 5 ("Screening") en orden decreciente según su peso *E*(*i*,*j*):

Poids	**Cooccurence**	**Terme i**	**Terme j**
0.180	3	Adenocarcinoma	Prostate
0.067	2	Postoperative complications	Risk Factors
0.058	5	Breast Cancer	Mammography

Como se observa, las asociaciones son más o menos fuertes, lo que quiere decir que su importancia en la clase o en la red de clases depende del valor de *E*(*i*,*j*). El otro valor a integrar en el análisis de las asociaciones es su "soporte", medido en número de documentos, que está indicado por el número de co-ocurrencias. Además, todas las palabras claves son ponderadas por un peso cuyo modo de cálculo está dado por la siguiente formula:

$$P_{(i)} = \frac{\left|O_{(i)}\right|}{\left|A_{\text{int} \wedge ext}\right|} \qquad (2)$$

En donde *O* es la ocurrencia del término *i* en las asociaciones internas y externas de la clase, y *A* el número total de asociaciones internas y externas de la clase. *P*(*i*) es entonces una medida de la centralidad del término *i* en la estructura de la clase. El término más central es automáticamente escogido por el algoritmo como la etiqueta de la clase. Aquí tenemos un ejemplo de *P*(*i*), bajo la etiqueta "*poids*", acompañado de la frecuencia del término que figura bajo etiqueta "Libellé": el ejemplo que sigue corresponde a los tres primeros términos de la clase 5 ("Screening").

Poids	**Fréquence**	**Libellé**
0.353	9	Screening
0.294	10	Prostate
0.294	43	Breast Cancer

En este ejemplo, el término “Screening” es el más central de la clase y como tal etiqueta la clase entera (clase n° 5 en las tablas 1 y 2).

Cada documento ocupa un rango en la clase, la lista de documentos clasificados se ordenan siguiendo un ranking y cuyo modo de cálculo es la formula (3):

$$R_{(d)} = \frac{\sum P_{(i)}}{T_{(d)}} \qquad (3)$$

Es decir, la suma de los pesos de las palabras claves *P*(i) indexando el documento *d* y presentes en la clase, dividido por el total *T* de las palabras claves indexando el documento *d*. Cuando un mismo documento ha sido clasificado en dos o más clases, entonces presentará diferentes valores de ranking *R*(*d*) en cada una de las clases. Este índice permite jerarquizar los elementos clasificados y ofrece al analista un indicador de la pertinencia de los documentos clasificados en la clase, como en el ejemplo que sigue:

Poids	Numéro	Titre
0.824	000518	Free PSA and prostate volume on the diagnosis of prostate carcinoma
0.823	000430	Breast cancer screening: physicians related issues.

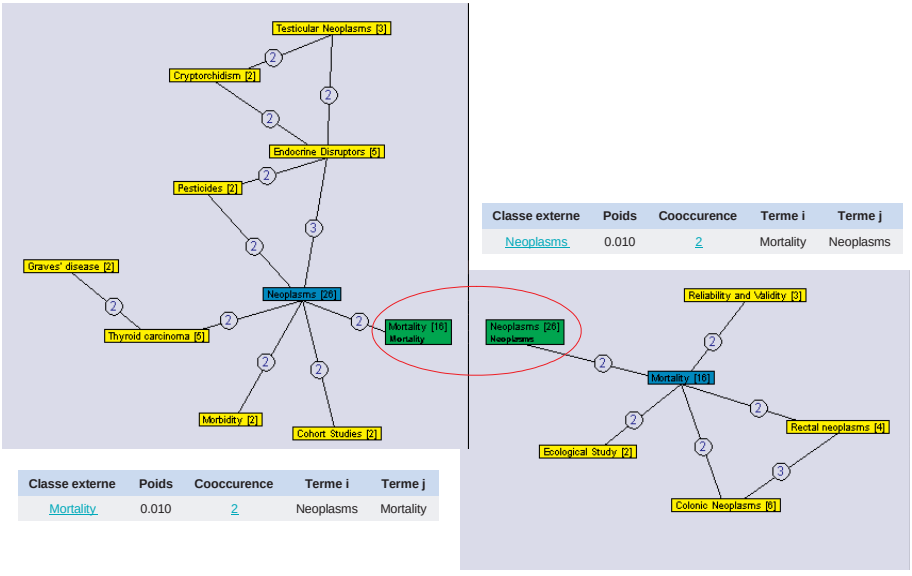
El ejemplo se refiere a los dos primeros documentos de la clase “Screening”. En STANALYST, tanto el número de identidad (“*numéro*”) de los documentos como los títulos permiten acceder a la referencia bibliográfica completa. La utilización de las bases SciELO permitirá, a partir de la referencia bibliográfica, acceder a los documentos “*full text*” correspondientes.

4.4.2. Grafos

Las clases son grafos (que se visualizan por medio de applets JAVA) en los cuales los nodos son los términos internos o externos de las clases y las relaciones o aristas son las asociaciones internas o externas. La figura 6 representa los grafos de las clases 4 y 7, es decir “Neoplasms” y “Mortality” respectivamente.

Se puede observar en la figura 8 la pareja de clases “Neoplasms” - “Mortality” en el mapa 2 bajo la forma de dos nodos etiquetados enlazados por una relación. Este hecho aparece desarrollado en la figura 6 como dos grafos enlazados por una sola asociación de valor *E*(*i*,*j*) = 0,010 y soporte = 2 documentos. En la lista de asociaciones externas de las dos clases esta relación bi-direccional aparece como se indica en la figura 6.

Figura 6. Grafos de la pareja de clases "Neoplamsms" - "Mortality"



Classe externe	Poids	Cooccurence	Terme i	Terme j
Neoplasms	0.010	2	Mortality	Neoplasms

Classe externe	Poids	Cooccurence	Terme i	Terme j
Mortality	0.010	2	Neoplasms	Mortality

Obsérvese además que los nodos del enlace son los términos más centrales de las dos clases (señalados en oscuro en los grafos). El círculo encierra los dos nodos, de modo que ellos son visualizados en los dos grafos distintamente: indicando los términos y su frecuencia (primera línea) y el nombre de las clases a las cuales ellos pertenecen (segunda línea). Las dos clases asociadas constituyen una red, como lo muestra la representación por grafos.

En el mapa temático producido a partir de los resultados de la clasificación, el conjunto de clases vía las asociaciones externas constituyen a su vez un grafo en el cual los nodos son las clases y las relaciones o aristas las asociaciones externas, de manera que se puede decir que se trata de un grafo de grafos o, en otras palabras, una red de redes (Polanco, 2002). Recordemos que las clases juegan el rol de poner en evidencia los temas de investigación o los focos de interés que estarían presentes en el corpus de datos. Entonces, las clases pueden ser analizadas como los nodos de una red de conocimiento. En la red, cada nodo representa un número de conceptos (términos) que constituyen redes o, si se quiere, subredes. Por otra parte, cada nodo-clase agrupa un número de revistas, artículos y autores, determinando la talla del sector de actividad que el nodo-clase representa. Además, como las clases presentan una densidad interna, siendo más o menos densas, ello permite de evaluar la cohesión de cada nodo-clase formando parte de la red de clases. Otra propiedad de las clases es su centralidad en la red de clases. En ella hay clases comparativamente más o menos centrales. Este valor de centralidad puede medirse según diversos criterios. La tabla 1 muestra dos formas: la primera es el valor medio (columna "Centralidad"), como se ha explicado en la sección 4.4.1, y la segunda es simplemente el número de relaciones entre las clases (y que figura en la columna As-ext). A estas dos formas podemos agregar una tercera, la suma de los valores $E(i,j)$ de las asociaciones externas (A-ext).

4.4.3. Mapas

La visualización de la información es el objetivo de los mapas. Como lo han dicho Gershon y Pages (2001), "la visualización de la información es un proceso que trans-

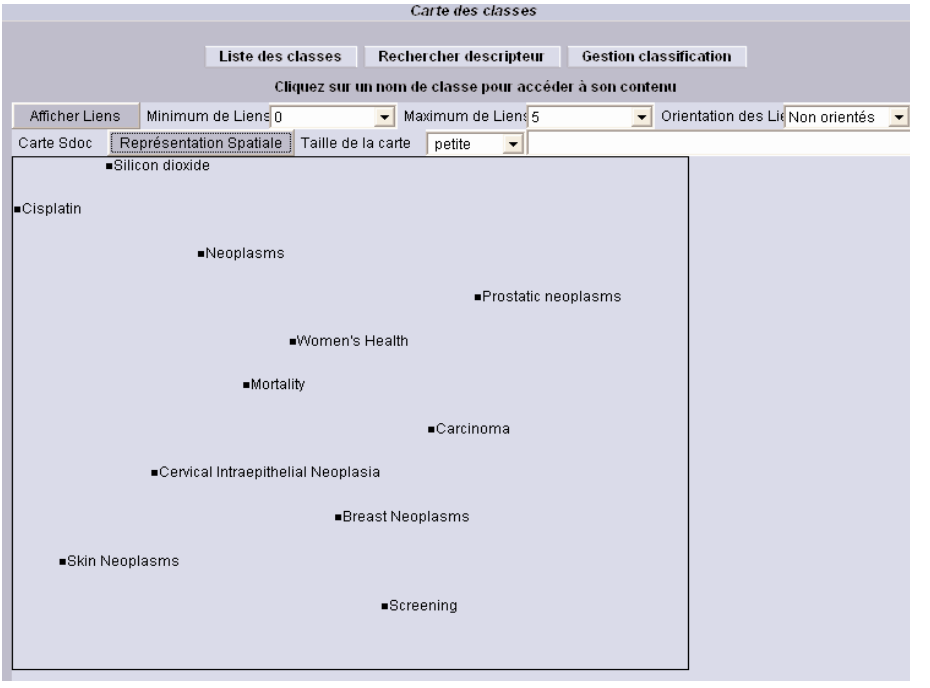
forma los datos, la información y el conocimiento en una forma que descansa en el sistema visual humano para percibir la información que incluye. Su objetivo es permitir al usuario / espectador observar, entender y hallarle sentido a la información". Por su parte, Carl, Mackinlay y Shneiderman (1999), afirman: "La visualización puede ser definida como el uso de representaciones visuales, interactivas y asistidas por computador para amplificar la cognición" y la "visualización de la información como el uso de representaciones visuales, interactivas y asistidas por computador de datos abstractos para amplificar la cognición".

En "knowledge discovery", se considera que la visualización implementa por sí misma un modelo cuya capacidad explicativa el usuario puede examinar. Cabe citar al respecto la observación hecha por Brachman y Anand (1996): "La exhibición apropiada de datos y sus relaciones puede dar al analista una perspectiva que es virtualmente imposible obtener mirando tablas de resultados o simples resúmenes estadísticos. De hecho, para algunas tareas, la adecuada visualización es lo único que se necesita para resolver un problema o confirmar una hipótesis, aun cuando habitualmente no pensamos en el trazado de una imagen como un tipo de análisis".

Las clases aparecen en el mapa como nodos etiquetados. La posición de las clases sobre el mapa es función de la densidad (Y) y la centralidad (X) en el plano de visualización. Las clases más densas se posicionan en la parte superior del mapa y las clases más centrales hacia la derecha del mismo. Los valores de estas dos propiedades que definen en el mapa las coordenadas (X, Y) se encuentran en la tabla 1 de la sección 4.4.1. Nótese que la clase 9 "Prostatic neoplasms" es la más central (0,081) y que la más densa es la clase 10 "Silicon dioxide" (0,496). Como puede apreciarse, el mapa 1 (figura 7) es la visualización de la información numérica de las columnas densidad y centralidad de la tabla 1. Como se ve en el mapa, la clase 5 "Screening" es la clase con menos cohesión interna (0,057), aunque su centralidad en el conjunto es mediana (0,027).

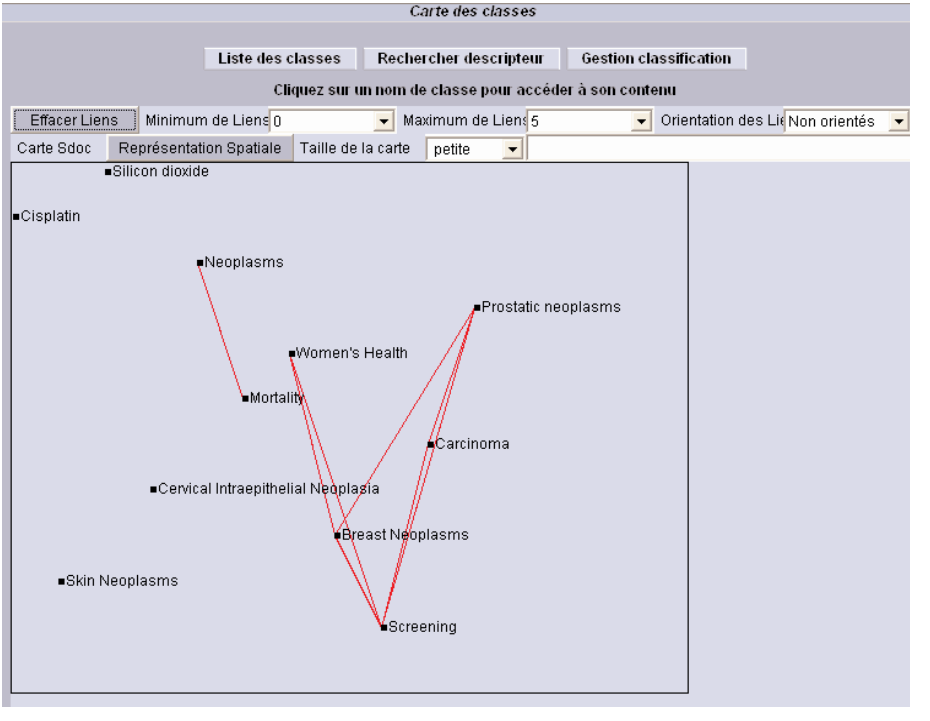
Como efecto de la activación del botón "Afficher Liens" (mapa 1) se produce la representación de los lazos entre las clases que exhibe el mapa 2. Aquí pasamos de la representación puramente espacial a la visualización de la red que estructura las clases, y en consecuencia los conocimientos a los cuales estas clases hacen referencia. Siendo las clases indicadores de temáticas de investigación, o de centros de interés científico, la red sugiere una cierta organización cognitiva del espacio de conocimientos y de los actores científicos (individuos y/o instituciones).

Figura 7. Mapa 1: representación espacial de las clases



370

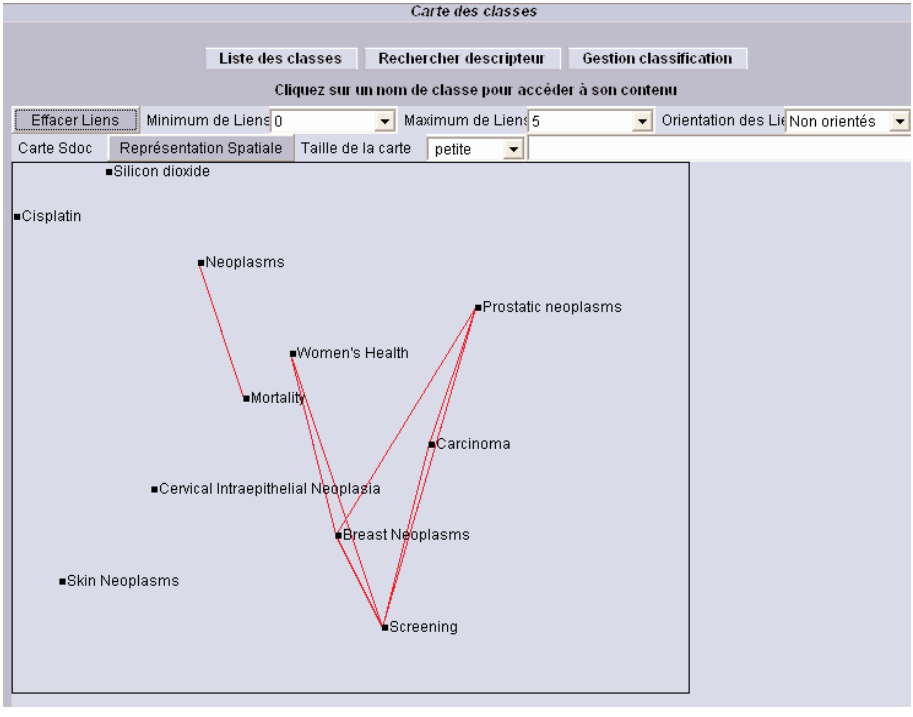
Figura 8. Mapa 2: visualización de la red de clases



En el mapa 2 (figura 8) se observan cuatro nodos (“Silicon dioxide”, “Cisplatin”, “Cervical intraepithelial neoplasms”, “Skin neoplasms”) que no entran en el sistema de relaciones constitutivo de la red: llamémoslos nodos-islas (su centralidad es 0.000 en la tabla 1). Se ven dos redes, una de sólo dos nodos-clases (“Neoplasms” - “Mortality”) y otra de cinco nodos-clases (“Prostatic neoplasms” - “Carcinoma” - “Screening” - “Breast neoplasms” - “Women’s health”), que es posible descomponer en tres subredes: 1) “Prostatic neoplasms” - “Carcinoma” - “Screening”, 2) “Prostatic neoplasms” - “Breast neoplasms” - “Screening” y 3) “Screening” - “Breast neoplasms” - “Women’s health”. Este trabajo de análisis de la información debe ceder la plaza al experto que, dados sus conocimientos y experiencia y asistido por estos indicadores, es capaz de realizar la interpretación científica de las clases y de sus posiciones relativas en el espacio de conocimiento “cáncer”. La interpretación del experto podrá validar o no, o agregar comentarios pertinentes. Es el trabajo de análisis que se ha llamado vigilancia científica o inteligencia estratégica (en las secciones 1.1 y 1.2).

Lo que en el mapa 2 (figura 8) se muestra como una pareja de nodos enlazados por una asociación, “Neoplasms” y “Mortality”, se ha visto en la figura 6 como grafos, así como la asociación que los enlaza y que es posible visualizar en el mapa. La articulación de estos dos modos de representación es necesaria. El mapa informa sobre la posición espacial de la pareja de nodos con respecto a los otros y los grafos informan acerca de la estructura interna de los nodos. STANALYST permite al usuario efectuar esta doble operación.

Figura 9. Mapa 3: Visualización del número de asociaciones entre las clases (nodos)



La figura 9 pone en evidencia el número de asociaciones que representan los enlaces entre los nodo-clases. La asociación entre dos clases puede ser múltiple. Activando directamente las clases, el usuario obtiene la visualización de las relaciones entre ellas. El número que se ve en el interior de un círculo sobre las relaciones corresponde al número de asociaciones que cada una representa. Aquí se han activado una a una sucesivamente todas las clases y las etiquetas en negro pasan a blanco, indicando así la activación de la clase por el usuario o analista. El mapa informa visualmente el número de asociaciones que existen entre las clases. Se observa que la relación entre las clases "Breast neoplasms" y "Prostatic neoplasms" es la más elevada (cinco).

Con un doble clic sobre la clase se despliega toda la información que ella representa, a la cual se ha hecho referencia en la sección 4.4.1. Es así que el mapa no sólo es un medio de visualización sino también de navegación en la información elaborada. Sobre estos recursos el experto científico puede apoyarse para analizar el espacio de información definido por el conjunto de artículos científicos.

4.4.4. Análisis de redes

Retomemos la red de cinco nodos-clases ("Prostatic neoplasms" - "Carcinoma" - "Screening" - "Breast neoplasms" - "Women's health") y siete relaciones, anotada como R = (5,7). Como se ha señalado, esta red puede descomponerse en tres subredes:

- "Prostatic neoplasms" – "Carcinoma" – "Screening"
- "Breast neoplasms" – "Prostatic neoplasms" – "Screening"
- "Women's health" – "Breast neoplasms" – "Screening"

Si se analiza la estructura se podrán extraer algunas observaciones. En primer lugar, la estructura de la red se compone de tres triángulos tangentes, que poseen un nodo común ("Screening"). En el análisis de las subredes es necesario considerar la densidad de los nodos-clases, que se da en orden decreciente: "Prostatic Neoplasms" 0,183, "Women's Health" 0,139, "Carcinoma" 0,134, "Breast Neoplasms" 0,116, "Screening" 0,057. En principio, la densidad mide la cohesión del nodo-clase, la cohesión de la micro-red que es el nodo-clase. Es en esos términos que los nodos (= micro-redes) entran en el análisis de las subredes y de la red como subconjuntos más o menos densos.

En segundo lugar, en la estructura triangular se distinguen siete parejas de nodos por efecto justamente de las siete relaciones (señaladas por una línea) que, como muestra la figura 9 (mapa 3), representan en realidad diecisiete asociaciones (suma de los números que etiquetan las relaciones). Las relaciones tienen, como se ha visto, dos propiedades esenciales: soporte y fuerza de asociación. Se llama soporte al número de documentos que sostienen la asociación, y fuerza al valor de la asociación dado por el coeficiente *E*(*i,j*) expuesto en la fórmula 1 de la sección 4.4.1.

Tabla 3. Soporte y centralidad de las asociaciones de la red R = (5, 7)

Nodo (i)	Nodo (j)	Núm. Docs.	Núm. As.	? E(i,j)	
Breast Neoplasms	Prostatic Neoplasms	12	5	0,149	15%
Prostatic Neoplasmas	Carcinoma	4	3	0,142	14%
Prostatic Neoplasmas	Screening	4	2	0,088	9%
Women's Health	Screening	4	3	0,067	7%
Carcinoma	Screening	2	1	0,021	2%
Breast Neoplasms	Screening	4	2	0,020	2%
Breast Neoplasms	Women's Health	1	1	0,007	1%

La centralidad puede expresarse de dos maneras: o bien simplemente por el número de asociaciones, que corresponde a la llamada centralidad de grado en el análisis de redes sociales, o bien por la suma de los valores *E*(*i,j*) de las asociaciones, que corresponde a los grafos valuados o ponderados (como se desarrolla en Polanco y San Juan, 2006, 2007). La tabla 3 expone los valores del soporte y de la centralidad de las asociaciones entre los nodos en orden decreciente de ?*E*(*i,j*).

Si se toma como ejemplo, en referencia al soporte de las asociaciones, las dos relaciones entre "Breast Neoplasms" y "Screening", "Mammography" - "Breast Neoplasms" y "Diagnosis" - "Breast Cancer", se verá que ellas adicionan un soporte de cuatro documentos, puesto que cada una tiene como soporte dos documentos.

0.013 2 Mammography Breast Neoplasms

1. Breast cancer's secondary prevention and associated factors

2. Cost estimate of mammographic screening in climacteric women

0.007 2 Diagnosis Breast Cancer

1. Trastuzumab in the treatment of advanced breast cancer: Our single-center experience and spotlights of the latest national consensus meeting

2. Prospective Study of The Ultrasound Features in the Diagnosis of Solid Breast Lesions

Se trata, en suma, de los artículos que proporcionan la significación científica detallada de la relación, visualizados en STANALYST como aquí se expone. Activando el número aparecen en la pantalla los dos títulos de los artículos correspondientes, comprobándose que las dos asociaciones tienen como soporte dos pares de documentos distintos. Luego, activando los títulos, se accede a la referencia bibliográfica completa. Este mecanismo permite al experto verificar la pertinencia y el sentido científico de la asociación, lo cual puede hacerse si es necesario para todas las asociaciones de todas las clases.

Si las dos relaciones tienen un soporte igual a dos documentos, no es el caso en lo que se refiere a la fuerza de ellas: la relación entre "Mammography" y "Breast Neoplasms" es casi el doble de fuerte (0,013) que la relación "Diagnosis" y "Breast Cancer" (0,007). Desde el punto de vista del análisis de la información esto quiere decir que la probabilidad de encontrar asociado en el corpus de documentos el término "Mammography" con el término "Breast Neoplasms es de 1,3 %, mientras que la probabilidad de la asociación "Diagnosis" y "Breast Cancer" es de solo 0,7%. Para la interpretación científica, es decir, el contenido y la significación científica de esta información, es necesaria la asistencia de una persona con competencia científica y clínica. En otras palabras, aquí es necesaria la intervención de un experto en la materia, como se hace habitualmente en la minería de datos o de textos y en el ámbito de la vigilancia científica.

Otro ejemplo es el de la pareja de nodos "Neoplasms" y "Mortality", que constituye en sí una red, registrada como R = (2,1), o sea, constituida por dos nodos y una relación. Los nodos tienen respectivamente una densidad 0,246 y 0,137, un volumen total de 32 y 19 documentos, y un volumen específico de 81% y 63% (véanse tablas 1 y 2), y están relacionados por una sola asociación cuyo soporte es dos y su fuerza igual a 0,010 (que define la centralidad igual de los dos nodos). Esta información permite a la vez comparar los nodos y evaluar la relación que los asocia, débil comparativamente a la fuerza de las asociaciones internas (densidad). Otro elemento de comparación, llámeselo sociológico, es el volumen en número de autores (actores científicos o investigadores): en este ejemplo, "Neoplasms" 130 y "Mortality" 53.

Aquí se debe repetir una vez más lo que ya se ha dicho en cuanto a la interpretación científica de estos indicadores: ella es competencia de científicos del dominio. El trabajo del analista de la información se detiene en la proposición y justificación de los indicadores que pone a la disposición del experto científico. Por cierto que se trata de un trabajo que debe llevarse a cabo en estrecha colaboración.

5. Conclusión

El análisis que se ha desarrollado en la sección 4 se basa en la información que STANALYST produce y ofrece al analista a partir de las bases SciELO. Por cierto que no es el sistema el que realiza el análisis. El sistema proporciona, por un lado, las herramientas de trabajo, los diferentes módulos y las operaciones que ellos ejecutan, los cuales el analista pone interactivamente en acción, y por otro lado, una información especialmente elaborada para asistir y alimentar el trabajo de análisis de los actores humanos.

De esta forma, STANALYST contribuye a la realización de la meta de las bases SciELO: dar visibilidad a la producción científica nacional y regional, a la difusión de la llamada "ciencia perdida". Este artículo pretende ser una ayuda para quienes se interesan por analizar la ciencia que se produce en los países de América Latina y que se encuentra colectada en las bases SciELO. En cuanto a las fuentes de información, además de las bases FRANCIS, PASCAL (del INIST/CNRS) y SciELO, otras bases podrán ser conectadas a STANALYST.

El objetivo de este trabajo fue presentar las características del sistema STANALYST - SciELO para el análisis de la información científica y técnica y, sobre todo, mostrar mediante un estudio de caso, el del cáncer, como se puede con STANALYST analizar el contenido de las bases SciELO. Lo que aquí se ha expuesto no es más que un ejemplo que debe completarse ciertamente por un trabajo de interpretación y validación por expertos del dominio científico considerado y que, por lo demás, puede extenderse a otras aéreas.

Como se dijo en la introducción, STANALYST constituye une tecnología de la inteligencia al servicio de tareas de vigilancia científica (*science watching*) o inteligencia estratégica en el campo de las ciencias y, en definitiva, una ayuda a la toma de decisiones en política científica.

Además de los indicadores estadísticos de RICYT, y de los que SciELO ya permite, se dispone ahora de una herramienta complementaria, que permite ir más allá de la simple información estadística de la ciencia. STANALYST abre así la posibilidad de realizar un análisis de contenido, apuntando a la representación del conocimiento contenido en los documentos y al mismo tiempo de los actores (individuos e instituciones) que se encuentran al origen de ellos. O como se ha dicho: "El descubrimiento del conocimiento refiere al proceso de descubrir conocimiento útil a partir de los datos" (Fayyad et al., 1996). En el enfoque que se ha presentado se ha buscado especialmente poner en evidencia mediante clases, grafos y mapas las redes del conocimiento presente en los documentos (datos), redes que se pueden considerar como las estructuras sociales y conceptuales de un frente de investigación.

Finalmente, cabe decir que este trabajo fue pensado como una forma de establecer el "acta de nacimiento" de STANALYST - SciELO. Quienes firman este trabajo confían en que así lo sea, ya que contribuyeron de una u otra manera a la realización del proyecto.

Bibliografía

ALTHUSSER, L. (1965): *Pour Marx*, Paris, Maspero.

BESAGNI, D., FRANÇOIS, C., POLANCO, X. y ROCHE, I. (2004): "Stanalyst®: Une station pour l'analyse de l'information", *Actes de Veille Stratégique Scientifique et Technologique* VSST2004, Toulouse, 25-29 octobre, pp. 319-320.

BRACHMAN, R. J. y ANAND, T. (1996): *The Process of Knowledge Discovery in Databases*, en Fayyad et al., c.2, pp. 37-57.

BROOKES, B. (1980): "The foundations of information science. Part I. Philosophical aspects », *Journal of Information Science,* vol. 2, pp. 125-133.

CALLON, M., COURTIAL, J-P. y PENAN, H. (1993): *La Scientométrie.* Paris, Presses Universitaires de France.

CARD, S. K., MACKINLAY, J. D. y SCHNEIDERMAN, B. (1999): *Readings in information visualization using vision to think*. San Francisco, Morgan Kaufmann Publisher.

COURTIAL, J-P. (1990): *Introduction à la scientométrie*. Paris, Anthropos Economica.

CRONIN, B. y DAVENPORT, E. (1993): "Social Intelligence", *Annual Review of Information Science and Technology*, vol. 28, pp. 3-44.

DAILLE, B., ROYAUTE, J. y POLANCO, X. (2000): "Evaluation d'une plate-forme d'indexation de termes complexes", *Traitement Automatique des Langues*, vol. 41, N° 2, pp. 395-422.

DEGENNE, A. y FORSE, M. (2001): *Les réseaux sociaux*, Paris, Armand Colin.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMITH, P. y UTHURUSAMY, R. (eds.) (1996): *Advances in Knowledge Discovery and Data Mining*, Menlo Park, Calif., AAAI Press & The MIT Press.

GERSHON, N. y PAGE, W. (2001): "What Storytelling Can Do for Information Visualization?", *Communication of the ACM*, vol. 44, N° 8, pp. 31-37.

GRIVEL, L. y FRANÇOIS, C. (1995): "Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique", *SOLARIS 2*, Presses Universitaires de Rennes, pp. 81-112, disponible en http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2grivel.html

GRIVEL, L., MUTSCHKE, P. y POLANCO, X. (1995): "Thematic Mapping on Bibliographic Databases by Cluster Analysis: A Description of the SDOC Environment with SOLIS", *Journal of Knowledge Organization*, vol. 22, N° 2, pp. 70-77.

JACQUEMIN, C., DAILLE, B., ROYAUTÉ, J. y POLANCO, X. (2002): "In Vitro Evaluation of a Program for Machine-Aided Indexing", *Information Processing & Management*, vol. 38, Issue 6, pp. 765-792.

KLEINBERG, J. (1998): "Authoritative sources in a hyperlinked environment", *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*. Extended version in *Journal of the ACM* 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.

MAIMON, O. y ROKACH, L. (eds.) (2005): *The Data Mining and Knowledge Discovery Handbook*, Berlin, Springer.

MENEGHINI, R., MUGNAINI, R. y PACKER, A. L. (2006): "International versus National Oriented Brazilian Scientific Journals. A Scientometric Analysis Based on SciELO and JCR-ISI Databases", *Scientometrics*, vol. 69, n°.3, pp. 529-538.

POLANCO, X. y SAN JUAN, E. (2007): "Hypergraph Modelling and Graph Clustering Process Applied to Co-Word Analysis", *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*, Madrid, Spain, June 25-27, vol. II, pp. 613-618: http://xavier.polanco.googlepages.com/home

POLANCO, X. y SAN JUAN, E. (2006): "Text Data Network Analysis Using Graph Approach", *Proceedings of the 1th International Conference on Multidisciplinary Information Sciences and Technology, InSciT2006*, Mérida, Spain, October 25-28, vol. II, pp. 586-592: http://www.instac.es/inscit2006/search.php?language=en

POLANCO, X. (2006): “STANALYST, un sistema de ayuda al análisis de la información”, en *El espacio público de las ciencias sociales y humanas*, Buenos Aires, Editores del Puerto, pp. 98-103.

POLANCO, X. (2002): “Clusters, Graphs, and Networks for Analysing Internet-Web Supported Communication within Virtual Community”, *7th International ISKO Conference*, Granada, Spain, 10-13 July, en *Advances in Knowledge Organization*, Vol 8, Würzburg, Ergon Verlag, pp. 364-371: http://xavier.polanco.googlepages.com/home

POLANCO, X. , FRANÇOIS, C., ROYAUTÉ, J. y BESAGNI, D. (2001): “STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology”, *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, July 16-20, Sydney, Australia, Vol. 2, pp. 871-873.

POLANCO, X. (1997a): “La notion d’analyse de l’information dans le domaine de l’information scientifique et technique”, *Colloque INRA*, 21-23 octobre, Tours. En P. Volland-Neil (coord.): *L’information scientifique et technique: Nouveaux enjeux documentaires et éditoriaux*, Paris, INRA, pp. 165-172.

POLANCO, X. (1997b): “Infometría” e ingeniería del conocimiento: exploración de datos y análisis de la información en vista del descubrimiento de conocimientos”, en H. Jaramillo y M. Albornoz (comps.): *El Universo de la medición*. Bogotá, TM Editores, COLCIENCIAS, RICYT, pp. 335-350.

POLANCO, X., Grivel, L. y Royauté, J. (1995): “How To Do Things with Terms in Informetrics: Terminological Variation and Stabilization as Science Watch Indicators”, *Proceedings of the Fifth International Conference of the International Society for Scientometrics and Informetrics*. Edited by M.E.D. Koening and A. Bookstein. Medford, NJ, Learned Information Inc., pp. 435-444.

POPPER, K. R. (1972): *Objective Knowledge*, Oxford, The Clarendon Press.

ROYAUTE, J. (1999): *Les groupes nominaux complexes et leurs propriétés: application à l’analyse de l’information*, Thèse de doctorat, Université Henri Poincaré, Nancy 1.

WASSERMAN, S. y FAUST, K. (1999) : *Social Network Analysis. Methods and Applications*, Cambridge University Press.