Web semántica, el futuro de las bases de datos de las colecciones de historia natural y cómo “4UColl” responde a este escenario

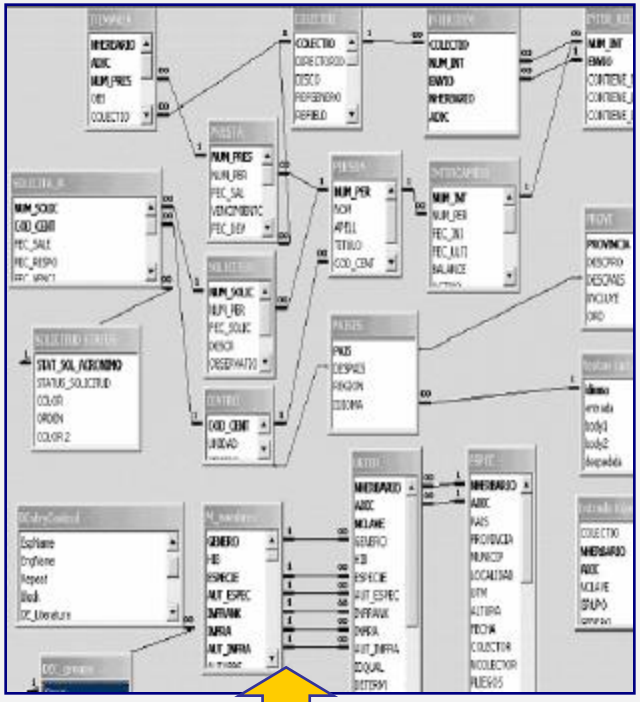
Francisco Pando

Guión

- Web semántica
- Datos enlazados

  (Los dioses antiguos y nuevos)
- Cambios en la red de GBIF y en contexto mundial en esa dirección
- Lo nuevo en 4UCOLL

Los dioses antiguos

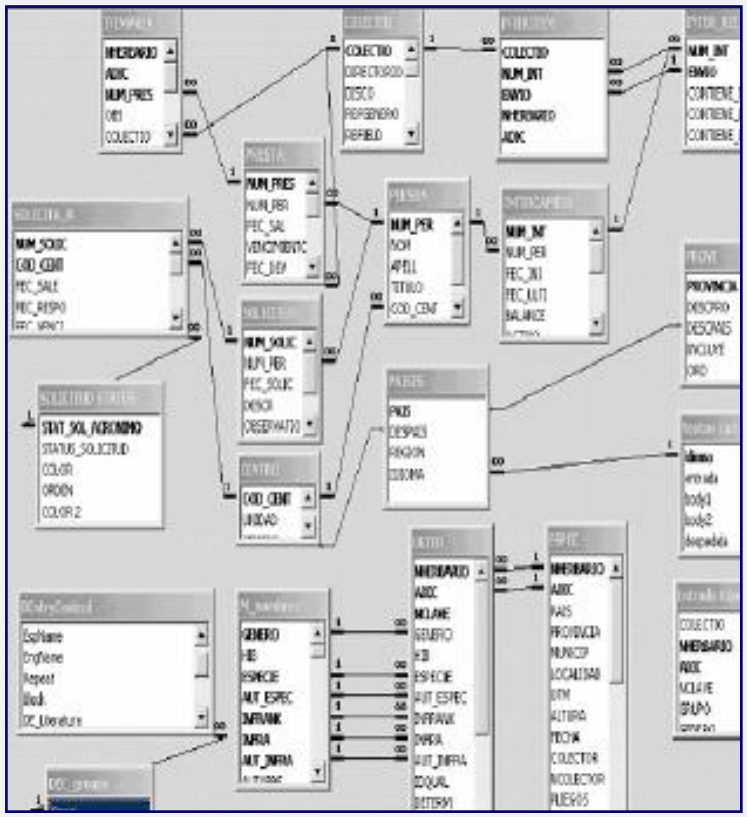


Attr#	Name
1	ScientificName
2	Kingdom
3	Phylum
4	Class
5	Order
6	Family
7	Genus
8	Species
9	Subspecies
10	InstitutionCode
11	CollectionCode
12	CatalogNumber
13	Collector
14	Year
15	Month
16	Day
17	Country
18	StateProvince
19	County
20	Locality
21	Longitude

CatalogNumber	Unique identifier for the specimen record within
Collector	The name of the collector or collectors that were observation) from the field.
Year	The year (four digit) in which the specimen was
Month	The month of the year (1..12) in which the speci
Day	The day of month that the specimen was collec
Country	The country or major political unit (ocean) from
StateProvince	The state, province or region (i.e. next political r collected.
County	The county (or shire, or next political region sma
Locality	The locality description (place name plus option specimen was collected.

La compatibilización entre sistemas se hace “a mano”

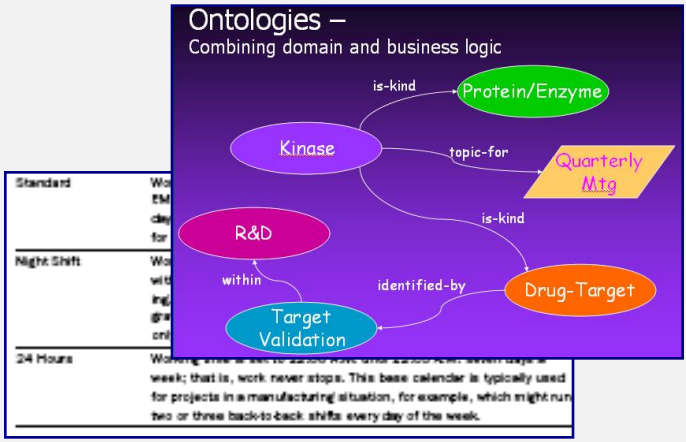
Los dioses nuevos:
Web semántica: comunicar contenidos, no documentos



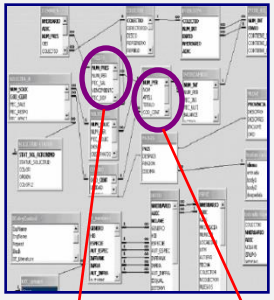
Las definiciones de las bases de datos se almacenan (o hacen uso) en un sistema:

- ***Accesible vía Web***
- ***Ininteligible por "maquinas"***
- ***Relaciona conceptos, relaciones e instancias***

→ *ontologías*



Web semántica: Triple storage, RDF*



Sujeto	*Predicado*	*objeto*
25322	genero	Inga
25322	especie	alba
25322	pais	COL
COL	Pais largo	Colombia

XML

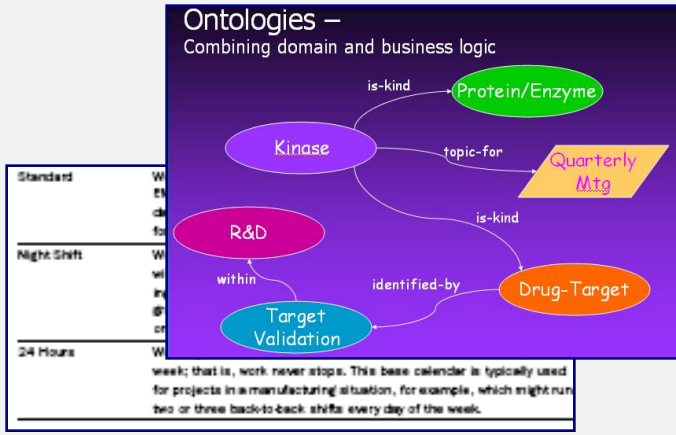
Especimenes

Campo	*valor*
Nmr_ejemplar	***25322***
genero	*Inga*
especie	alba
pais	COL

URIs

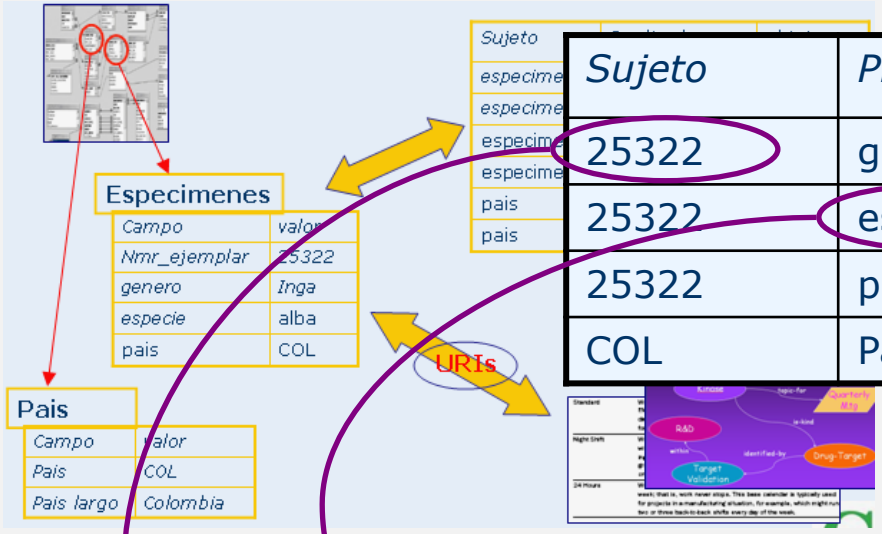
Pais

Campo	*valor*
Pais	***COL***
Pais largo	*Colombia*



**Resource Description Framework*

Web semántica: ídentificadores únicos



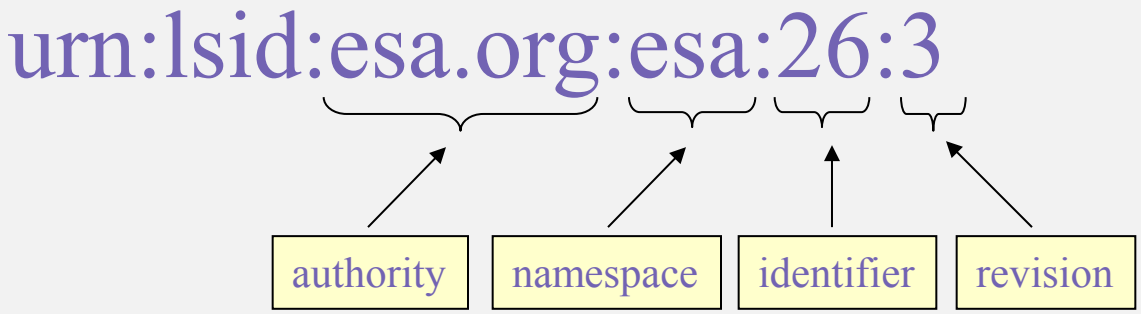
Sujeto	*Predicado*	*Objeto*
25322	genero	Inga
25322	especie	alba
25322	pais	COL
COL	Pais largo	Colombia

- ***Identificadores únicos de conceptos (ontología)***
- ***Identificadores únicos de objetos***

Web semántica: Identificadores únicos de objetos> LSIDs

Life Science Identifiers (LSID) son un tipo de identificadores globales

- Únicos
- Universales
- Se pueden resolver (esto es, los entienden las máquinas)
- Se pretende que sean persistentes
- Facilitar comunicación e intercambio
- Más robustos que URL
- Con metadatos
- Arquitectura que permite descubrir datos de fuentes independientes



Accepted by N. Evenhuis: 14 Apr. 2009; published: 14 Jul. 2009

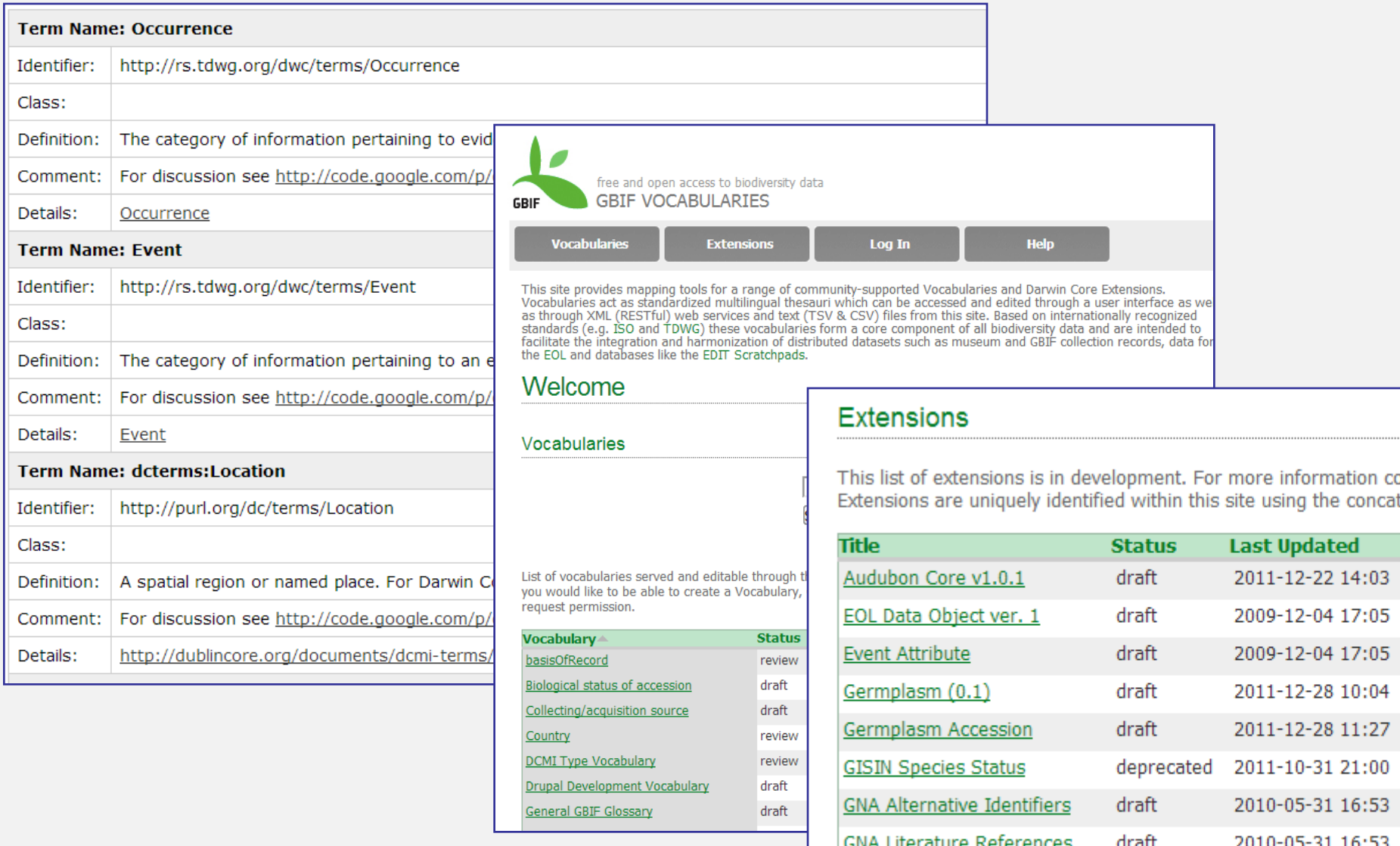
urn:lsid:zoobank.org:pub:E585D33C-262E-4796-A834-5FEAD656FE1F

Otros identificadores

- DOIs
  - p.ej.: http://doi.org/10.15470/gprffz
  - Centralizado
  - Metadatos no estandarizados

- UUIDs – 128 bit string, guaranteed unique
  - p.ej.: 58f202ac-22cf-11d1-b12d-002035b29092
  - Sin metadatos
  - Sin resolución

Web semántica: Identificadores únicos de conceptos > URIs; Vocabularios controlados



Term Name: Occurrence	
Identifier:	http://rs.tdwg.org/dwc/terms/Occurrence
Class:	
Definition:	The category of information pertaining to evid
Comment:	For discussion see http://code.google.com/p/
Details:	Occurrence
Term Name: Event	
Identifier:	http://rs.tdwg.org/dwc/terms/Event
Class:	
Definition:	The category of information pertaining to an e
Comment:	For discussion see http://code.google.com/p/
Details:	Event
Term Name: dcterms:Location	
Identifier:	http://purl.org/dc/terms/Location
Class:	
Definition:	A spatial region or named place. For Darwin C
Comment:	For discussion see http://code.google.com/p/
Details:	http://dublincore.org/documents/dcmi-terms/

Ontologías

An ontology is a way to represent knowledge, by describing the types or classes of entities within a given domain and the relationships among them. By providing standardized definitions for the terms used by scientists to represent these classes, and by defining the logical relationships among these terms, ontologies make information about content explicit for computers, allowing them to discover common meaning in diverse data sets. Thus, ontologies are an important component of many bioinformatics applications (Jensen and Bork, 2010), and they form the foundation of the semantic web (Berners-Lee et al.,

American Journal of Botany

American Journal of Botany 99(8): 1263–1275. 2012.

SPECIAL INVITED PAPER

ONTOLOGIES AS INTEGRATIVE TOOLS FOR PLANT SCIENCE[1]

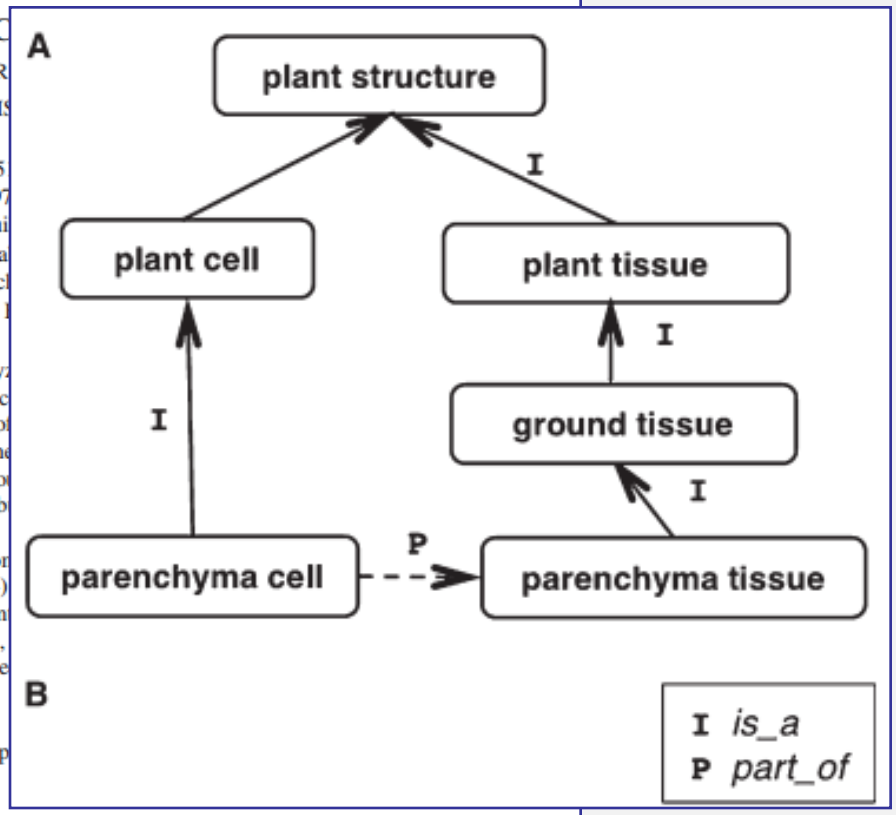
RAMONA L. WALLS[2,9], BALAJI ATHREYA[3], LAUREL C
MARIA A. GANDOLFO[4,9], PANKAJ JAISWAL[3,9], CHRISTOPHER
STEFAN RENSING[6], BARRY SMITH[7], AND DENNIS

[2]New York Botanical Garden, 2900 Southern Blvd., Bronx, New York 10458-5
Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, Oregon 97
Department of Plant Biology, Cornell University, 412 Mann Library Bui
[5]Berkeley Bioinformatics Open-Source Projects, Lawrence Berkeley National La
Berkeley, California 94720 USA; [6]Faculty of Biology, University of Freiburg, Sc
[7]Department of Philosophy, University at Buffalo, 126 Park Hall,

- *Premise of the study:* Bio-ontologies are essential tools for accessing and analyz and phenomic data. Ontologies provide structured vocabularies to support c framework for automated analyses and reasoning. They are a key component o
- *Methods:* This paper provides background on what bio-ontologies are, why th ontology development. It includes an overview of ontologies and related reso detailed description of the Plant Ontology (PO). We discuss the challenges of b (Viridiplantae).
- *Key results:* Ontologies can advance plant science in four keys areas: (1) cor development; (2) taxonomy and systematics; (3) semantic applications; and (4)
- *Conclusions:* Bio-ontologies offer a flexible framework for comparative plan standing. As genomic and phenomic data become available for more species, ontology terms will become less centralized, while at the same time, the nee common, causing more researchers in plant science to turn to ontologies.

Key words: bio-ontologies; genome annotation; OBO Foundry; phenomics; p plant systematics; semantic web.

Data overload is an issue for nearly every branch of plant growing larger and more complex. All this information creates



- http://www.amjbot.org/content/99/8/1263.full.pdf+html

TABLE 1. Ontologies and other related resources for plant science.

Resource (abbreviation)	Domain	References
Plant Ontology (PO)	Plant anatomical entities and plant structure development stages	(Pujar et al., 2006; Ilic et al., 2007)
Gene Ontology (GO)	Cellular components, biological processes, and molecular functions	(Gene Ontology Consortium, 2009) http://www.geneontology.org/
Chemical Entities of Biological Interest (ChEBI)	Molecular entities that are natural products or are synthetic products used to intervene in the processes of living organisms	(Degtyarenko et al., 2007; de Matos et al., 2009) http://www.ebi.ac.uk/chebi/
Protein Ontology (PR)	Proteins based on evolutionary relatedness, protein forms produced from a given gene locus, and protein-containing complexes	(Natale et al., 2007; Bult et al., 2011) http://pir.georgetown.edu/pro/
Ontology for Biomedical Investigations (OBI)	Scientific investigations, including the protocols and instrumentation used, the material used, the data generated, and the types of analysis performed	(Brinkman et al., 2010) http://obi-ontology.org
Phenotypic Quality Ontology (PATO)	Phenotypic qualities (properties). This ontology can be used in conjunction with other ontologies such as anatomical ontologies to refer to phenotypes.	(Mungall et al., 2010) http://obofoundry.org/wiki/index.php/PATO:Main_Page
Plant Trait Ontology (TO)	Phenotypic traits in plants; each trait is a distinguishable feature, characteristic, or quality of a plant	(Jaiswal, 2011) http://www.gramene.org/db/ontology/search?id=TO:0000387
Plant Infectious Disease Ontology (IDOPlant)	Plant infectious diseases, pathogens, and symptoms	(Walls et al., in press) http://purl.obolibrary.org/obo/idoplant.owl.
Extensible Observation Ontology (OBOE)	A suite of ontologies for modeling and representing scientific observations	(Madin et al., 2007) https://semtools.ecoinformatics.org/oboe
Environment Ontology (EnvO)	Environmental features and habitats	http://environmentontology.org/
NCBI Taxonomy	Biological taxa, based on the classification of the National Center for Biotechnology Information	(Wheeler et al., 2007) http://obofoundry.org/cgi-bin/detail.cgi?id=ncbi_taxonomy
BioPortal	Source for finding, searching and querying bio-ontologies	http://bioportal.bioontology.org/
Ontology Lookup Service	Source for finding and searching bio-ontologies	(Côté et al., 2006) http://www.ebi.ac.uk/ontology-lookup/
OntoBee	Source for finding, searching and querying bio-ontologies	(Xiang et al., 2011) http://ontobee.org

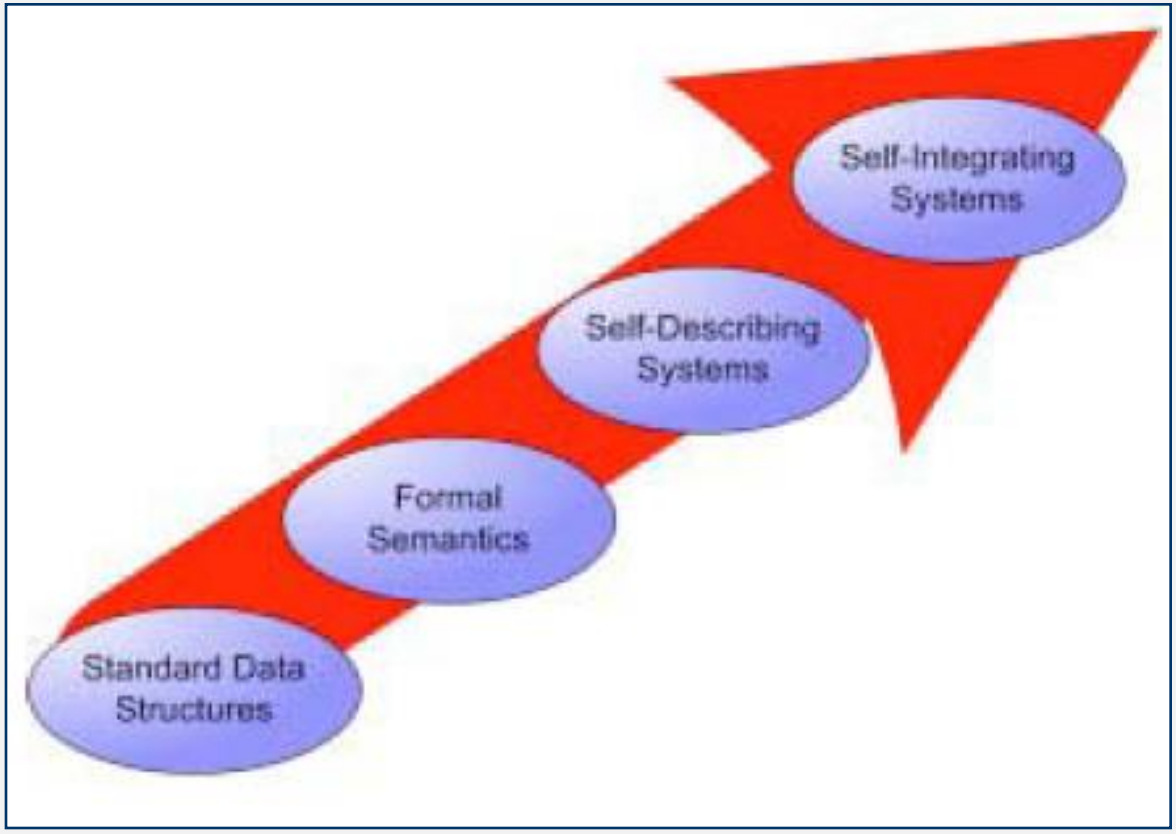
Ontologias hoy

- Retos:
  - Falta de vocabularios completos que cubran grupos amplios de organismos (p.e. plantas)
  - No todas las entidades observables pueden adscribirse “limpiamente” a las categorías establecidas (p.e. que es un septo de un fruto)
  - Homologías discutibles
  - Pérdida de detalle en las generalizaciones
  - Etc.

Web semántica: navegar los datos como el metro de una ciudad

- ***Bases de datos conectadas por ontologías y la Web***
- ***Conseguir que podamos inferir hechos a partir de datos distribuidos***

Hacia donde vamos



La idea es que la web funciona como un único sistema de información

XML

Los contenidos se tienden a codificar usando un metalenguaje llamado XML (Extensible markup Language .

La definiciones y relaciones se expresan usando una implementación de XML llamada XSD (*XML Schema Definition)*

XML

Documento XML

```
<?xml version="1.0" encoding="UTF-8"?>
<correo xmlns="http://www.idee.es" xmlns:xsi="http://www
instance"
xsi:schemaLocation="http://www.idee.es/correo.xsd">
        <origen>alice@mfom.es</origen>
        <destino>bob@mfom.es<destino>
        <asunto>Avances IDEE</asunto>
        <texto>La IDE de Andalucía se incorporará a la IDEE en breve.</texto>
</correo>
```

Esquema XML

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.idee.es"
xmlns=http://www.idee.es" elementFormDefault="qualified"
<xs:element name="correo">
  <xs:complexType>
   <xs:sequence>
        <xs:element name="origen" type="xs:string"/>
        <xs:element name="destino" type="xs:string"/>
        <xs:element name="asunto" type="xs:string"/>
        <xs:element name="cuerpo" type="xs:string"/>
   </xs:sequence>
  </xs:complexType>
</xs:element></xs:schema>
```

Web semántica en acción: “Linked (open) data”

¿Qué es Linked Data?

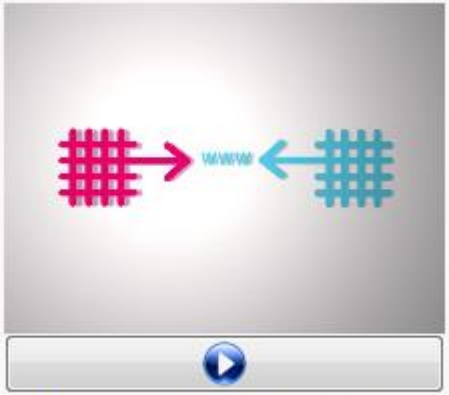
Los Datos Enlazados es la forma que tiene la Web Semántica de vincular los distintos datos que están distribuidos en la Web, de forma que se referencian de la misma forma que lo hacen los enlaces de las páginas web.

La Web Semántica no se trata únicamente de la publicación de datos en la Web, sino que éstos se pueden vincular a otros, de forma que las personas y las máquinas puedan explorar la web de los datos, pudiendo llegar a información relacionada que se hace referencia desde otros datos iniciales.

¿Cómo funciona?

Los Datos Enlazados, como parte de la Web Semántica, se basa en la aplicación de ciertos principios básicos y necesarios, que fomentarán el crecimiento de la Web, tanto a nivel de los documentos HTML (vista clásica de la Web), como a nivel de los datos expresados en RDF (vista de la Web Semántica).

1. Usar URIs para identificar las cosas
2. Usar URIs HTTP
3. Ofrecer información sobre los recursos usando RDF
4. Incluir enlaces a otros URIs



An introductory overview of Linked Open Data in the context of cultural institutions.

- http://www.youtube.com/watch?v=uju4wT9uBIA

Linked Data and the Charm of Weak Semantics
Introduction: The Strengths of Weak Semantics

by Thomas Baker and Stuart A. Sutton

Bulletin of the Association for Information Science and Technology – April/May 2015 – Volume 41, Number 4

Linked Data and the Charm of Weak Semantics

EDITOR'S SUMMARY

Logic and precision are fundamental to ontologies underlying the semantic web and, by extension, to linked data. This special section focuses on the interaction of semantics, ontologies and linked data. The discussion presents the Simple Knowledge Organization Scheme (SKOS) as a less formal strategy for expressing concept hierarchies and associations and questions the value of deep domain ontologies in favor of simpler vocabularies that are more open to reuse, albeit risking illogical outcomes. RDF ontologies harbor another unexpected drawback. While structurally sound, they leave validation gaps permitting illogical uses, a problem being addressed by a W3C Working Group. Data models based on RDF graphs and properties may replace traditional library catalog models geared to predefined entities, with relationships between RDF classes providing the semantic connections. The BIBFRAME Initiative takes a different and streamlined approach to linking data, building rich networks of information resources rather than relying on a strict underlying structure and vocabulary. Taken together, the articles illustrate the trend toward a pragmatic approach to a Semantic Web, sacrificing some specificity for greater flexibility and partial interoperability.

KEYWORDS

linked data
RDF
semantic networks
ontologies

Thomas Baker, an organizer of the Dublin Core Metadata Initiative, is an associate professor at Sungkyunkwan University in Seoul, South Korea. He can be reached at tb12<at>thbaker.org.

Stuart A. Sutton, associate professor emeritus in the Information School of the University of Washington, is managing director of the Dublin Core Metadata Initiative. He can be reached at sasutton<at>uw.edu.

When the meme first emerged in the late 1990s, *Semantic Web* stood for logical data processing on the foundation of World Wide Web technology. One of its roots reached back to the 1955 meme of *artificial intelligence*, with its notion "that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." [1, p. 12] The Semantic Web specifications developed by the World Wide Web Consortium from the late 1990s through the mid-2000s – the Resource Description Framework (RDF) and the Web Ontology Language (OWL) – were anchored in the notion of *ontology* as a "formal, explicit specification of a shared conceptualization" as supported by the field of ontology engineering.

Around 2006, Semantic Web was joined by the related, but more accessible and ultimately more popular meme of *linked data*. Starting with a cluster of databases linked to and from Wikipedia, the linked data movement took a more inclusive view of data technologies, with data serialized for Semantic Web-based interoperability as the five-star summit that providers of data in proprietary or application-specific document, database and record formats could by incremental steps ascend.

The contributions to this issue of the *Bulletin of the Association for Information Science and Technology* address, from five perspectives, how the shift to the idea of linked data at scale has changed the role of semantically precise ontologies.

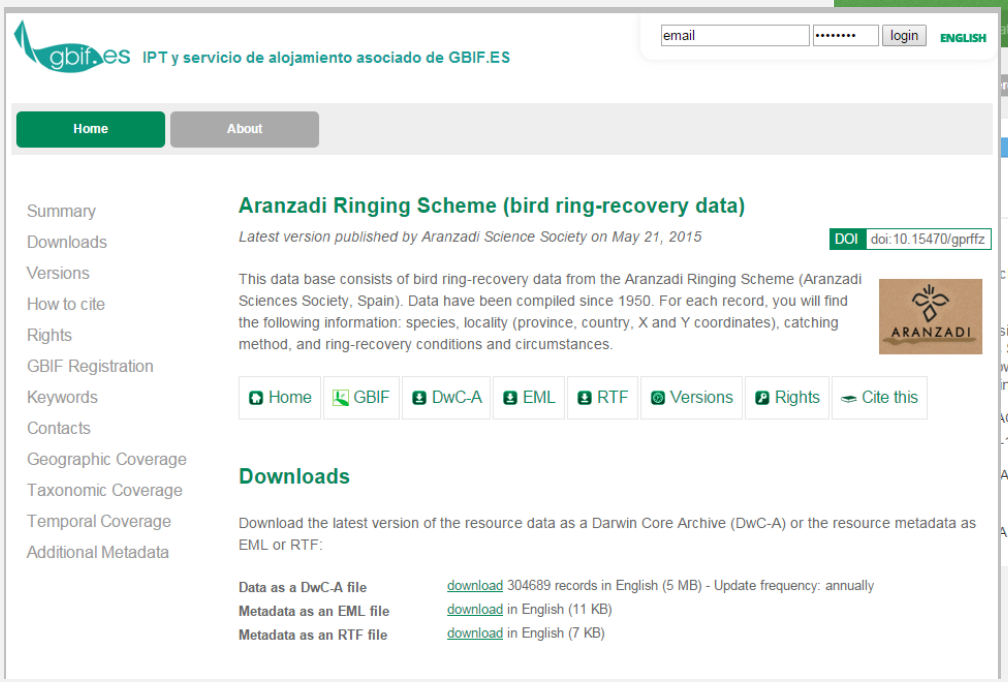
As **Oscar Corcho**, **María Poveda-Villalón** and **Asunción Gómez-Pérez** see it, linked data has put the field of ontology engineering into a new context. Where tradition has favored heavyweight ontologies that demonstrate deep understanding of a domain and enable sophisticated inferences, the

Cambios en la red de GBIF y en contexto mundial en esa dirección

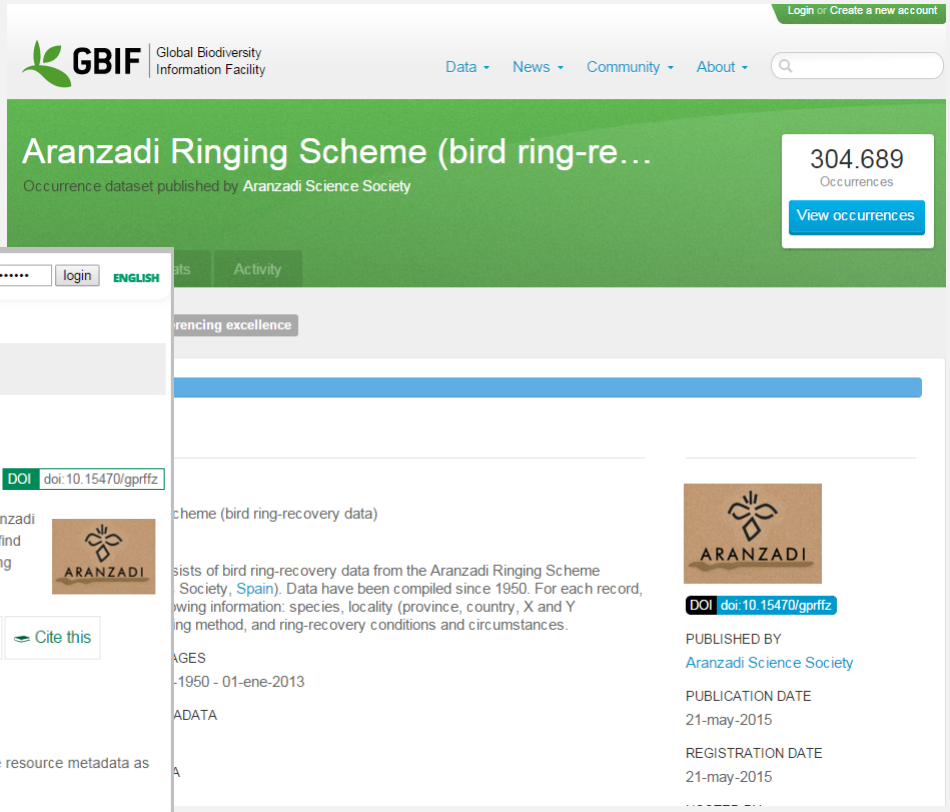
- Identificadores (UUPR)
  - Para juegos de datos
  - Para descargas
  - Para registros
- Licencias estandarizadas
- APIs

Identificadores para juegos de datos

DOI doi:10.15470/gprffz



http://www.gbif.es/ipt/resource?r=anillamiento-aranzadi

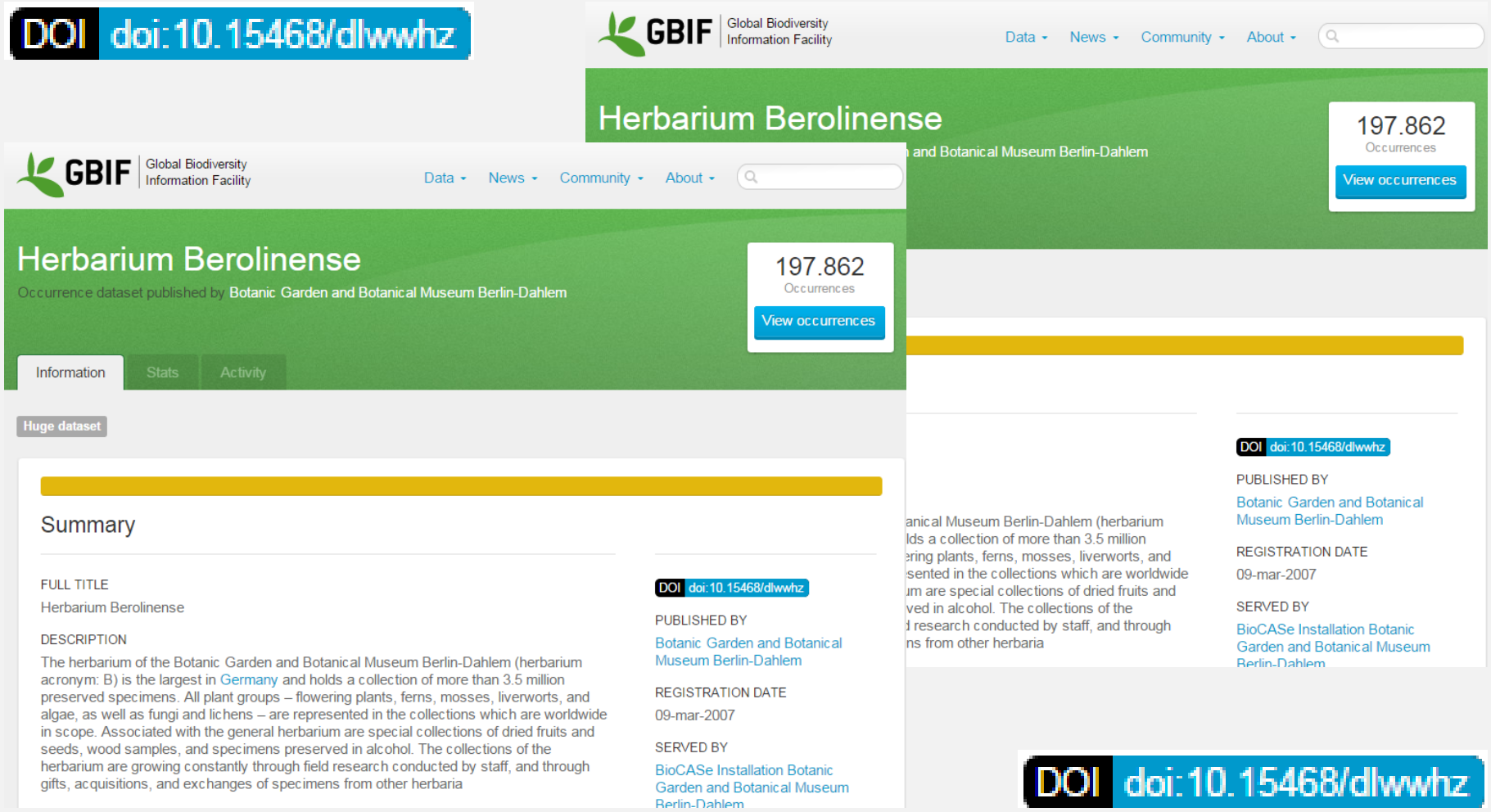


DOI doi:10.15470/gprffz

http://www.gbif.org/dataset/52f2051b-c47e-403a-8e32-04b2f2273c20

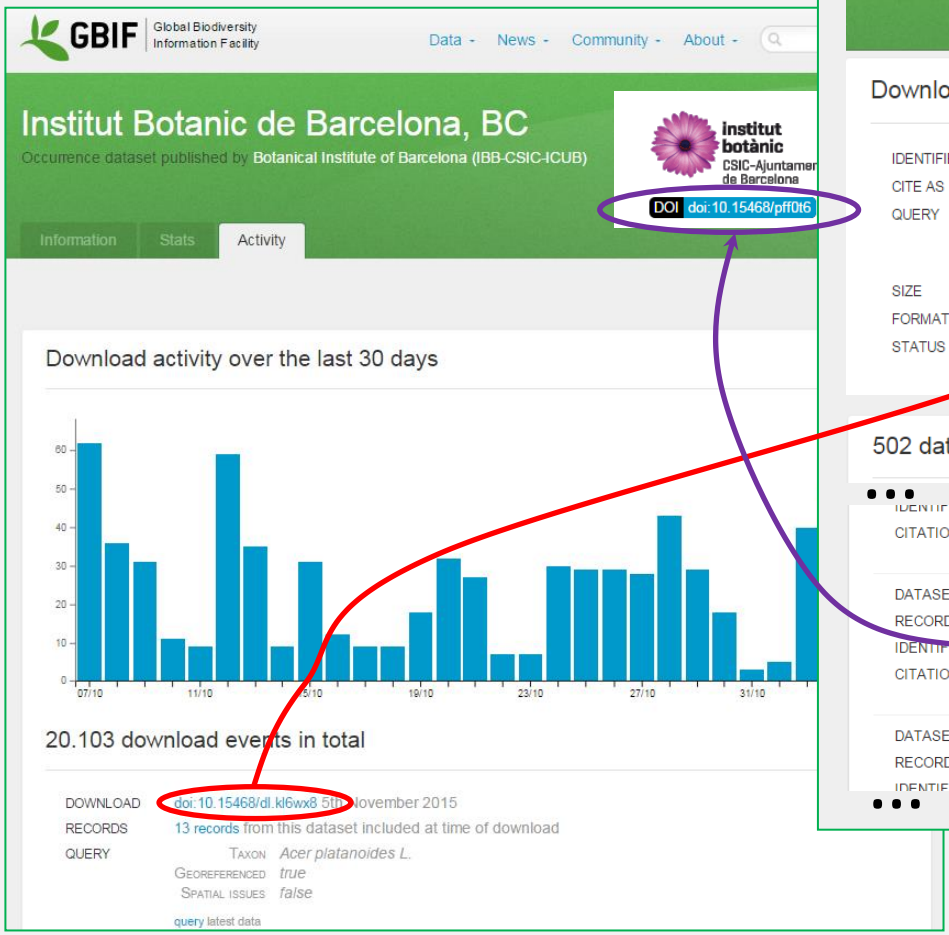
Juego de datos -- base de datos-- recurso

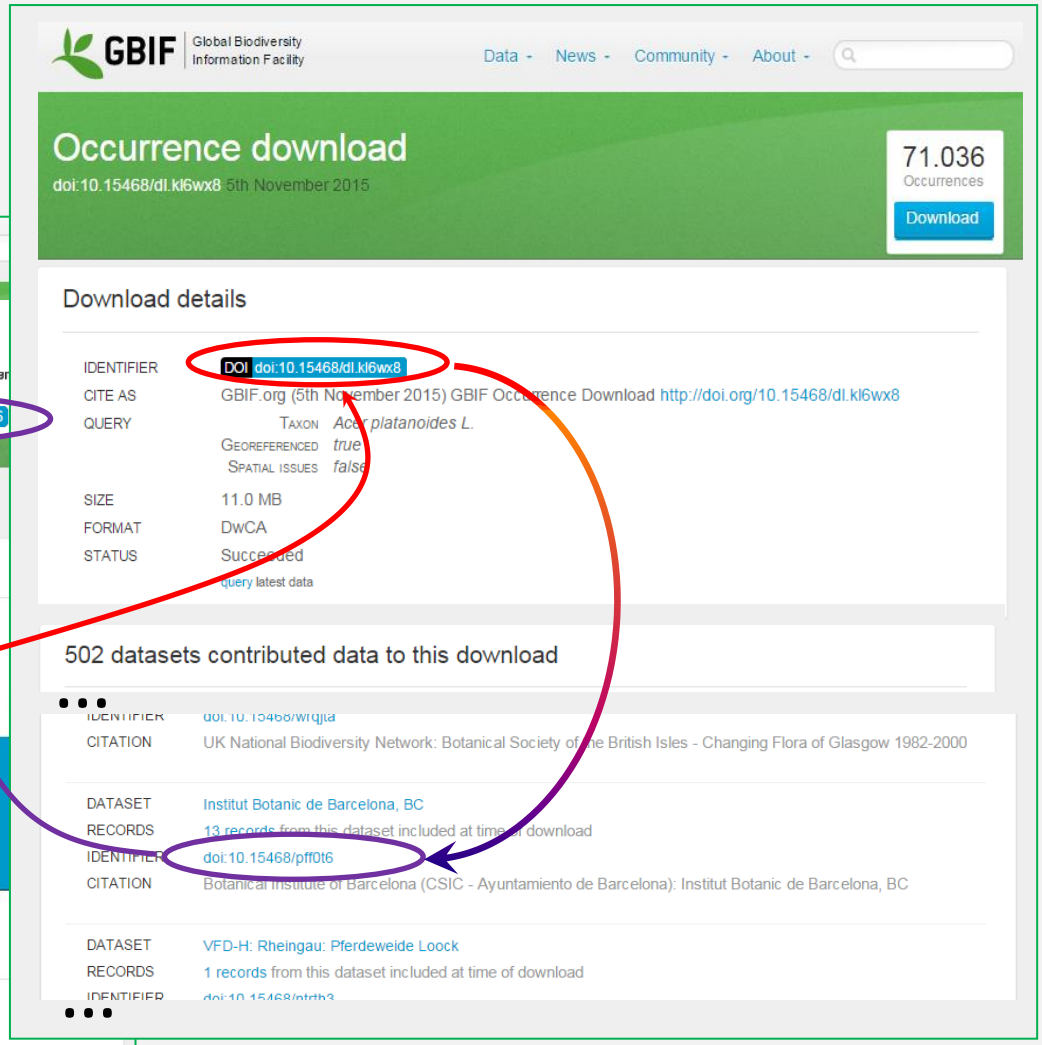
Identificadores para juegos de datos



http://www.gbif.org/dataset/85714c48-f762-11e1-a439-00145eb45e9a

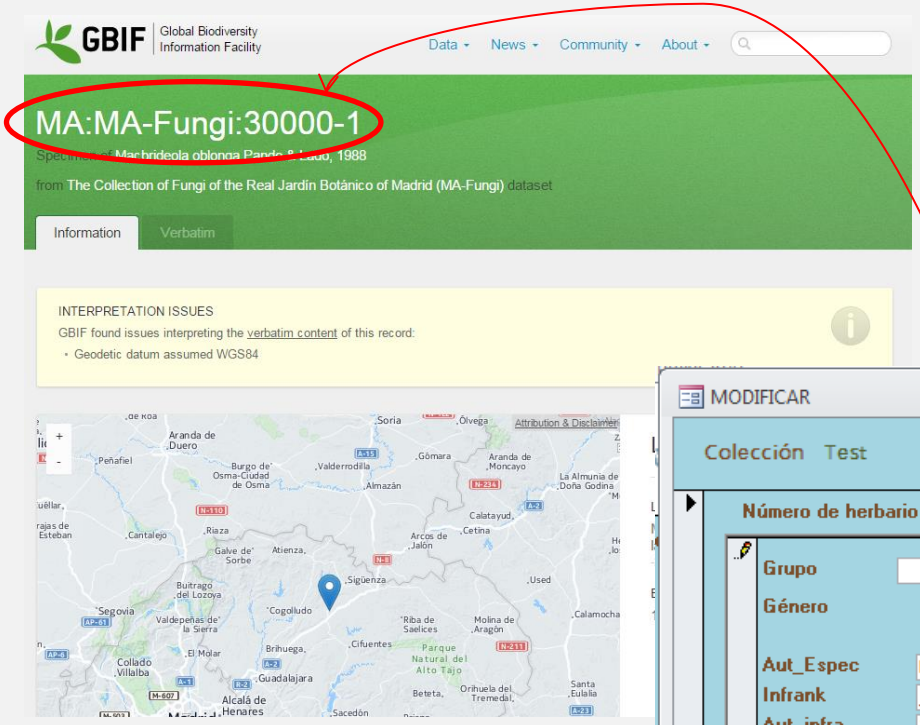
Identificadores para descargas





http://www.gbif.org/dataset/838475f4-f762-11e1-a439-00145eb45e9a/activity

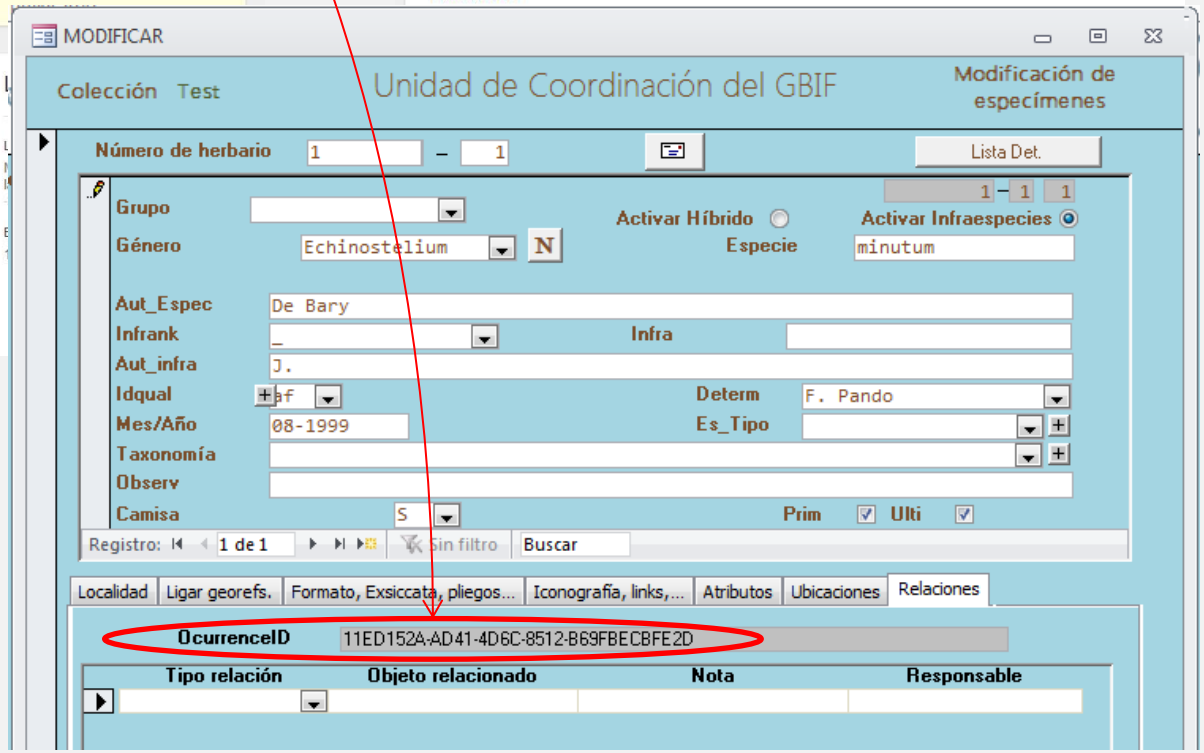
Identificadores para registros



http://www.gbif.org/occurrence/78504121

http://www.gbif.org/occurrence/85254856

Location



- Quitar el humano
- Hacerlos persistentes
- Hacerlos semánticamente agnósticos

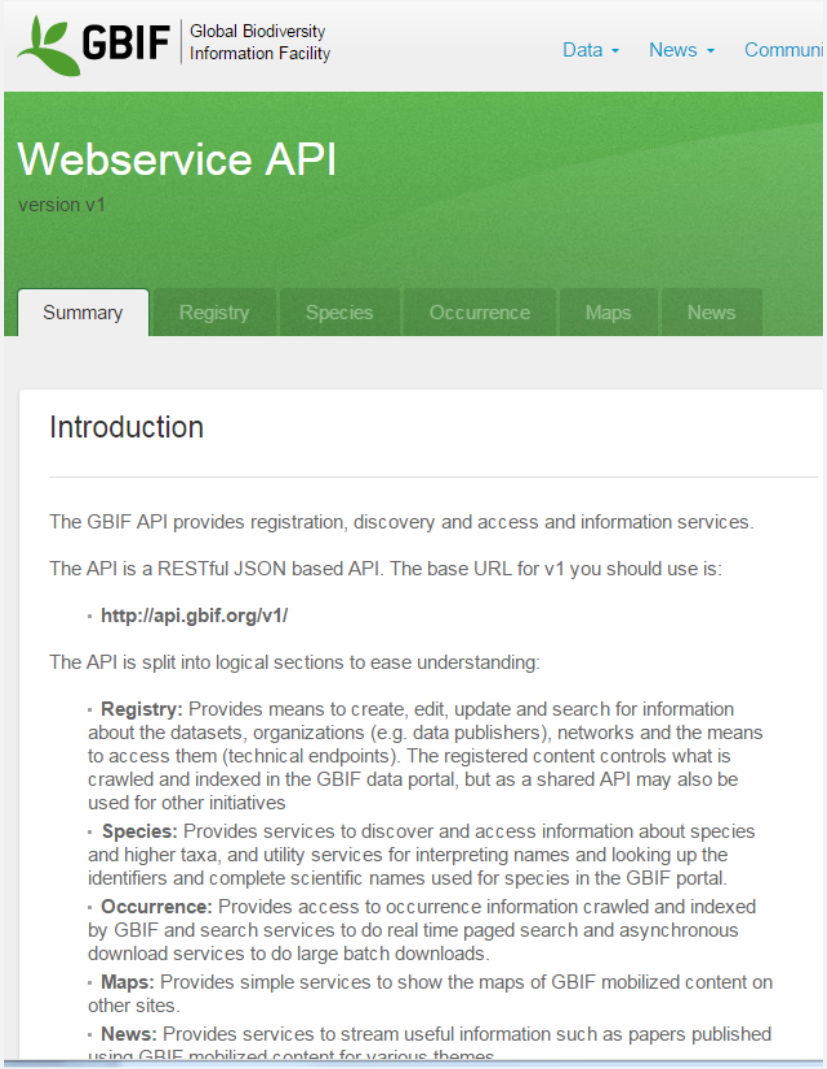
Licencias estandarizadas

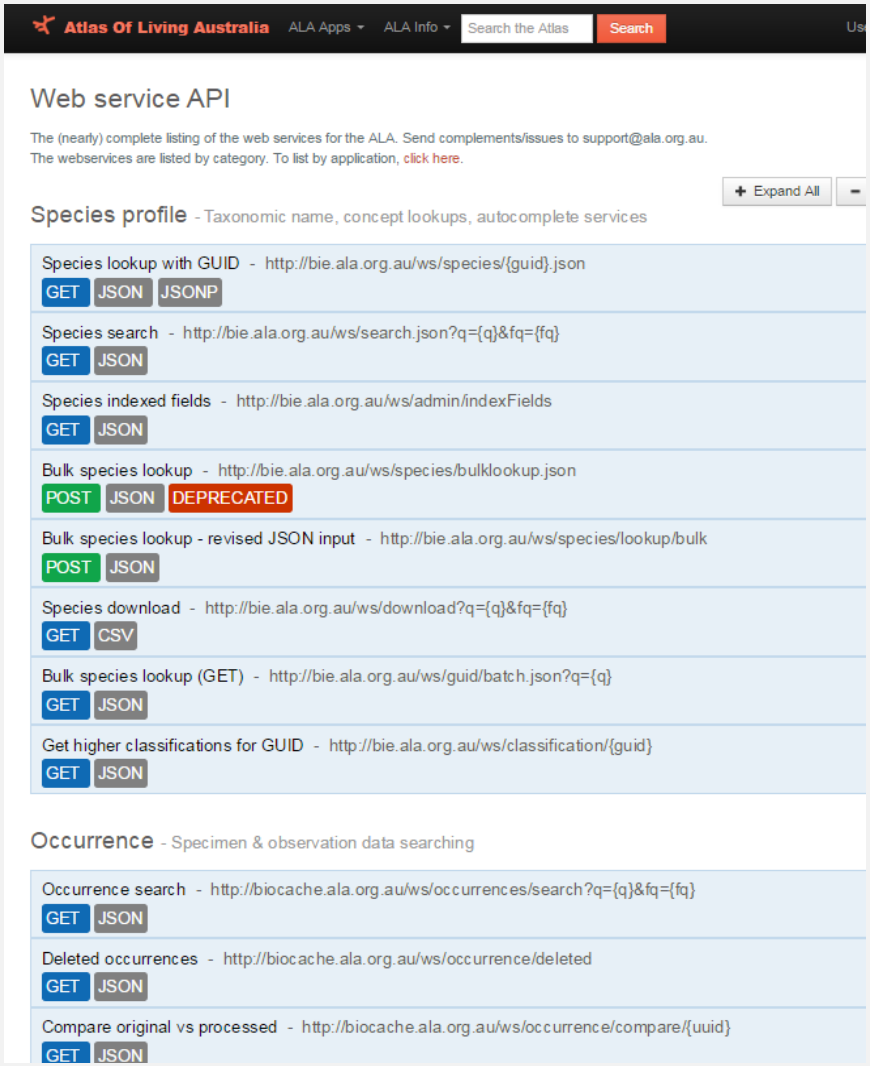
CC0, licencia de uso universal, sin ningún tipo de restricciones.

CC-BY, se permite copiar, redistribuir o editar el material, pero siempre citando adecuadamente la fuente e indicando los cambios realizados.

CC-BY-NC, igual que en el caso anterior pero únicamente para uso no comercial.

APIs





http://www.gbif.org/developer/summary

http://api.ala.org.au/

¿Y después?

- RDF

http://community.gbif.org/pg/search/?tag=rdf

Lo nuevo en 4UCOLL

1. Soporte para web semántica: identificadores persistentes: PURLS, UUIDs
2. Soporte para cumplir con el Protocolo de Nagoya
3. Nuevo interfaz, navegación por “Ribbon”
4. Soporte para códigos QR
5. Exportación de datos en Darwin Core Archive
6. Múltiples mejoras en toda la aplicación

Soporte para web semántica: identificadores persistentes

ESPEC

OcurrenceID	NHERBA
11ED152A-AD41-4D6C-8512-B69FBECBFE2D	
A4BB5EF3-5D93-486F-8897-095A44F1BA2C	
862A11F0-E901-4359-BB3B-EA5D6385BE2B	
73668F86-12	
A9E395F6-0D	
57DCB355-9C	
98FE4213-22	
8BD48C56-C	
D36D3AFA-E	
334BF31B-F8	
E63D9D0F-A	
D9DEE124-8D35-49A8-8597-008E5299D5A9	
7E30A1AA-00D1-4877-8184-049B	
6E2CB586-9424-47FC-0927-4B16	
D8540157-29C1-40B0-0874-34597	
2A89A656-0EE7-4D0F-B2C5-09BA	
AC3CD029-F12C-4162-9AB5-0D8	
72A5A46F-092D-4B90-BBFF-A73246CDA33E	1000 1
66D	

M_METADATA

UUID	Id
140-876n-436p-884h-954k	
195p-708h-617i-192n-356h	
366k-370p-981-350i-246m	
572n-148l-521j-660m-337j	
734l-621j-371n-112n-833n	
846o-630p-919i-725p-319l	

QR_Code

uuid	purl
1FACD4FE-C00F-4DB8-AEDF-CEB11B5F21FB	http://purl.org/MAHerbarium/MA
28695C4C-59F1-4521-B59A-50F579D8AA9C	http://purl.org/MAHerbarium/MA
429FA40F-C947-413D-8E39-B9623BBF9DEE	http://purl.org/MAHerbarium/MA
1FACD4FE-C00F-4DB8-AEDF-CEB11B5F21FB	http://purl.org/MAHerbarium/MA
28695C4C-59F1-4521-B59A-50F579D8AA9C	http://purl.org/MAHerbarium/MA
429FA40F-C947-413D-8E39-B9623BBF9DEE	http://purl.org/MAHerbarium/MA
1FACD4FE-C00F-4DB8-AEDF-CEB11B5F21FB	http://purl.org/MAHerbarium/MA
28695C4C-59F1-4521-B59A-50F579D8AA9C	http://purl.org/MAHerbarium/MA
429FA40F-C947-413D-8E39-B9623BBF9DEE	http://purl.org/MAHerbarium/MA
1FACD4FE-C00F-4DB8-AEDF-CEB11B5F21FB	http://purl.org/MAHerbarium/MA
429FA40F-C947-413D-8E39-B9623BBF9DEE	http://purl.org/MAHerbarium/MA

LOTES_ADJUNTOS

UUId	
74860554-E8C1-4A80-84C0-4AE45484FB8B	C:\Users\mcarmen.GBIF\Documents\katia1.pdf

Lotes

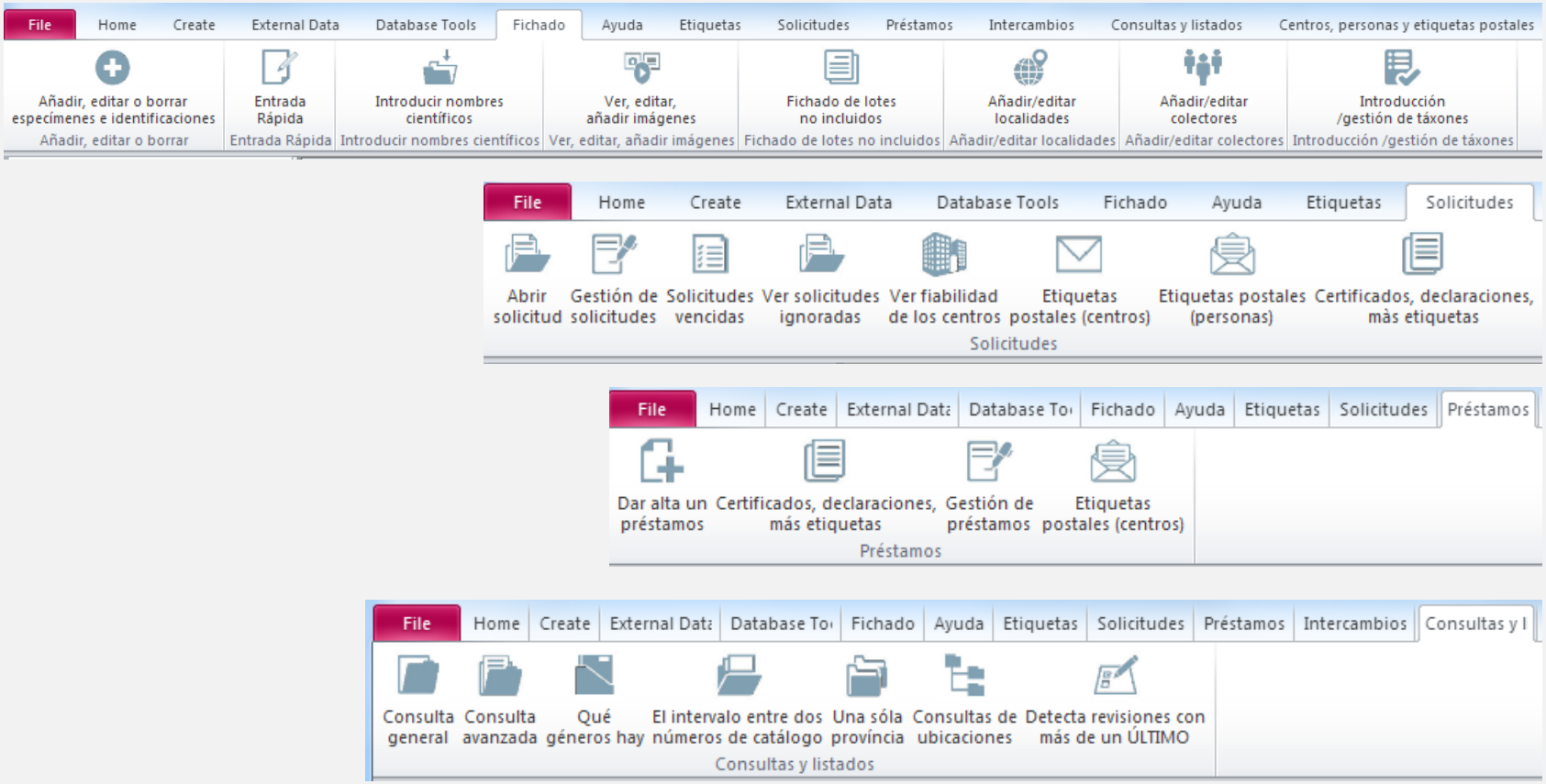
UUid	ESTANTERIA	COLECTI	RESPONSABL	CON_E	CON_NCOL
76B8A80E-F9CC-4D31-9F20-634E1DD287A0	P3/A5/E2/C1	Test	152	No	No
B6294C7D-6D82-4B49-8C39-66AFCCFCE17F	P3/A5/E1/C2/b	Test	151	No	No
74860554-E8C1-4A80-84C0-4AE45484FB8B			151	No	No
598BC16B-522B-41EE-A535-B6246438E942			150	No	No

Soporte para cumplir con el Protocolo de Nagoya

- 1.- Los ejemplares catalogados reciben un identificador único, global, persistente y resolvible, que permite su referencia y su trazabilidad hasta el origen.
- 2.- La aplicación permite asociar a cada ejemplar o lotes de ejemplares documentación; en este contexto, PICs, MTAs MATs.: pdf, hiperenlaces, imágenes…
- 3.- Las hojas de envío de material por las que se transfiere e intercambian ejemplares entre instituciones llevan un texto de referencia sobe como manejar y referenciar el material de acuerdo con Nagoya y las recomendaciones de CETAF:

  https://www.cbd.int/abs/submissions/icnp-3/EU-Taxonomic-practices.pdf

Nuevo interfaz, navegación por “Ribbon”



Soporte para códigos QR

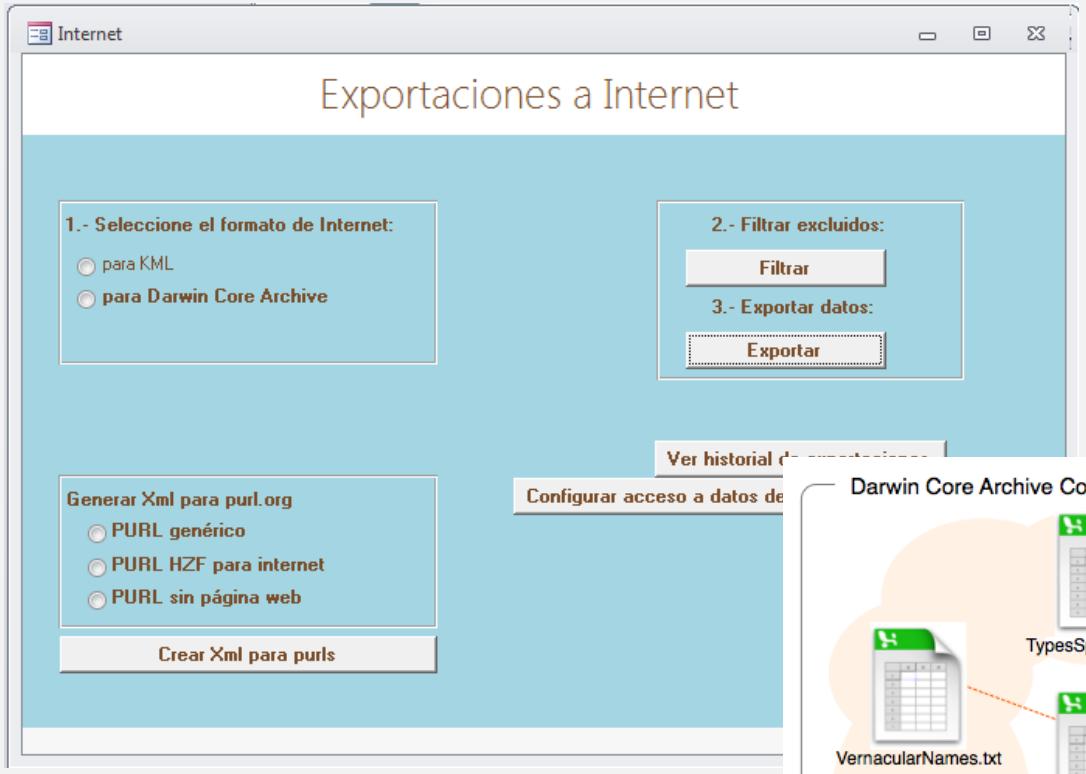
Hortus Regius Matritensis (MA-Fungi)

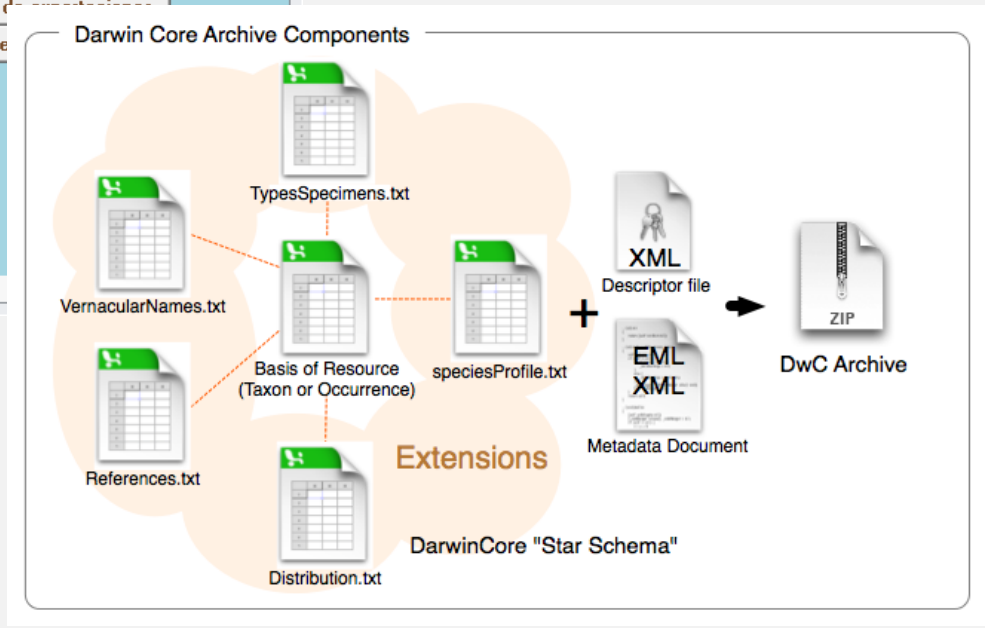
Battarrea stevenii (Libosch.) Fr.

ESP. Córdoba: carretera Comarcal 437 km 2,8; 30SUG4092; 05/10/1994; *S. Capello*, Det.: S. Capello.

MA-Fungi 35883

Exportación de datos en Darwin Core Archive





Francisco Pando

Unidad de coordinación, GBIF España
Real Jardín Botánico - CSIC
Claudio Moyano 1, 28014 Madrid, Spain
pando@gbif.es
www.gbif.es

GBIF-ES es el Nodo Nacional de Información sobre Biodiversidad patrocinado por el Ministerio Español de Economía y Competitividad, gestionado por el Consejo Superior de Investigaciones Científicas