

Adquisición de Datos (Búsquedas Bibliográficas)

Adquisición y selección de datos. En el proceso de KDD es importante la captura y creación de un almacén de datos sobre el problema bajo análisis. En esta etapa se debe considerar la estructura de las fuentes de información y sus características. En la actualidad existen bases de datos especializadas en todas las áreas científicas; cada una de ellas difieren en cobertura temática, criterios de selección de revistas y/o documentos, sesgos geográficos y lingüísticos y todas estas características deben analizarse de forma previa a la realización de un análisis.

De todas las publicaciones científicas-tecnológicas, solamente el artículo científico es considerado pieza clave para estudiar a la ciencia por medio de los análisis cuantitativos, es decir a través de los análisis bibliométricos. El indicador bibliométrico palabras comunes (Co-Word Analysis) elaborado a partir de los artículos seleccionados nos va a permitir detectar *conocimientos* que no existían explícitamente en ningún artículo de la colección, pero que surge de relacionar el contenido de varios de ellos.

MEDLINE®

MedLine es una base de datos bibliográficos producida por la Biblioteca Nacional de Medicina de los Estados Unidos. Entre sus ventajas como fuente de información está su amplia cobertura de revistas, pues contiene aproximadamente 15 millones de citas bibliográficas que provienen de más de 4,600 revistas que cubren los temas de la biomedicina, principalmente medicina, enfermería, odontología, oncología, medicina veterinaria, salud pública, ciencias preclínicas y de otras áreas de las ciencias de la vida [80]. En esta sección se describe en forma general como buscar y recuperar citas bibliográficas en MedLine.

MedLine asigna palabras claves a documentos que tratan algún tema de biomedicina específico. A este proceso de asignación se le conoce como *indización* y es simplemente la enumeración sucesiva de los diferentes términos del MeSH Vocabulary (Medical Subject Headings Vocabulary) que identifican el contenido o los contenidos de cada documento en MEDLINE. La indización es un proceso técnico que requiere de la aplicación de criterios uniformes como son la exhaustividad (multiplicidad), la especificidad, la coherencia, la imparcialidad, la fidelidad y el buen juicio [81].

El MeSH Vocabulary es un *tesauro* de palabras representativas sobre temas de biomedicina. Se integra por más de 33,000 palabras claves (términos), las cuales están clasificados en:

- Los **encabezados MeSH** (*MeSH Headings*) representan conceptos o temas generales que se encuentran en la literatura biomédica.
- Los **subencabezados MESH** (*MeSH Subheadings*), son palabras o frases, con las cuales, se califica un encabezado MeSH, esto es, estas palabras o frases se usan para caracterizar a los temas generales en sus aspectos más específicos.
- Los **conceptos suplementarios** (*Supplementary Concepts Records*) son palabras o frases usadas para detallar los efectos farmacológicos de algunos químicos. La guía de Conceptos Suplementarios forma un tesauro independiente del MeSH.

Un aspecto importante del MeSH es su estructura jerárquica. En esta estructura en forma de árbol, los términos del MeSH se ramifican en series de términos cada vez más concretos o específicos. La tabla 3 muestra las 15 ramas principales en las que se organiza el MeSH.

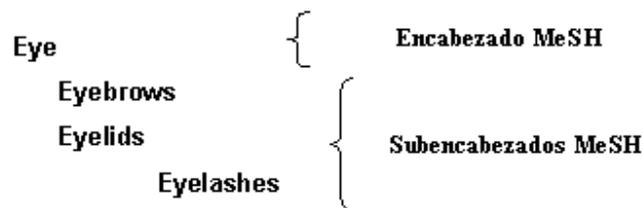
A	Anatomy
B	Organisms
C	Diseases
D	Chemical and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Biological Sciences
H	Physical Sciences
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology and Food and Beverages
K	Humanities
L	Information Science
M	Persons
N	Health Care
Z	Geographic Locations

Tabla 3: Categorías del MeSH Tree Structure

Veamos brevemente algunas categorías. La categoría **A** agrupa términos de anatomía referidos tanto a seres humanos como animales. La categoría **B** se refiere a organismos vivos. La categoría **C** agrupa enfermedades tanto experimentales como clínicas. Los términos relativos a una enfermedad se configuran en el siguiente orden: términos precoordinados como órgano/enfermedad («brain diseases», «skin diseases») o como organismo/enfermedad («salmonella infections», «trypanosomiasis»), órgano+término precoordinado u órgano+enfermedad («ileum, intestinal diseases», «conjunctiva, eye diseases»), síndrome+descriptivo («crying cat syndrome»), síndrome+epónimo («Korsakoff syndrome»), infecciones+términos generales precoordinados («Bordetella infections», «HIV-infections»), cáncer: tumor, cáncer y carcinoma son sinónimos; no se especifican diferencias entre tumores benignos y malignos; los tumores se indexan con términos que indican el tipo histológico («carcinoma, basal cell») y con términos que indican el órgano afectado («skin neoplasms»).

La categoría **D** agrupa sustancias químicas, endógenas y exógenas. La categoría **E** agrupa métodos para diagnóstico, terapéutica y equipamiento técnico, entre otros. Las técnicas y métodos se indexan solamente si son la materia principal de un artículo o si son tratados en detalle. Por ejemplo, un artículo que verse sobre el EEG en la epilepsia será indizado como «epilepsy» y «electroencephalography».

Veamos algunos ejemplos. La palabra clave “Eye” (*Ojo*) se localiza en la categoría **A**. Esta palabra es clasificada como encabezado MeSH. Las palabras “Eyebrows” (*Ceja*) y “Eyelids” (*Párpado*) son sus correspondientes subencabezados MeSH. También, se observa que “Eyelids” (*Párpados*) posee la palabra “Eyelashes” (*pestaña*) como subencabezado MeSH. (Vea Figura 1).



Químico	Aspirin <i>Anti-Inflammatory Agents, Non-Steroidal</i> <i>Cyclooxygenase Inhibitors</i> <i>Fibrinolytic Agents</i> <i>Platelet Aggregation Inhibitors</i>
Efectos	

Figura 1: Ejemplos

Mientras que la palabra “Aspirin” (*Aspirina*) se considera un químico. Algunos efectos farmacológicos que vemos en el cuadro de la figura 1 son los asignados a este químico por la guía de los Conceptos Suplementarios.

El Sistema Entrez– Pubmed

Una vez vistos en forma general, los elementos que son utilizados por la Biblioteca Nacional de Medicina durante el proceso de indización, pasemos a otro punto muy interesante: ¿Cómo buscar y recuperar citas bibliográficas mediante las *palabras claves*? Para ello, se expondrá en primer lugar el funcionamiento del sistema Entrez–Pubmed y en segundo lugar se da un ejemplo.

La Biblioteca Nacional de Medicina pone a disposición el sistema Entrez–Pubmed para la búsqueda y recuperación de las citas bibliográficas localizadas en Medline. Este servicio es gratuito y está disponible en la página electrónica de Pubmed.
<http://www.ncbi.nlm.nih.gov/entrez/>

El núcleo del sistema Entrez–Pubmed es el algoritmo llamado *Mapeo Automático de Términos*. Básicamente, el algoritmo se enfoca en encontrar coincidencias de términos o frases que son ingresados en los cuadros de búsqueda. Los términos o frases pueden ser nombres de temas, nombres de autores, nombres de revistas, instituciones, regiones, edades, términos técnicos, ISSUE, nombres químicos, etc. El funcionamiento del mapeo automático de términos es el siguiente: los términos o frases son comparados (en este orden) contra lo siguiente:

1. **Tabla de traducción MeSH:** contiene una lista alfabética de las palabras claves del MeSH; los sinónimos, las referencias cruzadas y los términos de entrada para las palabras claves del MeSH, los tipos de publicaciones, términos derivados del *Sistema de Lenguaje Médico Unificado*, los conceptos de nombres suplementarios correspondientes a los nombres de sustancias y sus sinónimos.

2. **Tabla de traducción de revistas:** contiene un listado alfabético de todos los títulos de revistas; abreviaturas de revistas en formato Medline; el número de identificación unívoco de una revista, (*International Standard Serial Numbers, ISSN*).
3. **Lista de frases:** contiene frases derivadas del *Sistema de Lenguaje Médico Unificado*; nombres de sustancias.
4. **Índice de autores:** contiene una lista alfabética con los nombres de los autores.

Si existe una coincidencia en cualquier etapa, el algoritmo se detiene y muestra los resultados. Si el algoritmo no obtuvo coincidencias en su primer intento de búsqueda, entonces entra en la fase de descomposición del término o frase mediante el operador AND. De nuevo, el procedimiento se repite, pero esta vez, el algoritmo empleará la instrucción *ALL Fields*.

Por ejemplo, supongamos que el algoritmo no encontró coincidencias para la frase *HIV Seropositive* en su primer intento. Ahora en su segundo intento descompondrá dicha frase en *HIV AND Seropositive*.. Debido a la opción descomposición se recomienda escribir entre comillas las frases que no queremos que sean descompuestas, por ejemplo, "*rheumatic diseases*".

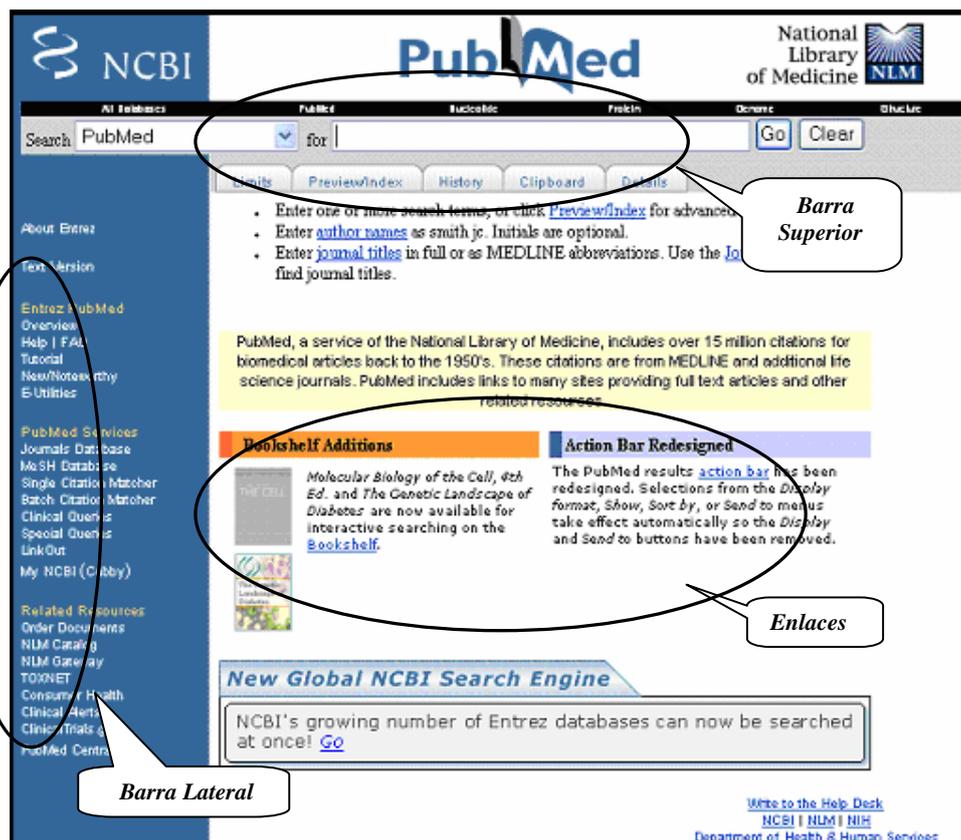
La instrucción *All Fields* obliga al algoritmo a buscar coincidencias en los campos que integran el formato MEDLINE. La tabla siguiente muestra algunos campos de dicho formato.

Campos del formato MEDLINE		
Tag	Nombre	Descripción
AB	Abstract	Resumen
AD	Affiliation	Filiación Institucional y dirección del primer autor
AU	Autor Name	Nombre de los autores
CY	Country	País de publicación de una revista
DP	Publication Date	Fecha en la que el artículo fue editado
EDAT	Entrez Date	Fecha en la que se incorporó en PubMed.
ID	Identification Number	Número que designa los trabajos financiados por la Agencia Americana del Servicio Público de Salud
IS	ISSN	Número de identificación unívoco de una revista.
JC	Journal Title Code	Código de identificación único compuesto de tres caracteres que adjudica Medline.
JID	NLM Unique ID	Número de identificación de revistas en el catálogo de la Biblioteca Nacional de Medicina.
LA	Language	Idioma del artículo
MH	MeSH Terms	Descriptor o encabezados
MHDA	MeSH Date	Fecha en la que el término MeSH fue incorporado a

		la cita
PG	Page Number	Páginas del artículo
PMID	PubMed Unique Identifier	Número de identificación unívoco asignado a cada registro Pubmed
PT	Publication Type	Tipo de artículo
RN	EC/RN Number	Número asignado por la Comisión de Encimas o por el Servicio de Resumen Químicos.
TI	Title Words	Título del artículo
UI	MEDLINE Unique Identifier	Número unívoco asignado a cada registro Medline
VI	Volume	Volumen de la revista

Tabla 4: Algunos campos del formato MEDLINE

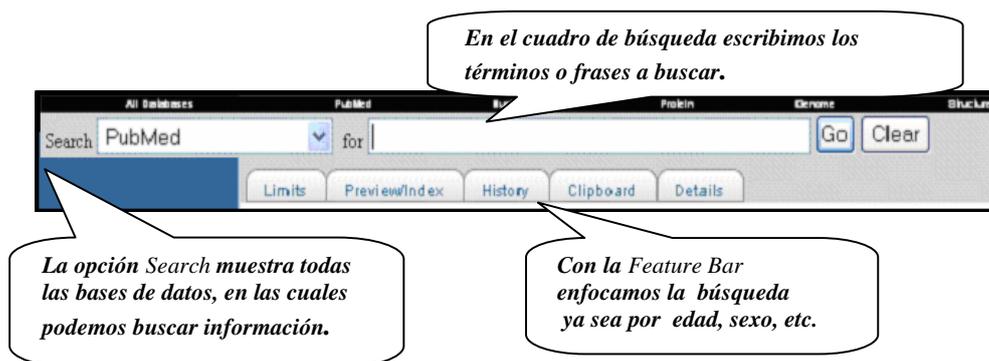
Ahora veamos algunos elementos que integran la página de Pubmed. En la figura siguiente están señalados tres elementos básicos para iniciar búsquedas y recuperaciones de citas bibliográficas:



La barra superior se utiliza para buscar términos, frases, autores, revistas, etc. La barra lateral ofrece servicios especializados de búsqueda como son el MeSH Databases, Bath Citation Matcher, Cubby, Fact Sheet Medline, Tutoriales, etc. Y por ultimo, está la opción de búsqueda a través de enlaces a todos los sitios Web de la Biblioteca Nacional de

Medicina. Algunos sitios son Pubmed Central, MedlinePlus y las bases de datos no bibliográficas de la NCBI, etc.

Veamos como funcionan estos tres elementos. En la barra superior solamente escribimos el término, la frase, el nombre del autor, etc. Y hacemos clic en Go para iniciar la búsqueda. En la figura siguiente están señaladas algunas opciones de la barra superior.



La barra lateral ofrece servicios especializados para todas aquellas personas que los requieran. Los servicios que ofrece se dividen en tres grupos: *Entrez Pubmed*, *Pubmed Services* y *Related Resources*. En la figura siguiente están señaladas algunas opciones de la barra lateral.



Enlaces a una amplia variedad de tutoriales para sacar el máximo provecho al sistema ENTREZ - PUBMED

Servicios para encontrar información detallada sobre descriptores, citas,

Enlaces a servicios disponibles en Pubmed Central, MedlinePlus®, TOXNET, etc.

Ahora veamos como salvar los resultados de una búsqueda. Cuando realizamos búsquedas por medio de cualquiera de los elementos anteriores, los resultados se muestran en una nueva página. Esta página contiene la barra de selección, la cual contiene todas las opciones necesarias para guardar nuestros resultados. Algunas opciones están señaladas en las dos figuras siguientes.

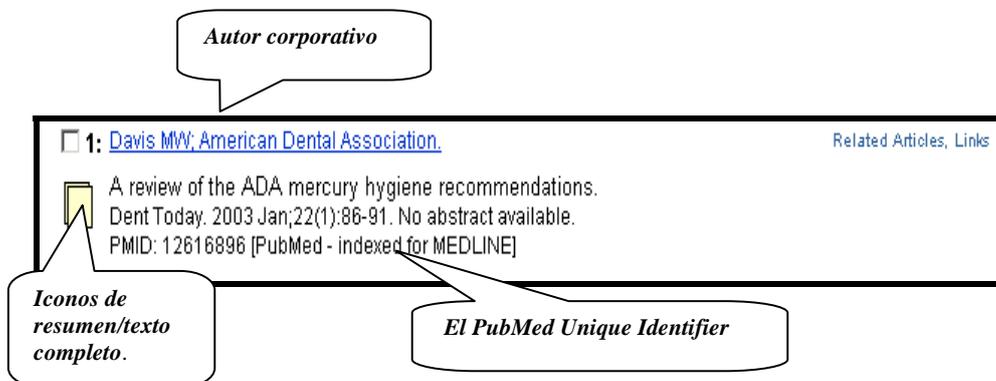
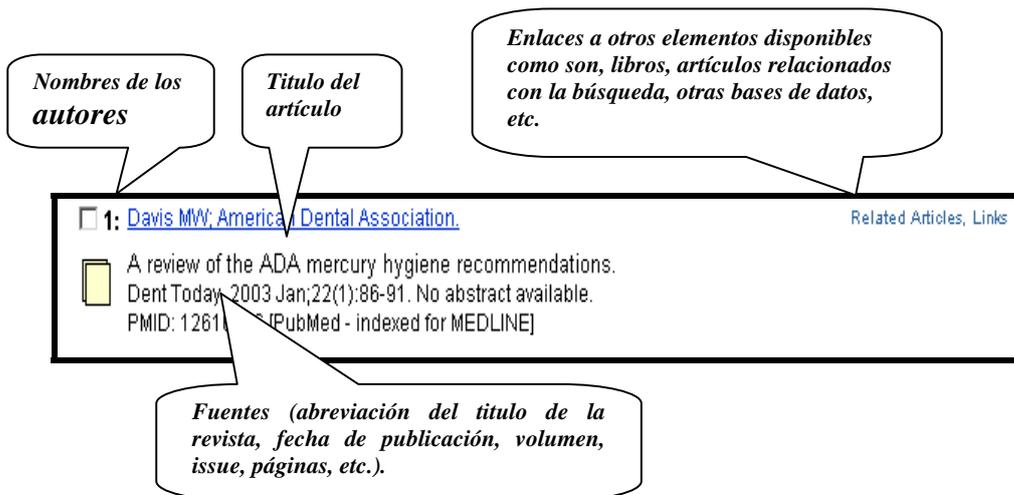


Clasificamos las citas

Opciones para guardar las



Los resultados de alguna búsqueda se muestran en el *formato Summary*. Éste expone los elementos bibliográficos que integran las citas en una forma sencilla, fácilmente entendible por cualquier usuario. Las dos figuras siguientes muestran algunos de estos elementos bibliográficos.



De todos estos elementos bibliográficos destacaremos solamente dos, *los iconos de resumen/texto completo* y el *PubMed Unique Identifier (PMID)*. El primero nos indica la presencia o ausencia del resumen, el cual es vital si deseamos comparar el contenido según el autor y el contenido según los términos del MeSH Vocabulary. La tabla siguiente muestra los iconos asociados para tal fin.

	Las citas no incluyen resumen.
--	--------------------------------

	Las citas incluyen resumen.
	El texto completo está disponible en PubMed Central (PMC).
	Hay un enlace al texto completo sin ser necesaria una suscripción.

Mientras que el *PubMed Unique Identifier (PMID)*, nos indica si la cita del artículo ha sido indizada con los términos del MeSH Vocabulary. La tabla siguiente muestra las etiquetas que acompañan al PMID:

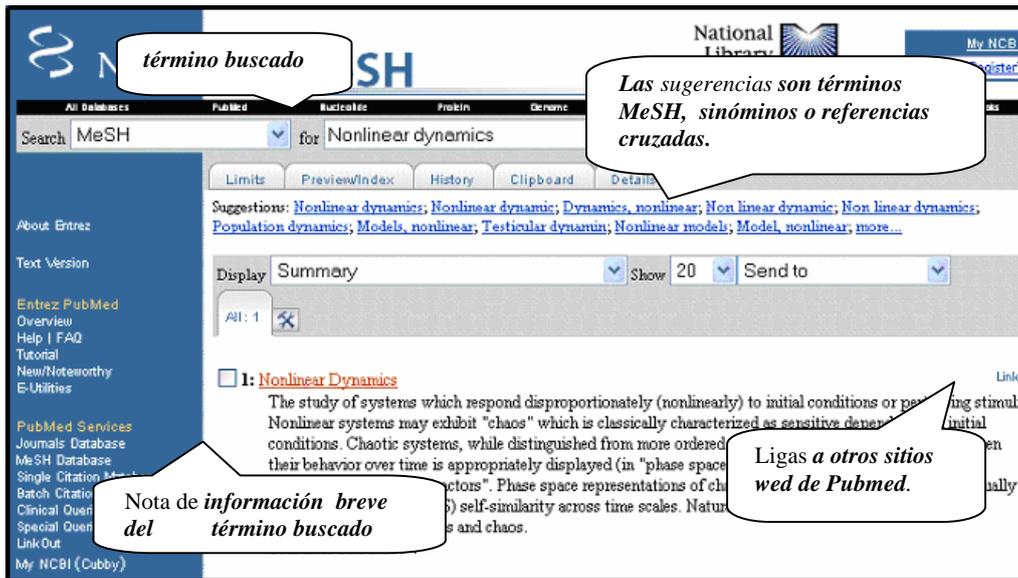
Publisher Supplied Citations	Son citas enviadas a PUBMED, por primera vez para su indización.
In Process	Son citas que están en el proceso de indización.
Indexed for MEDLINE	Son las citas indizadas exitosamente con el MeSH Vocabulary.
OLDMEDLINE for Pre1966	Son las citas localizadas en OLDMEDLINE provenientes de las décadas 1950s y 1960s. No están indizadas con el MeSH Vocabulary.
PubMed	Son las citas que no pudieron ser indizadas con el MeSH Vocabulary.

Ejemplo de Búsquedas

Finalmente veamos el ejemplo. La barra lateral ofrece el servicio *MeSH Database*, con este servicio se realizan búsquedas de citas indizadas con las palabras claves del MeSH Vocabulary. Para iniciar una búsqueda por medio del MeSH Database, en primer lugar debemos diseñar una estrategia de búsqueda, es decir seleccionar las palabras MeSH más idóneas, ya sea limitando las palabras a MeSH Major Topic, a encabezados MeSH o a Subencabezados MeSH o simplemente combinar encabezados MeSH con Subencabezados MeSH para concretar en uno o varios aspectos específicos el término deseado.

Los MeSH Major Topic son las palabras claves o combinaciones de palabras claves del MeSH Vocabulary que mejor se adaptan al contenido del artículo y son marcadas con un asterisco. La decisión de concederle a un MeSH ser un *major*, la toma el grupo de personas encargadas de la indización después de analizar el artículo.

Iniciemos: seleccionamos el MeSH Database en la barra lateral. Y en el cuadro de diálogo escribimos ***Nonlinear Dynamics*** (Dinámica No Lineal), por ejemplo, y hacemos clic en *Go*. En la imagen siguiente vemos los resultados obtenidos.



La frase que ingresamos es un MeSH, si no lo hubiera sido, el sistema nos habría arrojado los términos MeSH que estuvieran relacionados. Ahora hacemos clic en **Nonlinear Dynamics** para ver sus subencabezados, *Entry Terms* o sinónimos, etc., y su posición en el árbol del MeSH.

En la figura siguiente vemos una breve descripción del término **Nonlinear Dynamics**. Hay que destacar que este término no tiene subencabezados de algún tipo. Después vemos las tres opciones para limitar la búsqueda. Estas opciones son: *History*, *Restrict Search to Major Topic Headings Only*, y *Do Not Explode This Term*.

My NCBI (Cubby) Links

1: Nonlinear Dynamics

The study of systems which respond disproportionately (nonlinearly) to initial conditions or perturbing stimuli. Nonlinear systems may exhibit "chaos" which is classically characterized as sensitive dependence on initial conditions. Chaotic systems, while distinguished from more ordered periodic systems, are not random. When their behavior over time is appropriately displayed (in "phase space"), constraints are evident which are described by "strange attractors". Phase space representations of chaotic systems, or strange attractors, usually reveal fractal (FRACTALS) self-similarity across time scales. Natural, including biological, systems often display nonlinear dynamics and chaos.
Year introduced: 1994

Subheadings: This list includes those paired at least once with this heading in MEDLINE and may not reflect current rules for allowable combinations

history

Restrict Search to Major Topic headings only

Do Not Explode this term (i.e., do not include MeSH terms found below this term in the MeSH tree).

Entry Terms:

- . Dynamics, Nonlinear
- . Nonlinear Dynamic
- . Non-linear Dynamics
- . Dynamics, Non-linear
- . Non linear Dynamics
- . Non-linear Dynamic
- . Chaos Theory
- . Chaos Theories
- . Theories, Chaos
- . Theory, Chaos
- . Models, Nonlinear
- . Model, Nonlinear
- . Nonlinear Model
- . Nonlinear Models
- . Non-linear Models
- . Model, Non-linear
- . Models, Non-linear
- . Non linear Models
- . Non-linear Model

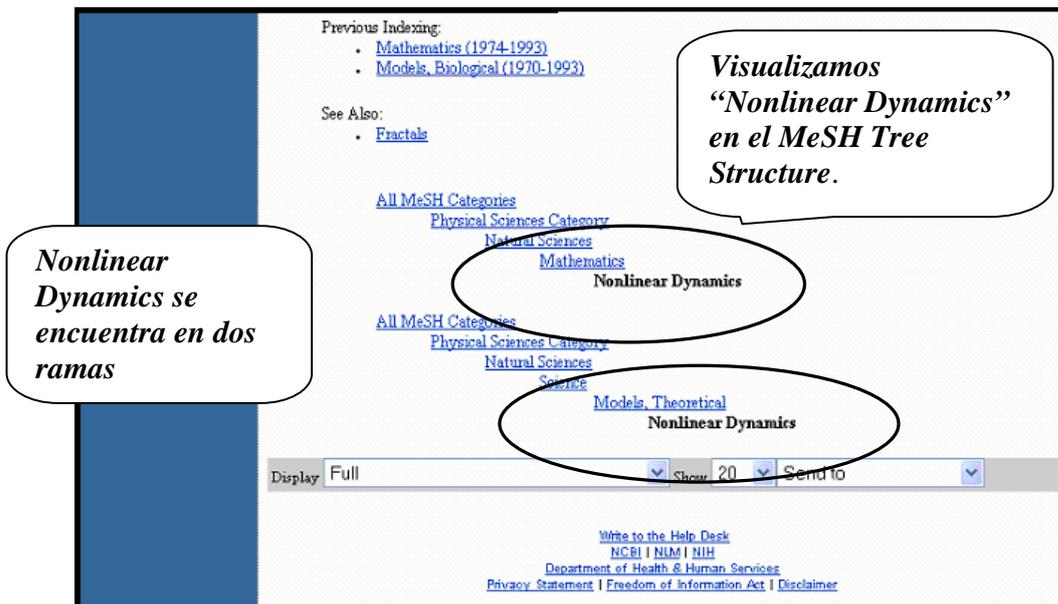
Sinónimos del

Activar alguna opción implica

Restringimos la búsqueda a "MeSH Major Topic" o activamos la opción "Explosión"

Como deseamos recuperar todas las citas que traten algún tema relacionado con **Nonlinear dynamics** no activaremos ninguna opción.

En esta figura vemos una parte del MeSH Tree Structure. Como se observa, para el NCBI, el término **Nonlinear Dynamics** pertenece a las Matemáticas, y a los Modelos Teóricos.



Además, vemos que, antes de ser aceptado el tema **nonlinear dynamics como un MeSH se indizaba como mathematics (1974-1993) y/o Models, Biological (1970-1993)**.

Como siguiente paso seleccionamos la casilla de chequeo y en el *combobox* seleccionamos a su vez, enviar a la caja de búsqueda con un AND (Search Box with AND). Esto nos permitirá agregar a nuestra petición buscar aquellas citas que tienen como MeSH **nonlinear dynamics**. En nuestro caso, es la primera sentencia de búsqueda que agregamos. La figura siguiente muestra lo anteriormente dicho.

Limits Preview/Index History Clipboard Details

Display Full Show 20 Send to

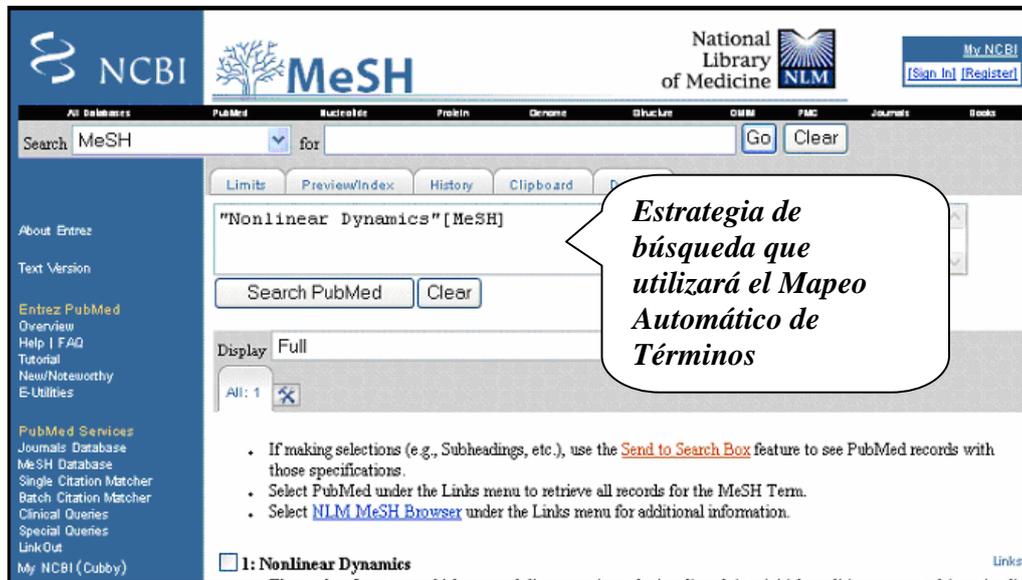
All: 1

- If making selections (e.g., Subheadings) to Search Box feature to see PubMed records with those specifications.
- Select PubMed under the Links menu for the MeSH Term.
- Select [NLM MeSH Browser](#) under the Links menu for additional information.

1: **Nonlinear Dynamics** [Links](#)

The study of systems which respond disproportionately (nonlinearly) to initial conditions or perturbing stimuli. Nonlinear systems may exhibit "chaos" which is classically characterized as sensitive dependence on initial conditions. Chaotic systems, while distinguished from more ordered periodic systems, are not random. When their behavior over time is appropriately displayed (in "phase space"), constraints are evident which are described by "strange attractors". Phase space representations of chaotic systems, or strange attractors, usually

Como resultado obtenemos la creación automática de la sentencia de búsqueda, la cual se agrega a la caja de búsqueda. En la figura siguiente vemos la sintaxis de búsqueda “**Nonlinear Dynamics**” [MeSH] que empleará el mapeo automático de términos.



Estrategia de búsqueda que utilizará el Mapeo Automático de Términos

- If making selections (e.g., Subheadings, etc.), use the [Send to Search Box](#) feature to see PubMed records with those specifications.
- Select PubMed under the Links menu to retrieve all records for the MeSH Term.
- Select [NLM MeSH Browser](#) under the Links menu for additional information.

1: Nonlinear Dynamics [Links](#)

Y finalmente, hacemos clic en el botón con etiqueta *Search PubMed* para iniciar la búsqueda de todas las citas que traten sobre **Nonlinear Dynamics**.

En la figura vemos los resultados en el *formato predeterminado*. Además, vemos que se encontraron un total de 4,220 citas sobre **Nonlinear dynamics**.

The screenshot shows the PubMed search results for "Nonlinear Dynamics" [MeSH]. The search bar at the top contains the query. Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Display" dropdown is set to "Summary", and the "Show" dropdown is set to "20". The search results show 4,220 items and 327 reviews. The first four results are listed, each with a checkbox, a citation number, the author name, and the full citation text including the journal name, volume, issue, pages, and PMID. The first result is by Tschumperle D, Deriche R. The second is by Fullgrabe C, Maillet D, Moroni C, Belin C, Lorenzi C. The third is by Park JH, Huh SH, Kim SH, Seo SJ, Park GT. The fourth is by Hayakawa T, Haddad WM, Hovakimyan N, Chellaboina V. There are two speech bubble annotations in Spanish: one pointing to the search results area saying "Visualizamos las citas sobre 'Nonlinear Dynamics'" and another pointing to the "Send to" button saying "Ya se encuentra indizada".

De esta forma recuperamos 4,220 citas que tratan algún tema relacionado con **Nonlinear dynamics**. Ahora debemos guardar estas citas en nuestra computadora. Para ello, en primer lugar, seleccionamos el *formato MEDLINE* disponible al lado derecho de *Display*. En segundo lugar, seleccionamos *File* y hacemos clic en *Send to*. Se abrirá un cuadro de dialogo en donde debemos dar un nombre y una ubicación a nuestro archivo. Con esto terminamos la etapa de adquisición de datos en este caso, recuperación bibliográfica.