Taller sobre calidad en bases de datos sobre biodiversidad

Aula de informática del Real Jardín Botánico (CSIC)
Madrid, 21-22 noviembre 2009

USO DE TESAUROS Y OTROS VOCABULARIOS CONTROLADOS

María Encinas
Unidad de Coordinación de GBIF España

ESQUEMA

- Introducción: los lenguajes documentales
  - ¿Qué son los lenguajes documentales?
  - Funciones, origen y tipos de los lenguajes documentales: los lenguajes controlados y los tesauros
- Características generales de los tesauros
  - Definición
  - Beneficios derivados del uso de tesauros en relación con las bases de datos
  - Estructura y elementos de un tesauro
  - Regulación de los tesauros
  - El proceso de elaboración de un tesauro
- Los tesauros y la calidad de las bases de datos sobre biodiversidad
  - El proceso de captura de datos
  - Ámbitos afectados y ejemplos de tesauros
    - Taxonómico (¿Qué?)
    - Geográfico (¿Dónde?)
    - Autoría y tiempo (¿Quién?, ¿Cuándo?)
    - Metodología y metadatos (¿Cómo?)

Introducción: los lenguajes documentales.

- ¿Qué son los lenguajes documentales?
  - Lenguaje documental: “Todo sistema artificial de signos normalizados que facilitan la representación formalizada del contenido de los documentos para permitir la recuperación, manual o automática de información solicitada por los usuarios” Blanca Gil Urdiciain
- Los tesauros son un tipo de lenguaje documental, de carácter controlado.

Introducción: los lenguajes documentales.

- **Funciones**, origen y tipos de lenguajes documentales
  - Lenguaje puente entre las informaciones contenidas en los documentos y las informaciones solicitadas por los usuarios.
  - Tienen que ver con los procesos de indización.
  - Diversificación terminológica del concepto “lenguaje documental”. No existe un vocabulario comúnmente aceptado para designar el concepto. La expresión más utilizada junto a “lenguajes documentales” es lenguajes de indización.

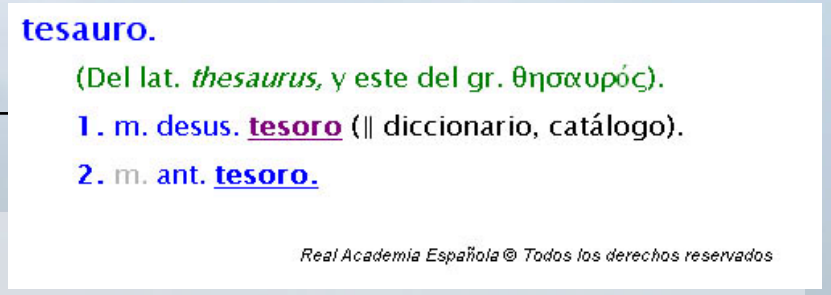
Introducción: los lenguajes documentales.

- Funciones, **origen** y tipos de lenguajes documentales
  - La historia de los lenguajes documentales está asociada a las bibliotecas y a la biblioteconomía: cuando la información depositada en las bibliotecas se hace tan grande que se es necesario organizarla (estandarizar y normalizar), no sólo ordenarla, para poder localizar la información.
  - No fue hasta finales del siglo XIX cuando se fragua el concepto moderno de lenguaje documental con las listas de encabezamiento de materia de Cutter y las clasificaciones (Clasificación decimal) de Dewey.

Introducción: los lenguajes documentales.

- Funciones, origen y **tipos** de lenguajes documentales
  - El lenguaje documental se puede clasificar en función de dos conceptos: control y coordinación.
    - Según el control
      - **Lenguajes libres**. Ej. Listas de palabras clave.
      - **Lenguajes controlados**. Ej. Tesauros.
    - Según la coordinación. Hace referencia al momento en el que se combinan los elementos que lo componen.
      - **Precoordinado**: cuando los elementos se combinan en el momento de la descripción. Ej. Clasificaciones y listas de encabezamiento.
      - **Postcoordinado**: cuando los elementos se combinan en el momento de la recuperación de información. Ej. Tesauros.
  - Los tesauros serían un tipo de lenguaje o vocabulario documental controlado y de carácter postcoordinado.

Características generales de los tesauros



tesauro.

(Del lat. *thesaurus,* y este del gr. θησαυρός).

1. m. desus. **tesoro** (‖ diccionario, catálogo).

2. m. ant. **tesoro.**

- Definición
  - Etimológicamente deriva del latín *thesaurus* y este a su vez del griego y significa *tesoro*.
  - "Vocabulario controlado y estructurado formalmente, formado por términos que guardan entre sí relaciones semánticas: de equivalencia, jerárquicas y asociativas. Se trata de un instrumento de control terminológico que permite convertir el lenguaje natural de los documentos en un lenguaje controlado, ya que se representa de manera unívoca el contenido de éstos, con el fin de servir tanto para la indización como para la recuperación de estos documentos" M.J. Lamarca
  - Normalmente se aplican a un dominio particular del conocimiento.

Características generales de los tesauros

- Beneficios del uso de los tesauros en relación con las bases de datos
  - Aumento de la **calidad de los datos almacenados**, pues aumentan los usos potenciales que se pueden dar a la información.
  - Aumento en la **interoperabilidad** de las bases de datos. Al cotejarse distintos grupos de datos de la misma naturaleza con los mismos parámetros, las posibilidades de realizar análisis conjuntos de los datos aumenta.
  - Se crean **bases de datos de conocimiento** que pueden reutilizarse en otras disciplinas
  - Facilitan la **consulta** y uso sistemático de datos, si se integran en los sistemas de explotación y consulta.

Características generales de los tesauros

- Estructura y elementos de un tesauro
  - Los tesauros se componen de unidades léxicas denominadas
    - **Descriptores o términos preferidos.**Expresiones o unidades lingüísticas que expresan conceptos. Un concepto se expresa con un único término y ese término responde a un único concepto.
    - **No descriptores o términos no preferidos**. Son sinóminos o cuasisinónimos de los descriptores. Son términos prohibidos, no se usan para la recuperación de información pero cada uno de ellos reenvía a un descriptor para representar los conceptos correspondientes.
  - Otros elementos: cualificadores y notas de alcance

Características generales de los tesauros

- Estructura y elementos de un tesauro
  - Las relaciones que se pueden dar entre los términos de un tesauro son:
    - **Relaciones de equivalencia.** Entre descriptores y no descriptores.
      - Operadores: UP, USE/UF, USE. Ej.
    - **Relaciones jerárquicas**. Es la relación vertical entre los descriptores de una misma clase, expresado en términos de subordinación de los conceptos.
      - Operadores: TG, TE/BT, NT. Ej.
    - **Relaciones asociativas**. Indican relaciones simétricas entre descriptores (uniones en la significación de los descriptores)
      - Operadores: TR/RT. Ej.

Ejemplo de relaciones jerárquicas

- TG Armas
- TE Armas blancas

- TG Genero
- TE Especie

Ejemplo de relaciones asociativas

- Política
- TR Políticos

Ejemplo de relaciones de equivalencia

- Comerciantes
  - UP Mercaderes
- Mercaderes
  - USE Comerciantes

- Ascensión vertical
  - UP Ascensor
  - UP Montacargas

Características generales de los tesauros

- Estructura y elementos de un tesauro
  - En la visualización de un tesauro en forma de listado, cada descriptor indicará, en general y si las tiene, las siguientes relaciones:
    - Notas de alcance
    - No Descriptores a los que sustituye (UP)
    - Términos genéricos (TG)
    - Términos específicos (TE)
    - Términos relacionados (TR)

Características generales de los tesauros

- Regulación de los tesauros
  - Distintos aspectos referentes a los tesauros se regulan a través de las Normas ISO y UNE
    - Tesauros monolingües ISO 2788-1986
    - Tesauros multilingües ISO 5964-1985
    - Directrices para el establecimiento de Tesauros monolingües UNE 50-106-90
    - Directrices para la creación y desarrollo de T. Multilingües UNE 50-125-1997
    - ANSI-ISO Z39.19.2003. Directrices para la construcción, formato y gestión de Tesauros monolingües.
  - Estas normas establecen formas de presentación de los términos del tesauro (unitérminos o términos compuestos), singular o plural, notas de alcance, etc.

Características generales de los tesauros

- El proceso de elaboración de un tesauro
  - Recogida de términos
  - Análisis de los términos: Agrupación por áreas y definición de los términos preferidos y no preferidos y establecimiento de las relaciones.
  - Revisión del tesauro

Los tesauros y la calidad de las bases de datos sobre biodiversidad

- El proceso de captura de datos

Captura y registro de los datos en el momento de la recogida/avistamiento

Manipulación de los datos previa a la digitalización

Identificación de la muestra y de su registro

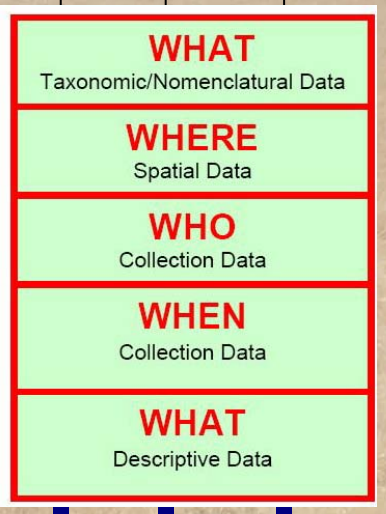
Digitalización de los datos

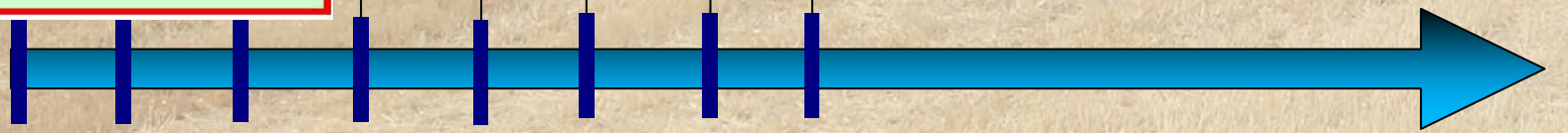
Documentación de los datos (metadatos)

Almacenamiento y archivo de los datos

Presentación de los datos y publicación

Analisis y manipulación de los datos (uso)





Origen: A. Chapman, 2005

Ámbitos afectados y ejemplos de tesauros

Listados taxonómicos

Species 2000 europa

GBIF

Referencias temporales

¿Qué?

¿Cuándo?

¿Dónde?

Registro Biológico

Directorios de investigadores/colectores

Nomenclator, *gazeteers*

IDEE

BioGeomancer

¿Quién?

¿Cómo?

Protocolos de recogida y tratamiento de datos

Ámbito taxonómico (¿Qué?)

- La nomenclatura y la taxonomía son ámbitos en los que la utilización de tesauros o listas de referencia mejor demuestra su valor.
- Muchas iniciativas tienen como objetivo la realización de listados taxonómicos, de los que nos podemos beneficiar a la hora de determinar y aumentar la calidad de nuestros datos.

Ámbito taxonómico (¿Qué?)

- Recursos globales

- **Species 2000**: Acceso a través de internet o en CD

http://www.sp2000.org/

- **ITIS**: *Integrated Taxonomic Information System*

http://www.itis.gov/

- **uBio**: *Universal Biological Indexer and Organizer*

http://www.ubio.org/

Ámbito taxonómico (¿Qué?)

- RECURSOS restringidos por ÁREA GEOGRÁFICA y/o GRUPO TAXONÓMICO

- **FAUNA EUROPAEA**

http://www.faunaeur.org/

- **FLORA EUROPAEA**

http://rbg-web2.rbge.org.uk/FE/fe.html

- **Euro+MED Plant Base**

http://www.emplantbase.org/

- **MarBef**

http://www.marbef.org/

Ámbito taxonómico (¿Qué?)

- RECURSOS restringidos por ÁREA GEOGRÁFICA y/o GRUPO TAXONÓMICO

- **FAUNA IBÉRICA: base de datos IBERFAUNA**

http://www.fauna-iberica.mncn.csic.es/

- ***FLORA IBERICA***

http://www.rjb.csic.es/floraiberica/

- **Vean también**

http://www.gbif.es/ProyBioEsp.php

http://www.gbif.es/recursos1.php

Geografía ¿Dónde?

- La multitud de ámbitos en los que se utiliza la INFORMACIÓN GEOGRÁFICA hace que los recursos donde se puede consultar esta información sean muy variados.
- Respecto a la información geográfica, los procesos más habituales a realizar son:
  - Comprobación de las localidades registradas (ortografía, etc.)
  - Asignación de datos geográficos precisos (coordenadas) a registros que carecen de esta información, lo que se denomina georreferenciación retrospectiva.

Geografía ¿Dónde?

- Recursos globales

http://www.biogeomancer.org/

http://www.museum.tulane.edu/geolocate/

Geografía ¿Dónde?

- Infraestructuras de datos espaciales:
  - Comprende los portales web, los servicios, los datos y metadatos y otro tipo de información geográfica que se ofrecen de manera integrada, en general asociada a un determinada área geográfica

Geografía ¿Dónde?

Infraestructura de Datos Espaciales de España – IDEE:
Incluye un servicio de nomenclátor, además de servidores de mapas y otros recursos geográficos.

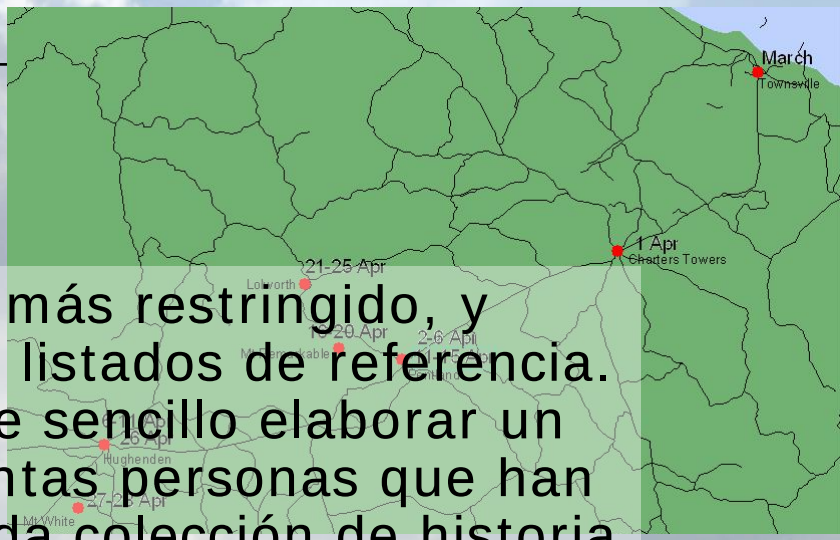
http://www.idee.es/

Otras infraestructuras de datos espaciales regionales:

- **Andalucía**: http://www.andaluciajunta.es/IDEAndalucia/IDEA.shtml
- **Asturias**: http://gis.princast.es/sitpacarto/
- **Cataluña**: http://www.geoportal-idec.net/geoportal/IDECServlet?idioma=cas
- **Castilla y León**: http://www.sitcyl.jcyl.es/
- **Galicia**: http://sitga.xunta.es/
- **Islas Canarias**: http://pre.sitcan.com/Visor/
- **Murcia**: http://www.sitmurcia.com/
- **Navarra**: http://idena.navarra.es/, http://sitna.tracasa.es/
- **La Rioja**: http://www.iderioja.org/

http://www.gbif.es/recursos1.php

Autoría y Tiempo: ¿Quién? ¿Cuándo?



- Es sin duda un campo mucho más restringido, y donde es más difícil encontrar listados de referencia. Sin embargo, es relativamente sencillo elaborar un listado restringido de las distintas personas que han contribuido en una determinada colección de historia natural, o proyecto de investigación: colectores, determinadores, etc.
- Es en trabajos sobre la historia de la biología o donde se analizan determinadas expediciones históricas o la biografía de un determinado investigador donde esta información puede adquirir gran relevancia: **cotejar los rangos temporales con la distribución geográfica de las muestras puede ayudarnos a localizar lagunas e inconsistencias**.

▲ *Karel Domin* (1882–1953), profesor botaniky Karlovy univerzity, světově proslulý odborník a nestovatel, autor monografií o Ordech a Kokořínsku.

Autoría y Tiempo: ¿Quién? ¿Cuándo?

Algunas referencias genéricas

- **Base de datos mundial de taxónomos**

  http://www.eti.uva.nl/tools/wtd.php

- **Informe de colecciones de historia natural en España** (BioCASE – GBIF España)

  http://www.gbif.es/ic_BusquedaPersonas.php

- **Index herbariorum**: a guide to the location and contents of the world's public herbaria, Part 2: Collectors. 7 volúmenes.

  http://sciweb.nybg.org/science2/IndexHerbariorum.asp

- García-Valdecasas, A., Bello, E. & Becerra, J M., 1994. **Directorio de Taxónomos (DIRTAX)**. *Graellsia*. Monografía nº 1: 1-233.

Metodología y metadatos: ¿Cómo?

- A nivel de METODOLOGÍA podemos registrar información a distintos niveles del proceso de tratamiento de los datos, si disponemos de la misma:
  1. Los métodos utilizados para construir cada juego de datos del sistema.
  2. Los atributos registrados en cada juego de datos o para cada unidad biológica.
  3. Los métodos que se han utilizado para obtener el valor de cada atributo.
  4. Las personas responsables del desarrollo y/o ejecución de estos métodos.

Metodología y metadatos: ¿Cómo?

- Esta información nos abre muchas posibilidades:
  - Permite realizar una evaluación de la precisión del juego de datos, dependiendo de la descripción de sus métodos y atributos.
  - Permite repetir una determinada metodología, y aceptar/rechazar los datos obtenidos con anterioridad.
  - Puede convertirse en una referencia para otros estudios similares.
  - Permite contactar con quien realiza este tipo de labores para obtener más información

Metodología y metadatos: ¿Cómo?

La página web del **Sistema de Información sobre Biodiversidad de Colombia** ofrece gran cantidad de información sobre métodología, pero también sobre los otros temas tratados en esta presentación:

http://www.siac.net.co/

Por último...

- Sólo mencionar que Las aplicaciones de software que se desarrollan en la Unidad de Coordinación de GBIF España disponen de herramientas de ayudas a la introducción de datos y de comprobación basadas en tesauros y vocabularios controlados.

http://www.gbif.es/recursos1.php

Los recursos utilizados a la hora de cotejar, corregir, ampliar… los datos, merecen el debido **reconocimiento** y el respeto a sus derechos de **propiedad intelectual**.

Algunos documentos y páginas web consultados

- Arano, S. y Codina. L. 2004**.** La estructura conceptual de los tesauros en el entorno digital: ¿nuevas esperanzas para viejos problemas? Jornades Catalanes d'informació i documentació.
- Chapman, A.D. Principles of Data Quality. 2005. Report for the Global Biodiversity Information Facility, Copenhagen.
- Craven, T. 2008. Thesaurus construction. Faculty of Information and Media Studies, The University of Western Ontario. http://publish.uwo.ca/~craven/677/thesaur/main00.htm (Consulta 17-11-2008)
- J. Paul Getty Trust. (web en línea). Thesaurus of Geographic Names Online. http://www.getty.edu/research/conducting_research/vocabularies/tgn/?find=Tajo&place=river&nation=Spain&prev_page=1&english=Y (Consulta 17-11-2008)
- Jiménez, A.G. 2004. Instrumentos de representación del conocimiento: tesauros versus ontologías. Anales de Documentación, 7:79-95.
- Lamarca Lapuente, M.J. 2008. Hipertexto, el nuevo concepto de documento en la cultura de la imagen. Tesauros. < http://www.hipertexto.info/documentos/tesauros.htm> (Consulta 17-11-2008)
- Martínez García, L. y García García-Castro, C. 2008. Universidad Carlos III de Madrid. http://es.geocities.com/ontologias_y_tesauros/introduccion_a_los_tesauros.htm (Consulta 17-11-2008)
- Méndez. E. 2002. Soporte a la construcción de Tesauros en Internet. Dto. De Biblioteconomía y documentación. Universidad Carlos III de Madrid. http://rayuela.uc3m.es/~mendez/tesauro.htm (Consulta 17-11-2008)
- National Biological Information Infraestructure. (web en línea). Biocomplexity Thesaurus. http://thesaurus.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1797&mode=2&in_hi_userid=2&cached=true (Consulta 17-11-2008)
- Queensland University of Technology. Brisbane, Australia. (web en línea) http://www.imresources.fit.qut.edu.au/vocab/ (Consulta 17-11-2008)
- Santana, O. y Mayor, O. 2000. Construcción de tesauros. http://protos.dis.ulpgc.es/docencia/seminarios/rit/Construccion_de_tesauros/ (Consulta 17-11-2008)
- Universidad de León. (web en línea) http://www3.unileon.es/dp/abd/tesauro/pagina/conceptos/conceptos.htm (Consulta 17-11-2008)

Taller sobre calidad en bases de datos sobre biodiversidad

Aula de informática del Real Jardín Botánico (CSIC)
Madrid, 21-22 noviembre 2009

Más información

http://www.gbif.es/

Muchas gracias

María Encinas
Unidad de Coordinación de GBIF España