ATLAS DE LA CIENCIA ESPAÑOLA: PROPUESTA DE UN SISTEMA DE INFORMACIÓN CIENTÍFICA

Félix Moya-Anegón[1], Víctor Herrero-Solana[1], Benjamín Vargas-Quesada[1], Zaida Chinchilla-Rodríguez[1], Elena Corera-Álvarez[1], Francisco Muñoz-Fernández[1], Víctor Guerrero-Bote[2], Carlos Olmeda-Gómez[3]

Resumen: En el presente trabajo se propone un sistema de información cuyo objetivo general consiste en representar gráficamente la investigación científica española. Dicha representación gráfica se concibe como una colección de mapas –de ahí el término atlas– que persigue tres objetivos fundamentales:

- Facilitar a la comunidad científica española un instrumento para el análisis de la estructura que forman los diferentes campos científicos y sus correspondientes frentes de investigación, con el fin de mejorar su capacidad de interacción con otros dominios de conocimiento e institucionales pertenecientes al Sistema de Ciencia en que se integran.
- Brindar un interfaz gráfico que permita funciones de navegación a través de los espacios semánticos que forman los diferentes mapas. Este interfaz permitirá el acceso a la información documental disponible al modo de los sistemas denominados Bibliotecas Digitales.
- Representar la evolución de la investigación en los dominios institucionales y de conocimiento objeto de estudio a través de mapas dinámicos que mejoren la capacidad de la comunidad científica para analizar tendencias en el desarrollo de futuras líneas de investigación.

Palabras clave: mapas de la ciencia, sistemas de recuperación de la información, interfaces visuales de recuperación de información.

Abstract: In this paper we propose an information system that allows the graphic representation of the Spanish research. The representation is composed by a collection of maps –an atlas– that follow three basic objectives:

- To give to the the researchers a tool for the analysis of the scientific fields structure and research fronts. The principal goal of the system is to provide a powerful interface with the knowledge domains.
- To develop a graphic interface as a navigation tool through semantic spaces in the maps. This interface gives access to a specific systems called Digital Libraries.
- To represent the evolution of the production in both institutional and knowledge domains, with dynamic maps, that can be useful to analyse the future trends in scientific research.

Keywords: science maps, information retrieval systems, visual information retrieval interfaces (VIRI).

[1] Departamento de Biblioteconomía y Documentación. Universidad de Granada. Correo-e: victorhs@ugr.es.
[2] Departamento de Informática. Universidad de Extremadura.
[3] Departamento de Biblioteconomía y Documentación. Universidad de Carlos III de Madrid.
Recibido: 22-5-03; 2.ª versión: 29-10-03.

1 Introducción

1.1 Antecedentes al Atlas de la Ciencia

Como no podía ser de otra manera quien primero comenzó a hablar de la posibilidad de contar con un Atlas de la Ciencia fue el creador del Science Citation Index y fundador del Institute for Scientific Information, Eugene Garfield. Los trabajos de Garfield en torno a la posibilidad de describir y representar a la ciencia en función de las citas bibliográficas que realizan los científicos se remontan a la década de los 50 (1) y entroncan con los planteamientos teóricos de su precursor Derek de Solla Price (2).

Si bien los índices de citas comienzan a ser una realidad tangible en la propia década de los 60, es a mediados de los 70 cuando Garfield comienza a plantear, un tanto tímidamente si se quiere, la posibilidad de representar un verdadero Atlas de la Ciencia con la información que ya se había logrado acumular en casi diez años de existencia (3). La propuesta puede ser calificada de tímida ya que no brinda demasiados elementos para ser llevada a la práctica, particularmente por el problema de los métodos y materiales necesarios para un desarrollo óptimo de la idea, el cual se constituyó en una constante en los años posteriores.

A principios de los años 80, Garfield retoma la idea y anuncia la creación de un prototipo de atlas correspondiente al campo temático compuesto por la Bioquímica y la Biología Molecular (4). Por aquel entonces, Garfield exponía las ventajas de contar con una herramienta tan potente como el Atlas: determinar los trabajos más citados de cada campo temático, construir una red de relaciones entre estos trabajos, a partir de estas relaciones establecer el comportamiento de los diferentes subcampos (102 en total), etc. La técnica de representación propuesta era el «mapa de clusters» *(cluster mapping)*, la cual había sido abordada en trabajos inmediatamente anteriores (5) (6). Cada uno de los 102 subcampos o frentes de investigación constituye un capítulo del atlas, el cual está a su vez compuesto por cuatro componentes: un mini-review, un cluster-map que muestre la conectividad de los documentos núcleo *(core)* del frente, una bibliografía completa de los trabajos que constituyen dicho núcleo, y una lista de los trabajos más importantes *(key)* citados por el núcleo.

El principal elemento que introduce Garfield son los *mini-reviews*, pequeñas entradas, al estilo de una enciclopedia, que describen brevemente las características de cada frente, y que estaban a cargo de los autores más destacados de cada núcleo. Este elemento, si bien es sumamente útil a la hora de caracterizar cada frente, introduce un sesgo propio de autor en detrimento de la información estructural que brinda la interrelación de trabajos. Garfield parece haber dado a estos *reviews* demasiada importancia, incluso llega a afirmar que más que un atlas se propone construir una verdadera «Enciclopedia de la Bioquímica». Incluso refuerza esta idea al mencionar que él mismo siente una inclinación por este tipo de documento como reflejo de la historia de la ciencia, y recuerda que trabajó como asesor de la Enciclopedia Americana. Para Garfield, la principal ventaja de este tipo de enciclopedia radica en el hecho de que se encontraría estructurada temáticamente en función de la libre interrelación de los trabajos científicos y no basada en clasificaciones y taxonomías a priori, como en el caso de las enciclopedias tradicionales.

Con respecto a la naturaleza de los análisis desarrollados, Garfield plantea la necesidad de superar el método del «bibliographic coupling» de Kessler (7), con el método

de cocitación de autores (ACA) desarrollado al mismo tiempo por Henry Small y Belver Griffith en Estados Unidos (8) (9) y Valentina Marshakova en la Unión Soviética (10). Por otra parte, no hace mucho hincapié en el método a utilizar para la representación gráfica.

A lo largo de la década de los 80, el ISI sólo publicó dos prototipos de atlas de dos áreas temáticas: Biotecnología y Genética Molecular (1978-80) (11) y Bioquímica y Biología Molecular (12). El propio Garfield, en su columna del Current Contents, solo habla esporádicamente de la evolución de estos atlas. En 1987, y en la misma columna, Garfield relanza esta idea y promete un cambio radical en esta línea de productos (13). Presenta un plan editorial para el periodo 1987-1990, en el cual se proyecta la publicación de 12 diferentes atlas en áreas que van desde la Medicina Clínica hasta las Ciencias Sociales. Con respecto al formato de esta nueva línea de atlas, parece no haber cambios con relación a los dos prototipos anteriores. La principal diferencia consiste en el hecho de que los nuevos atlas se plantean como publicaciones periódicas trimestrales, acumulables anualmente. Esta línea editorial parece no haber fructificado durante los años '90. En la actualidad, si accedemos al sitio web del ISI (http://www.isinet.com) y realizamos una búsqueda por la palabra «atlas», no obtendremos ningún registro. La línea de productos basada en los «atlas de la ciencia», parece no haber sido continuada, ignoramos por que razones.

Garfield no entra en detalles sobre las metodologías de representación para la construcción de mapas, tarea que parece haber delegado en Henry Small, investigador del propio ISI. Small sí lo hace y describe los métodos necesarios para representar gráficamente la información bibliográfica, valorando positivamente la metáfora del mapa para establecer relaciones, propuesta anteriormente por Price (14).

Small utiliza técnicas estadísticas multivariantes, análisis de cluster y Escalamiento Multidimensional (MDS), con las cuales construye los mapas de elementos (15). Estos mapas resuelven de manera aceptable la representación de relaciones entre elementos de un campo temático específico (perspectiva micro). Sin embargo, parecen no ser tan efectivas a la hora de representar relaciones entre grandes grupos temáticos (perspectiva macro). En el caso de los atlas de Garfield, esto no constituye un problema, debido a que de antemano se ha partido la totalidad del conocimiento científico en diferentes fragmentos, relativamente fáciles de representar. De esta forma se renuncia a las representaciones globales, y se potencian las parciales. Esto a todas luces constituye un inconveniente y sería equivalente a querer tener una visión del planisferio mediante la suma de mapas comarcales. Esta característica es la que nos lleva a afirmar que el término atlas en estos productos del ISI es un tanto pretencioso e incluso excesivo.

Si bien podemos detectar una discontinuidad en la línea de atlas ISI durante los años 90, Small ha seguido trabajando intensamente durante estos años, con el afán de avanzar en la construcción de mapas. A mediados de la década, presentó una aplicación para la generación de mapas denominada SCI-MAP (16). En este caso, en lugar de trabajar con MDS utiliza una metodología un tanto similar basada en la triangulación de clusters. No obstante, en este trabajo Small aún no aborda el problema de los niveles de representación. Ha sido recientemente en 1997, cuando aporta una solución a la representación mediante mapas de grandes espacios documentales, en dos (17) e incluso tres dimensiones (18). Para ello, propone en método que el denomina «Humpty-Dumpty», mediante el cual es posible realizar mapas en diferentes niveles de forma tal que pueden ser visualizados como si los más específicos fueran un zoom de los más generales. A

pesar de proponer diferentes niveles, no propone metodologías de representación diferentes para cada nivel, asumiendo que pueden representarse todos ellos con el mismo método. Por otro lado, los niveles que permite generar esta metodología son los más generales, difícilmente se podría utilizar las mismas técnicas para continuar el desarrollo de los niveles más específicos hasta llegar a la representación del investigador individual.

Finalmente, podemos concluir que el desarrollo de los atlas ISI no ha sido del todo satisfactorio por diferentes razones: a) porque en la práctica fueron concebidos como enciclopedias de la ciencia, y no como atlas, b) porque los atlas fueron desarrollados fragmentariamente, como soluciones puntuales a un área relativamente pequeña del conocimiento, con la correspondiente pérdida de las interrelaciones entre las disciplinas y el carácter universal del conócimiento, c) porque además de fragmentaria fue sesgada, a través de los reviews realizados por ciertos científicos, d) porque no se concibieron como interfaces electrónicos de acceso a las bases de datos del ISI, sino que lo hicieron como simples mapas impresos, e) porque las metodologías de representación nunca fueron un motivo de preocupación del equipo de desarrollo, f) no se planteó la representación de dominios institucionales, sino solo dominios de conocimiento, y por último g) tampoco se representaron los aspectos dinámicos de la investigación, ni la evolución de los frentes de investigación.

Recientemente, han aparecido otras experiencias que han incursionado en la utilización de mapas de la ciencia, en calidad de interfaz. Wormell presenta la posibilidad de integrar mapas de cocitación de autores, basados en la técnica análisis de redes, en un sistema de información que consiste en un portal temático especializado en Ciencias Sociales (19). Otra aplicación en esta línea es la presentada por Sotolongo, en este caso la técnica de representación utilizada no es el análisis de redes sino las redes neuronales. En concreto el autor presenta una variante del modelo de mapas a (SOM), denominado ViBlioSOM (20).

1.2 Atlas de la Ciencia Española

Aunque el desarrollo de los llamados Atlas de la Ciencia ha sido un objetivo largamente acariciado por los investigadores, no ha sido hasta que las tecnologías de la información y comunicación (TIC) han llegado al suficiente grado de madurez, cuando este viejo proyecto ha empezado a cristalizar. Un proyecto de esta naturaleza pretende representar gráficamente la investigación en un determinado dominio científico. Dicha representación gráfica se concibe como una colección de mapas, de ahí el término atlas.

El desarrollo de lo que en definitiva se concibe como un Sistema de Información Científica y una herramienta de análisis, se basa en la utilización combinada de metodologías basadas en técnicas de Análisis Multivariante y técnicas basadas en Análisis Estructural y de Redes. En gran medida es objeto del proyecto determinar en qué niveles de la construcción del atlas resultan más eficaces unos métodos de representación u otros y en qué forma deben ser parametrizados y combinados en su caso cada uno de los métodos con el fin de lograr el interfaz y la herramienta de análisis que mejores resultados obtenga en la fase de evaluación.

El Atlas se concibe como una doble estructura de mapas que mantienen enlaces entre sí a lo largo de sendas redes de conexiones. La estructura es doble porque existirá

una serie de mapas temáticos y otra serie de mapas institucionales, ambas representando la ciencia española en su totalidad, aunque en la actualidad solo se ha trabajo en los primeros. En el prototipo que hemos diseñado (http://www.atlasofscience.net), actualmente solo existe la estructura temática, aún no la institucional. Esto se debe en gran medida al problema de control de autoridades que conlleva el trabajo con este tipo de información bibliográfica. Los mapas no sólo están enlazados entre sí constituyendo un sistema de navegación, sino que permitirán el acceso a todo tipo de información –indicadores bibliométricos, registros bibliográficos y texto íntegro de los trabajos- a partir de las elecciones realizadas por los usuarios.

Inicialmente el proyecto restringe su ámbito de actuación a la investigación española de la década de los noventa y comienzo de la actual, a partir de los documentos referenciados en la base de datos Science Citation Index Expanded (SCI-E) del ISI. Esto quiere decir que el Atlas refleja la investigación española visible internacionalmente durante el periodo mencionado en las llamadas ciencias duras. Las metodologías utilizadas permiten con facilidad extender el proyecto de las ciencias sociales y humanas tras el proceso de identificación de las correspondientes bases de datos.

El esquema adjunto representa el proceso general de desarrollo. En este esquema es posible determinar la existencia de cuatro partes diferenciadas:

1. Fase de carga, nomalización y vectorización de la información fuente.
2. Fase de procesamiento de la información vectorizada para la obtención de los atributos y relaciones de los nodos de la red (etiquetas, posiciones, tamaños, relaciones internas y relaciones externas).
3. Fase de construcción del sistema de navegación gráfico y enlace con la base de datos
4. Fase de evaluación formal y de usabilidad.

En cada fase del proceso se puede observar la mención al punto correspondiente del apartado metodología en el cual se amplia la tarea con mayor grado de detalle. El proyecto se encuentra actualmente en la tercera fase (ver figura 1).

2 Objetivos del sistema

- Facilitar a la comunidad científica española un instrumento para el **análisis** de la estructura que forman los diferentes campos científicos y sus correspondientes frentes de investigación, con el fin de mejorar su capacidad de interacción con otros dominios de conocimiento e institucionales pertenecientes al Sistema de Ciencia en que se integran.
- Desarrollar un **interfaz** gráfico que permita funciones de navegación a través de los espacios semánticos que forman los diferentes mapas. Este interfaz permitirá el acceso a la información documental disponible al modo de los sistemas denominados Bibliotecas Digitales. Asimismo, dará acceso a un amplio elenco de indicadores bibliométricos convencionales que el sistema es capaz de calcular en tiempo real y como respuesta a las peticiones realizadas por el usuario.
- Representar la evolución de la investigación en los dominios institucionales y de conocimiento objeto de estudio a través de mapas dinámicos que mejoran la capa-

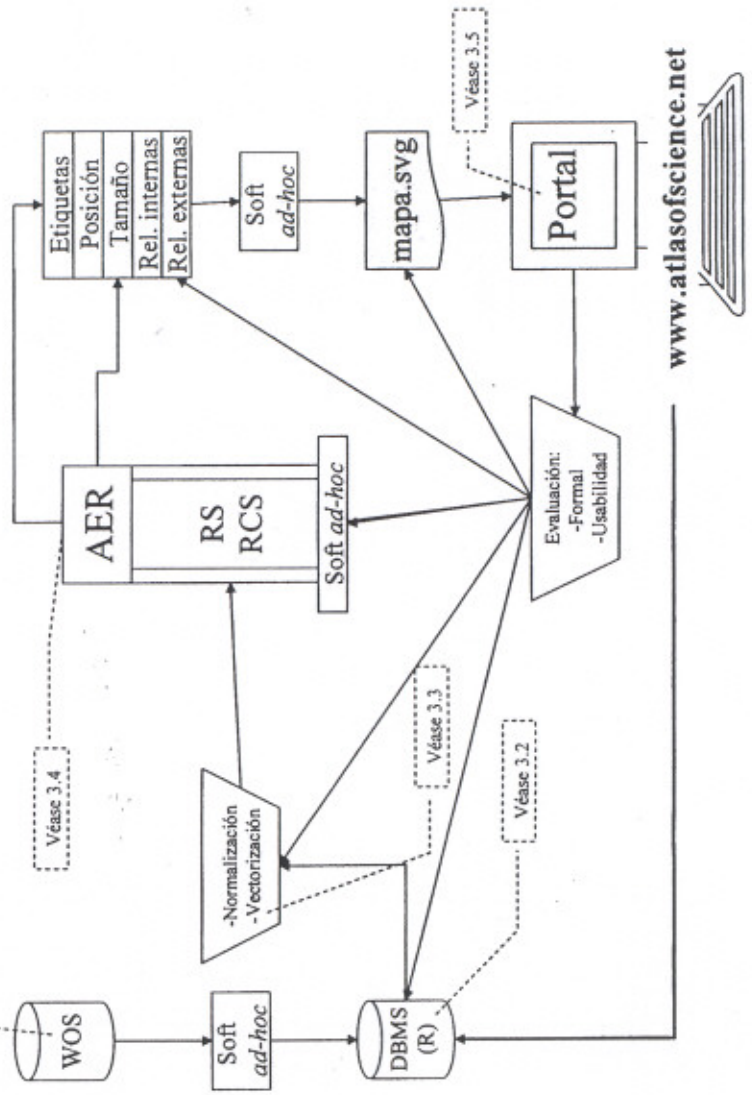

Figura 1
Proceso de elaboración del sistema

cidad de la comunidad científica para analizar tendencias en el desarrollo de futuras líneas de investigación.

3 Materiales

3.1 Origen de los datos

A partir de la información suministrada por las bases de datos Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI) y Arts & Humanities Citation Index (A&HCI) del Institute of Science Information (ISI), se han recuperado todos los registros con al menos una dirección española en el campo *Address Word* para el periodo 1995-2002 a través del Web of Science. Además se extrajo información del Journal Citation Report (JCR) para analizar cuestiones relativas a la adscripción de las revistas a las categorías temáticas, factor de impacto, etc.

Tradicionalmente, la historia de las bases de datos ISI ha estado plagada de críticas relacionadas con el sesgo en la cobertura de las revistas en términos de disciplinariedad y nacionalidad. No obstante, estudios recientes (21) que comparan la cobertura del SCI con la del Ulrich's International Periodicals Directory (U-S&T), demuestran que esto no es así. El conjunto de revistas SCI-JCR presenta un balance equilibrado con respecto al del U-S&T a nivel macro, por lo que afecta al menos a países y disciplinas. En contra de la creencia general, no existe un sesgo ISI a favor de países como Estados Unidos o campos temáticos como la Biomedicina, en algunos casos incluso existe una infra-representación. Las excepciones en cuanto a cobertura por disciplinas se centran en Alemania y en concreto en la agricultura y en lo referente a editores, destaca Francia. En general hay una sobre-representación de los principales editores en el SCI-JCR, pero en cualquier caso, este fenómeno no afecta a los objetivos de este estudio. Por otra parte, hay que tener en cuenta que es la única fuente de datos multidisciplinar e internacional que ofrece información institucional de todos los autores. A esto hay que añadir que la selección de estas fuentes se encuentran en consonancia con la evaluación de la investigación española en todos los campos científicos excepto en Derecho y Jurisprudencia, Historia, Arte, Filosofía, Filología y Lingüística, propuesta en las últimas convocatorias para la dotación de incentivos a los investigadores. Por tanto, consideramos que la fuente de datos es la apropiada y que los datos de partida reflejan de una manera consistente la investigación española visible internacionalmente.

3.2 Estructura de la base de datos

Los datos bibliográficos de origen requieren un tratamiento previo en parte automático y en parte manual. Por un lado, fue necesario desarrollar un software *ad-hoc* con el que se cargan los registros a una base de datos relacional (véase figura 2). La base de datos resultante contiene los campos con la información estructurada de los documentos y con las relaciones establecidas a priori, así como información adicional que se añade por procedimientos semiautomáticos. Por otra parte, es necesaria una normalización de los campos directamente relacionados con los distintos niveles objeto de estudio (disciplinas, instituciones, revistas, autores, etc.) para su identificación y posterior análisis.

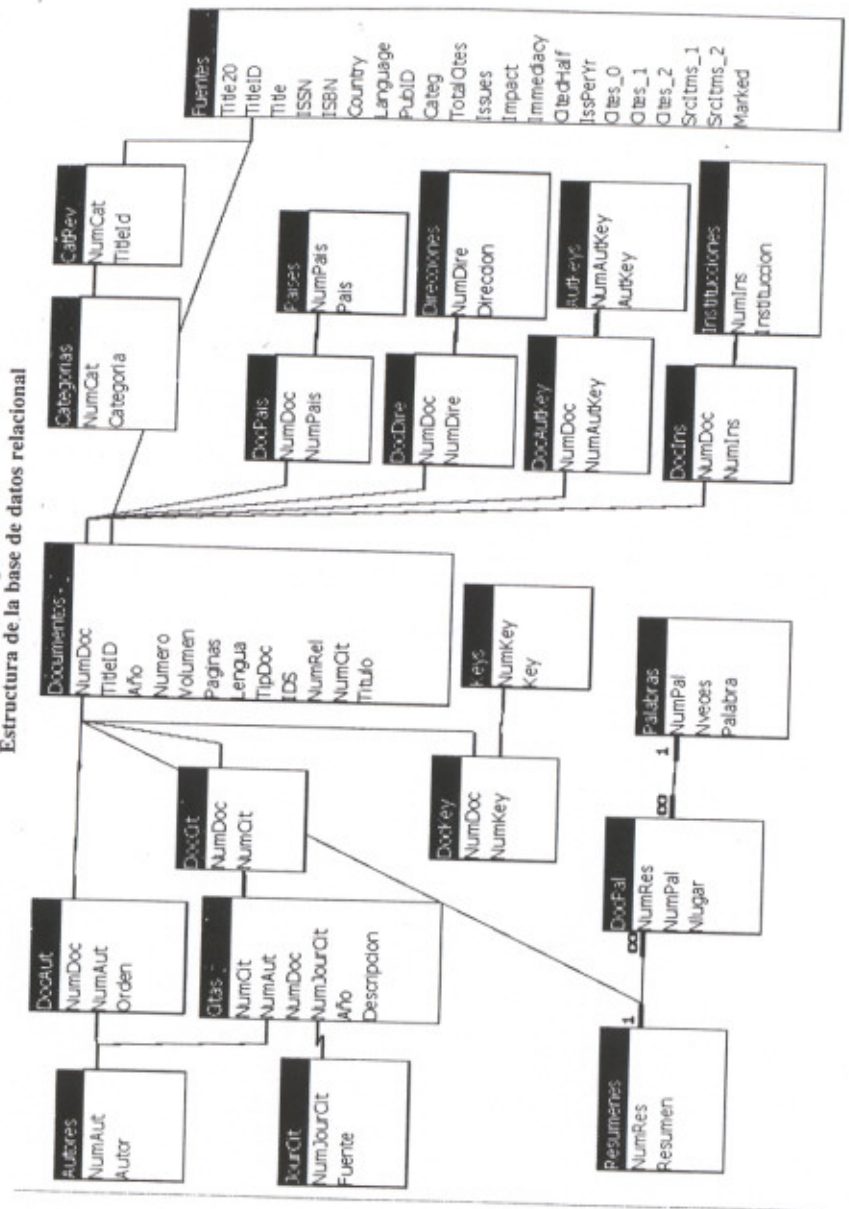

Figura 2
Estructura de la base de datos relacional

3.3 Tratamiento de los datos

En los últimos años han aparecido en la literatura de la especialidad, descripciones de proyectos o sistemas pilotos que intentan integrar diversas técnicas provenientes del campo de la bibliometría en diversos tipos de sistemas de información. Con relación a la inclusión de indicadores bibliométricos, encontramos un trabajo interesante sobre una experiencia desarrollada en el CINDOC (22).

Los campos susceptibles de normalizar vienen determinados por los niveles de agregación hasta los que se quiera descender. Desde el campo dirección es posible estudiar la producción científica usando países, ciudades y organizaciones principales como unidades de investigación. La información que proporcionan los otros campos del registro bibliográfico permite limitar el estudio para un cierto período de tiempo usando el año de publicación del artículo, o el nombre de la revista. Las revistas se pueden ordenar dentro de categorías temáticas. La clasificación de los artículos basada en las categorías de las revistas sirve para definir los campos en los que se trabaja. El JCR-ISI clasifica las revistas adscribiéndolas a una o más categorías temáticas, y ofreciendo datos como el factor de impacto de cada una de ellas.

La estructura del campo institucional *(address word)* contiene en la mayoría de las veces cuatro partes: la organización principal, un departamento de la organización, la ciudad y el país (Univ Granada, Fac Ciencias, Dept Quim Fis, Granada 18071, Spain). En muchos casos, sólo hay tres niveles, excluyendo el nivel departamental o el institucional. El país suele estar bien normalizado y la información sobre la ciudad puede normalizarse fácilmente al eliminar los códigos postales. El primer nivel institucional es decir, la organización principal puede tener un gran número de variantes y lo mismo ocurre con el segundo nivel institucional, el nombre de los centros o facultades.

Algunos institutos están intentando estandarizar las direcciones para hacer posible a gran escala el análisis de citas y de co-autoría de la producción de artículos institucionales. En algunos casos, codifican las principales organizaciones dentro de sectores generales, tales como universidades, institutos de investigación, industria, etc., y de esta forma permiten los estudios entre los sectores (23). Esto hace posible un análisis de dominio institucional partiendo de los datos convenientemente normalizados. En este caso la normalización se centró en el primer nivel institucional, es decir, organizaciones principales ya que descender a otros niveles conlleva un gran consumo de tiempo que sólo se justifica si profundizamos en una organización en particular. En una primera aproximación, se trata de tomar como unidad de análisis la organización principal en vez de los autores. Por tanto, el primer paso es refinar la información de las direcciones para permitir el análisis. Los errores que hay que subsanar tienen que ver con las variantes ortográficas y la adscripción de cada uno de los documentos a un centro. Por otro lado hay que contar con una serie de irregularidades en los datos de origen como puede ser las direcciones relacionadas con los hospitales o la ausencia en la dirección del primer nivel institucional.

En cuanto a un tercer nivel institucional, el nivel departamentos en el caso de las universidades, la dirección del departamento indica la disciplina de una institución particular, por ejemplo, DEP PHYS, DEP PHYSIOL, DEP CHEM. Aunque la información departamental está incompleta y poco normalizada y no siempre se corresponde con los departamentos, unidades o centros (24), hay trabajos que estudian la definición de los campos científicos a través de la denominación de los departamentos en el campo

address word (25) (26) (27). Este método no parece ofrecer de una manera consistente una conexión entre los límites que define el nombre del departamento y la disciplina temática, y la depuración que se requiere para tal objetivo excede las pretensiones de este proyecto. Por tanto, para la adscripción de los documentos a una determinada disciplina científica, en este trabajo se partirá de la información de la revista y de su categorización en base al JCR. Esto hará posible un análisis de dominio temático una vez hechas las correcciones correspondientes.

Otro problema de normalización que tienen las bases de datos ISI es la inconsistencia de algunos títulos abreviados de revistas en el listado del JCR y la forma en la que se presentan en las referencias de los artículos citados. Esto deriva en una pérdida de información a la hora de trabajar en el análisis de citas y cocitas que se asume de antemano. Con respecto al control de las citas hay que añadir que el ISI sólo registra el primer autor en la referencia. Para salvar esta deficiencia, se realizan los cálculos utilizando todos los autores de cada trabajo.

Todo esto nos lleva a afrontar el problema de la normalización desde frentes distintos y complementarios según los niveles de estudio. La problemática de cada nivel exige un tratamiento distinto en los datos. Una vez que se determinan las cuestiones y definiciones específicas que se necesitan para normalizar hay que trabajar sobre la base de datos.

3.4 Análisis de redes

Los últimos años se ha desarrollado una nueva forma de estudiar las estructuras sociales: el llamado análisis de redes *(network analysis)*. Esta nueva aproximación representativa ha alcanzado altos niveles de sofisticación metodológica y técnica y ha mostrado su altísimo valor en un amplísimo abanico de aplicaciones. El análisis de redes no es una mera técnica más o menos sofisticada para el análisis de fenómenos sociales, sino que es también una nueva aproximación teórica (28) (29).

La diferencia principal entre las explicaciones aportadas por el análisis estructural de redes con otras aproximaciones analíticas es la inclusión de conceptos e información acerca de las relaciones entre unidades. Ya sea queriendo estudiar el comportamiento individual en el contexto de relaciones estructuradas o queriendo estudiar directamente estructuras sociales, el análisis de redes maneja las estructuras en términos de redes de enlaces entre unidades. Las regularidades en estos enlaces dan lugar a estructuras representables. En el análisis de redes los atributos de los objetos de estudio son interpretados en términos de pautas o estructuras relacionales entre las unidades. Los ligámenes relacionales entre los objetos son interpretados en términos de pautas o estructuras relacionales entre unidades. Los enlaces relacionales entre objetos son lo primordial, los atributos son secundarios. Las relaciones pueden ser expresadas mediante una representación gráfica de los elementos, y consideramos que esta representación es susceptible de ser utilizada en la construcción de una interfaz visual con la base de datos (30).

Las relaciones que pueden ser representadas dentro de una comunidad científica son muy diversas, aunque nosotros nos centraremos en las relaciones derivadas de la cocitación de trabajos científicos. De esta manera, es posible analizar diversos dominios, ya sean estos temáticos, institucionales, e incluso personales. White ha clarificado estos últimos en cuatro grandes grupos (31):

a) las relaciones de coautoría,
b) la «identidad» del autor, basada en un análisis de las co-referencias realizadas por este en sus trabajos,
c) la «imagen» del autor, compuesta por un análisis de cocitas de este autor con el resto de la comunidad científica,
d) la denominada «creadores de imagen» (image makers), compuesta por todos los autores que han citado a un determinado autor y que por tanto han creado la imagen visible del mismo.

Las representaciones se crean mediante un software ad-hoc para el análisis de redes sociales, capaz de generar una red de los pares de elementos más cocitados. Así, en lugar de representar puntos en el espacio de forma un tanto ambigua, el resultado es una representación espacial de dichos elementos, de sus relaciones, de los grupos que forman debido a su alto grado de cocitación, y de las relaciones que existen entre esos grupos. Para lograr una representación de este tenor, utilizamos el algoritmo de Kamada-Kawai (32).

La red resultante debe ser podada mediante algún tipo de algoritmo que optimice su representación. Un algoritmo muy utilizado es el denominado Minimun Spannig Tree (MST), cuya aplicación es bastante sencilla. Otro algoritmo, algo más complejo, es el denominado Pathfinder o también conocido como PFNet (33). Con PFNet la similaridad entre los vértices del mapa se representa como un parámetro de distancia r, que puede ser fijado hasta infinito mientras que los pares de cocitación de autores sean un número ordinal. Para evitar una amalgama de enlaces y mejorar la visualización, un segundo parámetro q, restringe el número de enlaces entre los elementos. La interfaz no está limitada a ser un mero elemento visual-informativo, por lo que además de mostrar redes de cocitación, permitirá la interacción con el usuario mediante técnicas de zoom con las que se podrá acceder a grupos de elementos, además de regenerar la red partiendo de un área seleccionada.

El análisis de las redes de cocitación o co-referencias, así como de sus relaciones, pondrá de manifiesto no sólo las distintas perspectivas de una disciplina en función del autor seleccionado, sino aspectos hasta ahora invisibles para los sistemas tradicionales de representación de la información tales como: pautas o conductas de citación, relaciones entre autores y los grupos que conforman, colegios invisibles, etc.

3.5 Portal de información

Nuestro atlas es en realidad un complejo sistema integral de mapas, que se encuentran accesibles a través de un portal web. Todas las metodologías descritas en los puntos anteriores obtienen como salida un conjunto de datos, gracias a los cuales es posible construir una representación gráfica o mapa.

Como hemos dicho en la introducción, estos mapas tienen una doble finalidad: como elementos para el análisis de dominio y como interfaces con la propia información bibliográfica. Por esta última razón, no nos limitaremos a colgar imágenes del web, ni trabajaremos con aplicaciones prototípicas, sino que ponemos en línea la totalidad de la producción científica española.

El gran volumen de datos a gestionar nos lleva a plantearnos seriamente cuales serán las herramientas tecnológicas necesarias para un funcionamiento óptimo del sistema. En un primer momento, se desarrollará un prototipo que tendrá la información en formato

de base de datos relacional, por lo que el hospedaje del web se hará en una plataforma NT como Windows 2000 Server, y como servidor el IIS (Internet Information Server). La conexión con la base de datos se realizará mediante ODBC (Object Data Base Connection) a través de la tecnología ASP (Active Server Pages). El motor de la base de datos es, por tanto, de tipo relacional.

Independientemente de la plataforma utilizada, tenemos también varias alternativas a la hora de implementar tecnológicamente los distintos tipos de mapas. Los mapas son representaciones gráficas interactivas que permiten el acceso a otros mapas o conjuntos de datos, simplemente seleccionando alguna de sus zonas. En primer lugar elegimos los mapas sensitivos basados en formatos gráficos tales como JPG y GIF, dada su facilidad de implementación. No obstante, creemos que una solución más consistente contemplaría la utilización de un formato de gráficos vectoriales. El más indicado es el SVG (Scalable Vector Graphics), un estándar de dominio público desarrollado por el W3C (World Wide Web Consortium). La ventaja de este tipo de formatos es que pueden ser construidos mediante código generado automáticamente.

Estas mismas ventajas presenta otro estándar: el VRML (Virtual Reality Modeling Language). En este caso el VRML será utilizado para representar estructuras tridimensionales. El agregado de una dimensión adicional a un mapa bidimensional, concede la posibilidad de enriquecer la representación con información complementaria. No obstante, hay que tener en cuenta que las ventajas potenciales de las representaciones tridimensionales siempre estarán mediatizadas por el entorno de periféricos con que se cuente. Tanto en el caso del VRML como el SVG, se experimentará con su capacidad de representar movimiento, con el fin de reflejar aspectos dinámicos de la ciencia.

4 Los tres niveles del acceso temático

Cómo hemos dicho, el principal punto sobre el que se ha trabajado es el acceso a la información mediante una estructura temática. Este se logró a través de tres niveles que conducen al usuario desde los temas más generales a los más particulares. En la figura 3 podemos observar el mapa principal o de primer nivel. En él aparecen diferentes esferas, donde cada una de ellas representa una de las áreas temáticas de la ANEP (Agencia Española de Evaluación y Prospectiva), 25 en total. El tamaño de cada esfera es proporcional al volumen de artículos incluidos en cada área ANEP. Los enlaces muestran las relaciones de cocitación entre cada área. Lo interesante de este mapa se encuentra en la ubicación de las esferas, que permiten reconocer al menos cuatro grandes zonas: 1) la correspondiente a las Ciencias de la Vida, donde la Medicina es la más productiva y a la que se suman la Biología Celular y Molecular, las Ciencias Agrícolas, y la Psicología; 2) la Física y la Química; 3) las Ingenierías; 4) las Ciencias Sociales y Humanidades.

Es interesante apreciar la conexión de las Ciencias Sociales y Humanidades (caracterizadas por el SSCI y el A&HCI), con el resto de las disciplinas científicas (provenientes del SCI-E). La conexión se da a través de dos «puentes», el primero lo constituye la Psicología y Ciencias de la Educación, que se encuentra relacionada con la esfera de las Ciencias Sociales. El segundo puente es de carácter metodológico, ya que se realiza a través de las Matemáticas cuya incidencia en las Ciencias Sociales queda de manifiesto. Los enlaces dentro de cada agrupación también ponen de manifiesto relaciones más sutiles, sin embargo, la riqueza de estas relaciones se pone de manifiesto cuando accedemos a los mapas de segundo nivel.

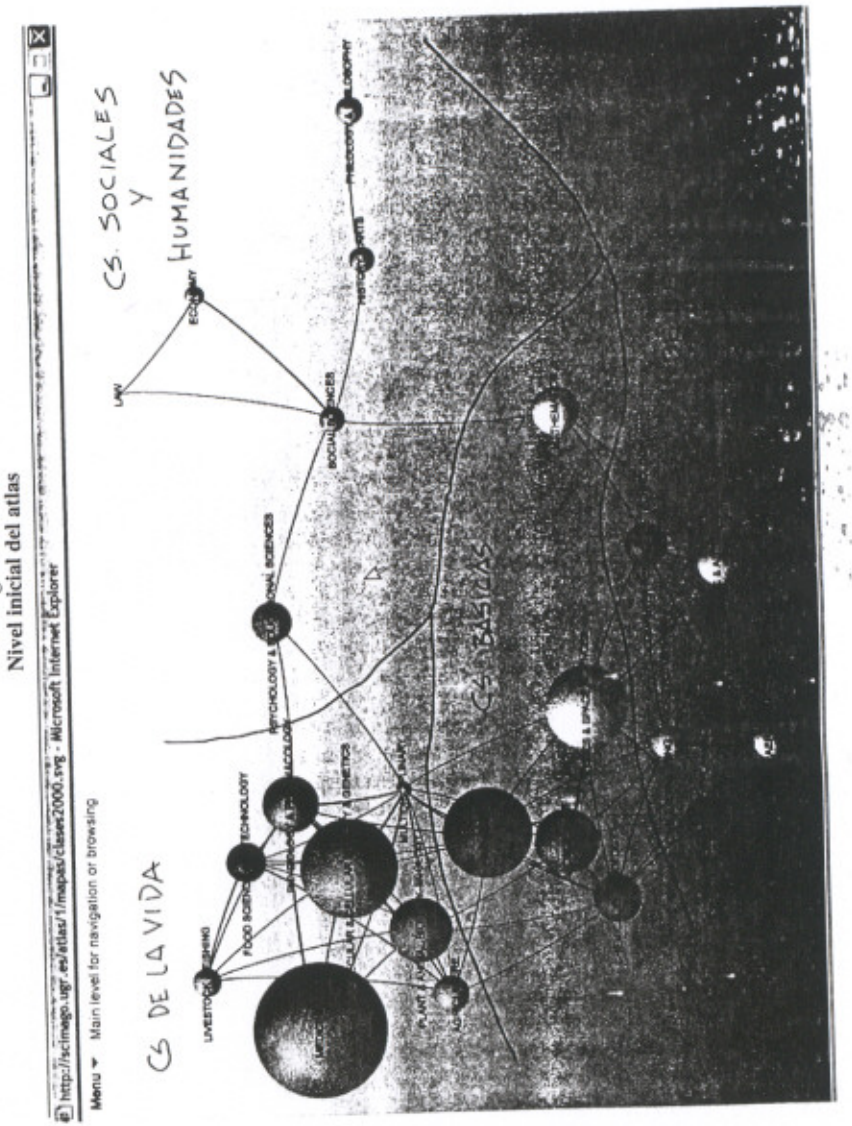

Figura 3
Nivel inicial del atlas

En el segundo nivel encontramos un mapa por cada esfera del nivel inicial. Aquí encontramos las categorías temáticas (subject categories) con las que el ISI clasifica su propia base de datos. Existen en total unas 247 categorías y en 222 de ellas aparece producción española. En cada mapa aparece un número variable de estas categorías, siempre que se encuentren adscritas al área ANEP correspondiente. Esta asignación la hace la propia ANEP, aunque nosotros ampliamos la asignación con algunas categorías que no habían sido oportunamente tenidas en cuenta.

En alguno de los mapas, como por ejemplo en Medicina, la cantidad de esferas representadas hace imposible la inclusión de todos los enlaces de relación. Por esta razón hemos optado por incluir un elemento de interactividad que permite visualizar sólo los enlaces de la esfera que se encuentran por el cursor. En la figura 4 podemos observar un mapa de este estilo, en este caso el correspondiente a Ciencias Sociales. Aquí el cursor se encuentra sobre la categoría *Management* y por esa razón solo aparecen los enlaces que parten de ella. Esto permite apreciar con más claridad las relaciones entre esferas. Además, como los enlaces visualizados son pocos, se ha introducido una ponderación de tal forma que la intensidad de relación entre categorías temáticas es directamente proporcional al ancho de la línea. En el caso de la figura, la relación mayor que presenta *Management* es con la categoría *Operations Research & Management Science*. En este nivel, la cantidad de trabajos existentes no solamente es proporcional al tamaño de la esfera, sino que además aparece entre paréntesis junto al nombre. Cuando seleccionamos cualquiera de las esferas, accedemos al mapa del tercer nivel.

En los mapas de tercer nivel también se representan categorías temáticas mediante esferas, aunque la naturaleza del mismo es diferente de la anterior. Aquí aparece la categoría seleccionada en el centro de la representación. De ella parten los enlaces hacia otras esferas de forma radial. Lo interesante es que las categorías representadas pueden formar parte o no de la misma área ANEP, además, la intensidad de relación se mide aquí por la longitud del enlace. En la figura 5, por ejemplo, encontramos el mapa correspondiente a la categoría *Information Science & Library Science,* y en él se aprecian las relaciones de esta categoría con otras de la misma área, como por ejemplo *Communication o Management,* pero también con categorías pertenecientes a otras áreas ANEP. En este caso, la categoría que se encuentra ligada con más intensidad es *Computer Science, Information Systems,* a pesar de pertenecer a otra área ANEP. Esto indica una fuerte relación que trasvasa las fronteras de cada área. La forma de identificar las categorías de la misma área y las que no lo son es mediante los colores de las esferas, que aunque no pueda apreciarse en la figura, es fácilmente identificable en la pantalla del ordenador.

Cuando seleccionamos la categoría central, obtenemos un listado de la producción completa española discriminada por años. Si en cambio seleccionamos una de las esferas de la periferia, accedemos a un nuevo mapa donde la categoría seleccionada es central y aparecen las nuevas categorías relacionadas en la periferia. De esta forma obtenemos un efecto de desplazamiento temático lateral. Una opción innovadora que hemos introducido, consiste en la posibilidad de seleccionar los enlaces, en lugar de las esferas, de forma tal que puede ser listada toda la producción que es común a dos categorías temáticas.

Este sistema de mapas puede ser examinado hacia delante, hacia atrás, de manera lateral, etc., a través de los nombres que figuran en la parte superior pero además pueden seleccionarse en cualquier momento una serie de opciones, a través de los términos

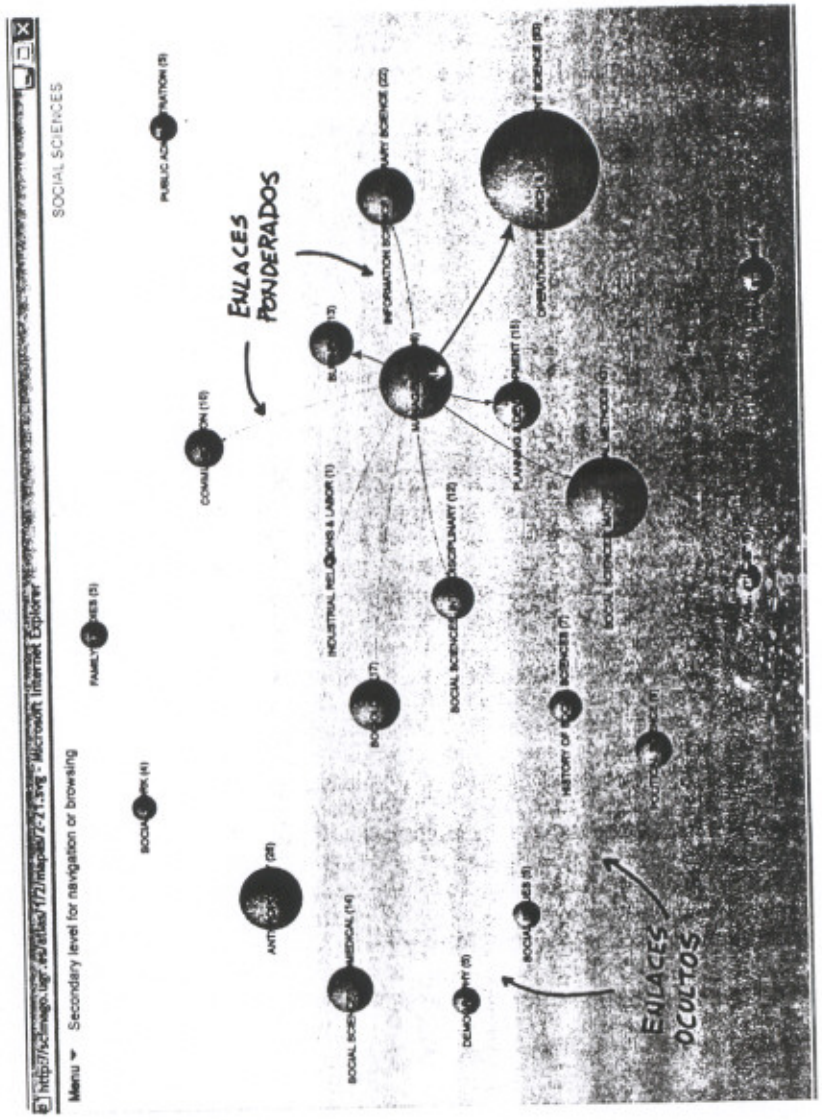

Figura 4
Mapa de segundo nivel

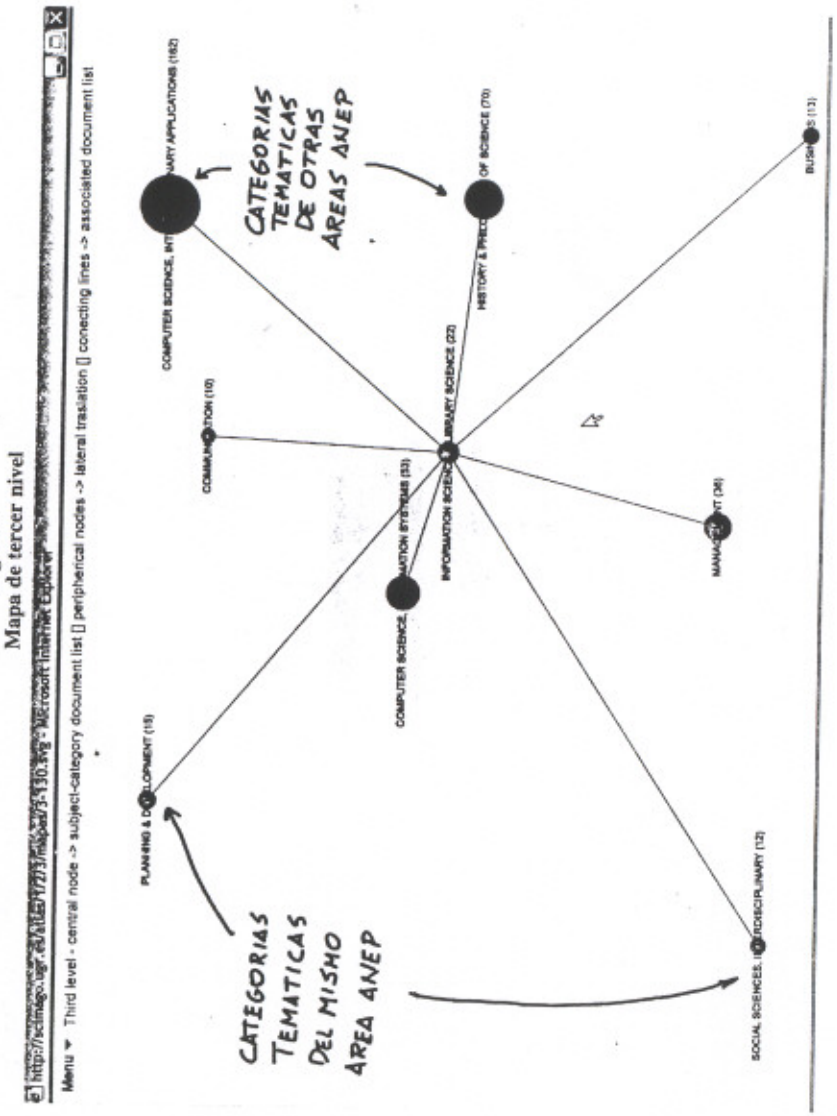

Figura 5
Mapa de tercer nivel

que se encuentran en la parte superior del mapa y que son también conocidos como «migas de pan». Además, cuenta con un menú en el que pueden seleccionarse una serie amplia de indicadores. Esto permite que al final no obtengamos el simple listado de la producción correspondiente a cada esfera, sino que en su lugar obtendremos la información en forma de un determinado indicador. Entre los indicadores disponibles se encuentran: listados de producción por instituciones españolas, listados de producción de instituciones extranjeras, producción por revista, producción por editor, producción por autor, coautoría, colaboración internacional, revistas más citadas, documentos más citados, producción por tipos de documentos, entre otros. Los indicadores propuestos están siendo considerados constantemente con el fin de hacerlos más precisos e incrementarlos.

5 Conclusiones

Hasta el momento, el sistema ha sido desarrollado en su estructura temática, sin embargo, las contribuciones científico-técnicas parecen ser prometedoras, no solamente porque constituyen un avance dentro de la especialidad, sino porque brindan una herramienta invalorable a todas las disciplinas científicas en su conjunto. Las líneas de aplicación son, asimismo, tan diversas como interesantes. Entre los sectores potencialmente interesados en las ventajas del Atlas, se encuentran todas la universidades, el Consejo Superior de Investigaciones Científicas, los distintos niveles de la administración pública con competencias en torno a la política científica (comunidades autónomas, gobierno central, autoridades europeas, agencias de evaluación, empresas privadas, etc.).

Una de las líneas de acción a desarrollar en un futuro cercano, está relacionada con la implantación de un proceso continuado de evaluación. En este caso hemos planteado dos tipos de evaluación: formal y de usabilidad. En el primer caso, se deberá realizar un completo control del sistema a medida que se va desarrollando, con el fin de detectar, por ejemplo, falta de representación de los documentos, mapas temáticos confusos o erróneos, etc. Este proceso de control se dará en todos los puntos del proceso, por lo que un error en el producto final (mapa) puede afectar directamente a cualquier fase del trabajo, con el fin de corregirla y reformularla.

El segundo tipo de evaluación apunta a medir el grado de usabilidad que el sistema tendrá para el usuario potencial. Este tipo de evaluación se realizará tardíamente, ya que para ello es necesario contar con un sistema lo más completo posible. En cuanto a los usuarios potenciales, debemos diferenciar claramente dos grupos. En primer lugar nos encontramos con los usuarios comunes, cuyo objetivo será simplemente la recuperación de una determinada información. En segundo lugar tendremos a los científicos expertos en una determinada área temática. Éstos, además de actuar como simples usuarios del sistema, también nos podrán decir hasta qué punto el Atlas es útil como herramienta de análisis de su propia especialidad científica.

Bibliografía

1. GARFIELD, E. Citation indexes for science. *Science*, 1955, n.º 122, p.108-11.
2. PRICE, D. S. Big Science, Little Science. New York: Columbia University Press, 1963.

3. GARFIELD, E. ISI's Atlas of Science may help students in choice of career in science. *Current Contents*, 1975, n.º 29, p. 5-8.
4. GARFIELD, E. Introducing the ISI Atlas of Science: Biochemistry and Molecular Biology. *Current Contents*, 1981, n.º 42, p. 279-87.
5. GARFIELD, E. ABCs of cluster mapping. Part 1. Most active fields in the life sciences in 1978. *Current Contents*, 1980, n.º 40, p. 5-12.
6. GARFIELD, E. ABCs of cluster mapping. Part 2. Most active fields in the physiscal sciences in 1978. *Current Contents*, 1980, n.º 41, p. 5-12.
7. KESSLER, C. Bibliographic coupling between scientific papers. *American Documentation*, 1963, vol. 14, n.º 1, p. 10-25.
8. SMALL, H. Co-citation in the scientific literature a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 1973, vol. 24, n.º 4, p. 265-69.
9. SMALL, H.; GRIFFITH, B. C. The structure of scientific literatures. I: identifying and graphing specialities. *Science Studies*, 1974, vol. 4, p. 17-40.
10. MARSHAKOVA, V. System of document connections based on references. *Nauchno-Tekhnichescaya Informatisya:* Series II, 1973, vol. 6, p. 3-8.
11. INSTITUTE FOR SCIENTIFIC INFORMATION. ISI Atlas of Science: Biochemistry and Molecular Biology 1978/80. Philadelphia: ISI, 1981.
12. INSTITUTE FOR SCIENTIFIC INFORMATION. ISI Atlas of Science: Biotechnology and Molecular Genetics 1981/82. Philadelphia: ISI, 1984.
13. GARFIELD, E. Launching the ISI Atlas of Science: for the new year, a new generation of reviews. *Current Contents*, 1987, vol. 1, p. 1-6.
14. PRICE, D. S.; BEAVER, D. Collaboration in an invisible college. *American Psychologist*, 1966, vol. 21, n.º 11.
15. SMALL, H.; SWEENEY, E.; GREENLEE, E. Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 1984, vol. 8, n.º 5-6, p. 321-40.
16. SMALL, H. A sci-map case study: building a map of AIDS research. *Scientometrics*, 1994, vol. 31, Nº 4, p. 229-41.
17. SMALL, H. Update on science mapping: creating large document spaces. *Scientometrics*, 1997, vol. 38, n.º 2, p. 275-93.
18. SMALL, H. A general framework for creating large scale maps of science in two or three dimensions: the SciViz system. *Scientometrics*, 1998, vol. 41, n.º 1-2, p. 125-33.
19. WORMELL, I. Bibliometric navigation tools for users of subject portals. *Journal of Information Science*, 2003, vol. 29, n.º 3, p. 193-201.
20. SOTOLONGO-AGUILAR, G.; GUZMÁN-SÁNCHEZ, M. V.; CARRILLO, H. ViBlioSOM: visualización de la información bibliométrica mediante el mapeo autoorganizativo. *Revista Española de Documentación Científica*, 2002, vol. 25, n.º 4, p. 77-484.
21. BRAUN, T.; GLANZEL, W.; SCHUBERT, A. How balanced is the Science Citation Index's journal coverage? a preliminary overview of macrolevel statistical data. En: Cronin, B.; Barsky, H. (eds.). *The web of knowledge: a festschrift in honor of Eugene Garfield*. Atkins Medford: ASIS, 2000.
22. FERNÁNDEZ, M. T.; CABRERO, A.; ZULUETA, M. A.; GÓMEZ, I. Constructing a relational database for bibliometric analysis. *Research Evaluation*, 1993, vol. 3, n.º 1, p. 55-62.
23. MELIN, M.; PERSSON, O. Studying research collaboration using co-authorships. *Scientometrics*, 1996, vol. 36, n.º 3, p. 363-77.
24. ZULUETA, M. A.; CABRERO, A.; BORDONS, M. Identificación y estudio de grupos de investigación a través de indicadores bibliométricos. *Revista Española de Documentación Científica*, 1999, vol. 23, nº 3, p. 333-47.
25. DE BRUIN, R.H.; MOED, H.F. Delimitation of scientific subfields using cognitive words form corporate addresses in scientific publications. *Scientometrics*, 1993, vol. 26, n.º 1, p. 65-80.

26. BOURKE, P.; BUTLER, L. Institutions and the map of science: matching university departments and fields of research. *Research Policy*, 1998, vol. 26, n.º 6, p. 711-18.
27. LEWISON, G. The definition and calibration of biomedical subfields. *Scientometrics*, 1999, vol. 46, n.º 3, p. 529-537.
28. SCOTT, J. Social network analysis: a handbook. London: Sage, 1992
29. RODRÍGUEZ, A. Análisis estructural y de redes. Madrid: Centro de Investigaciones Sociológicas, 1995. (Cuadernos Metodológicos, 16)
30. WHITE, H. D.; BUZYDLOWSKI, J.; LIN, X. Co-cited author maps as interfaces to digital libraries: designing Pathfinder Networks in the humanities. En: IEEE International Conference on information visualization, 2000.
31. WHITE, H. D. Author-centered bibliometrics through CAMEO's: Characterizations automatically made and edited online. *Scientometrics*, 2001, vol. 51, n.º 3, p. 607-37.
32. KAMADA, T.; KAWAI, S. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 1989, vol. 31, n.º 1, p. 7-15.
33. SCHVANEVELDT, R. W. Pathfinder associative networks: studies in knowledge organization. Norwood N.J.: Ablex, 1990.