Introducción al Portal personalizable de GBIF con extensiones

María Mora
Administradora del Nodo GBIF
Costa Rica

Agenda

- Introducción
- El Portal personalizable de GBIF
  - Objetivo
  - Arquitectura
  - Tecnologías utilizadas
- Opciones de software para proveer datos a la red
- Extensiones al Portal GBIF (GBIF.ES, IABIN)

Introducción

- Misión de GBIF ->Hacer que los datos primarios de la biodiversidad del mundo estén disponibles de forma libre y universal por medio de desarrollar una red de bases de datos con acceso por medio de Internet.
- La red debe ser flexible para solventar las necesidades de los proveedores de datos y de los usuarios.

Introducción

- 4 puntos a tomar en cuenta:
  - Premisa → Se desea compartir información similar entre proveedores.
  - Objetivo → Intercambio e integración de información entre diferentes instituciones.
  - Problema → Fuentes de datos tienen estructuración de datos muy diversas
  - Solución → Exponer datos de forma estándar.

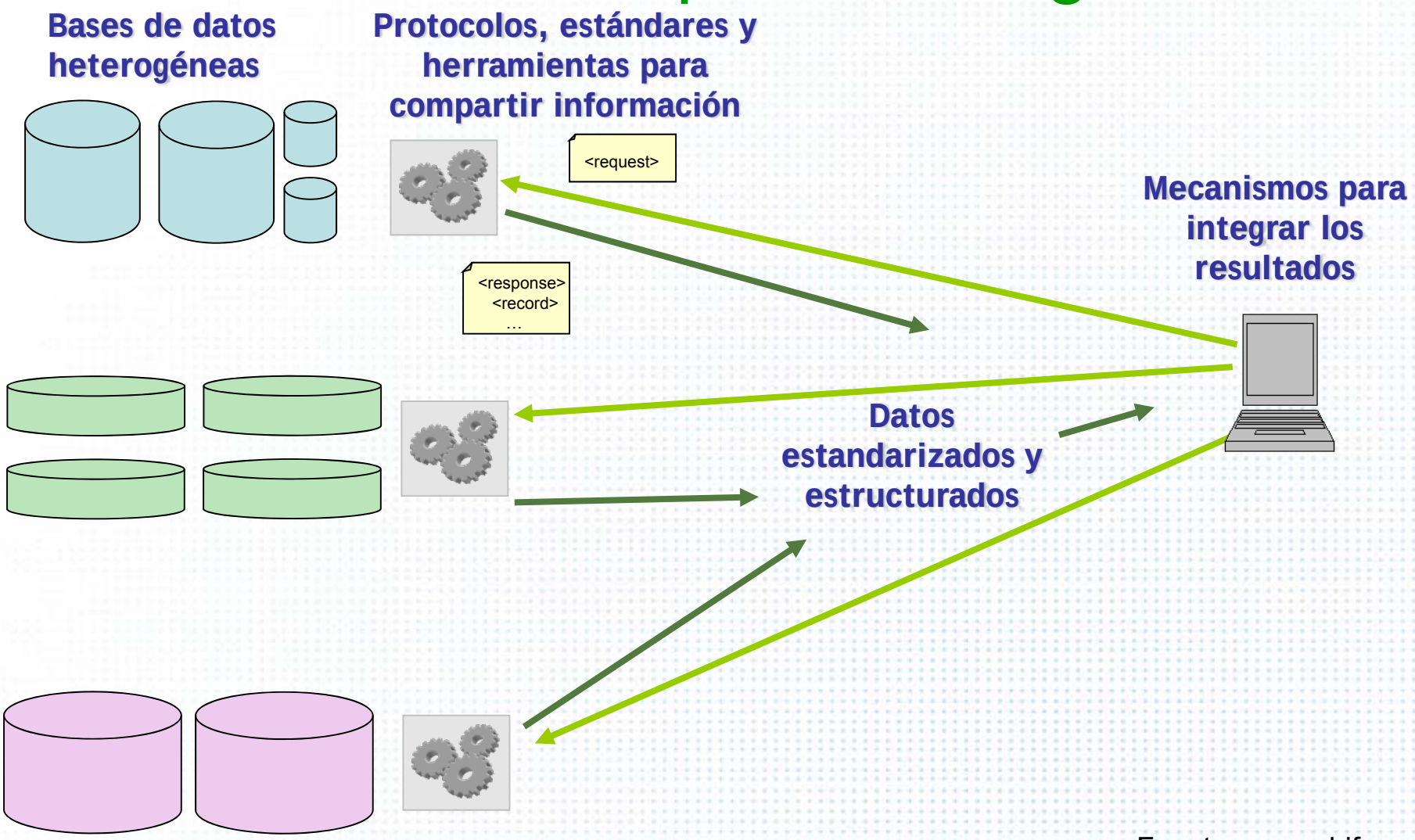
Introducción

- Portal personalizable de GBIF
  - Un desarrollo con el soporte de una institución con amplia experiencia en el manejo de información biológica.
  - Gratis y open source.
  - En constante mejora por parte de los desarrolladores.

Introducción

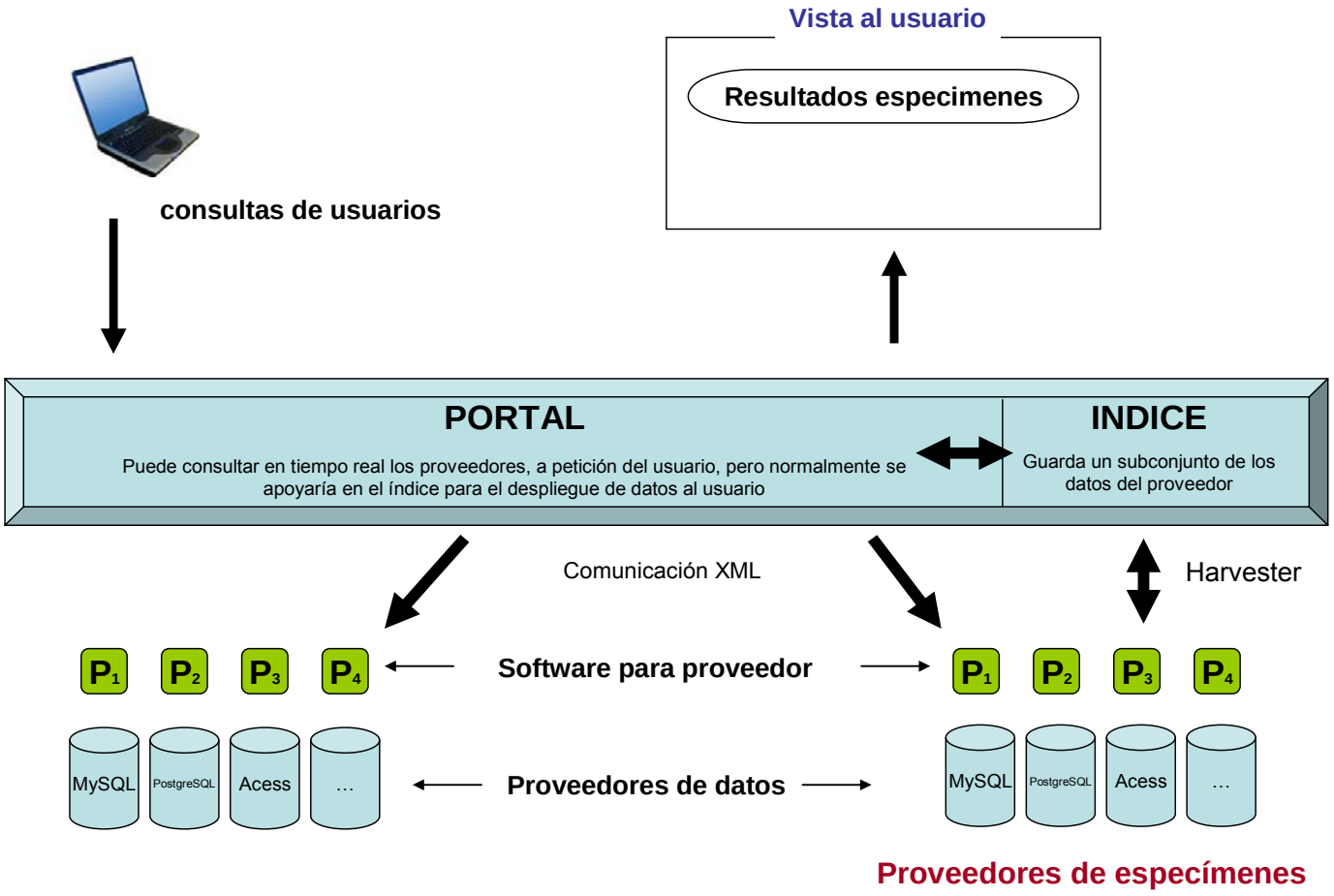
- Portal personalizable de GBIF
  - Dos tipos de datos
    - Registros de ocurrencias de especies
    - Nombres y clasificaciones de organismos
- Búsqueda por país, recurso
- Despliegue de mapas
- Soporte para DiGIR/DwC
- Soporte para TAPIR e IPT

Architectura para la integración



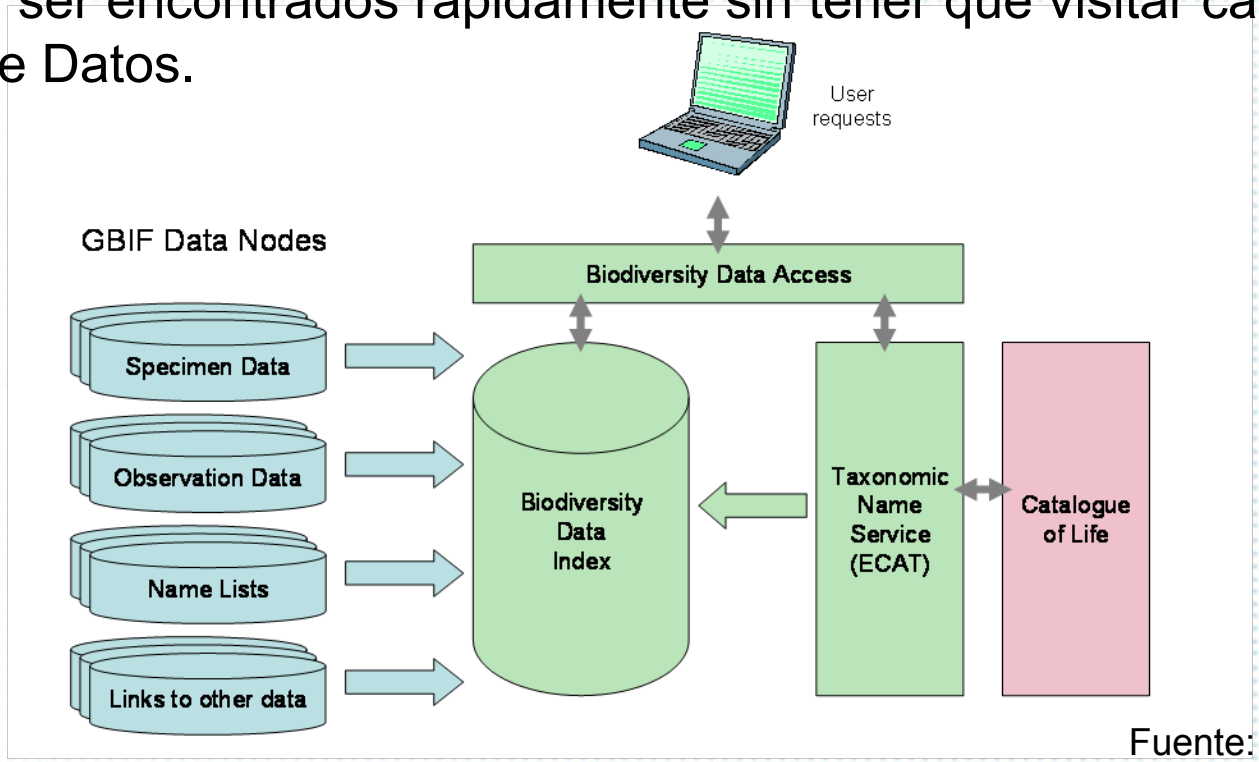
Fuente: www.gbif.org

Arquitectura General



Indice central del Portal GBIF

Una consideración que ha implementado GBIF en la arquitectura actual ha sido la generación de un índice como componente central de la red, lo que asegura que los registros de datos relevantes pueden ser encontrados rápidamente sin tener que visitar cada Nodo de Datos.

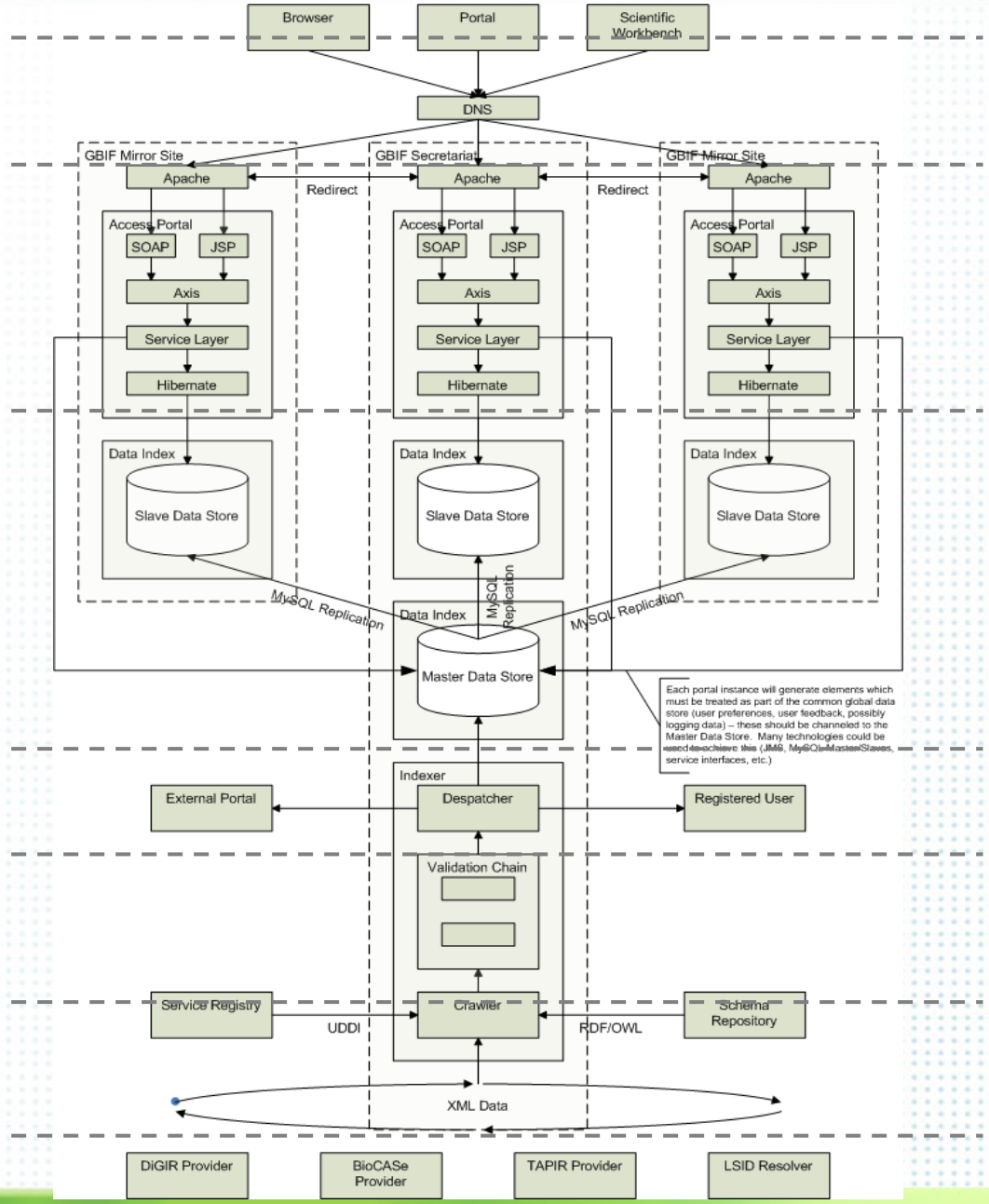


Fuente: www.gbif.org

Indice del Portal

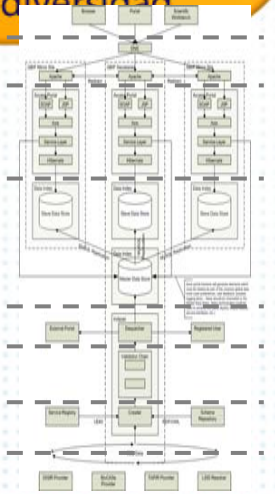
- Metadatos de especimenes/ocurrencias
- Metadatos de especies
- Información taxonómica asociada a un taxon
- Nombres comunes y grupos nomenclaturales
- Metadatos de los proveedores de datos
- Bitácora de la actividad de los usuarios en el portal

Portal architecture (new version)



Fuente: www.gbif.org

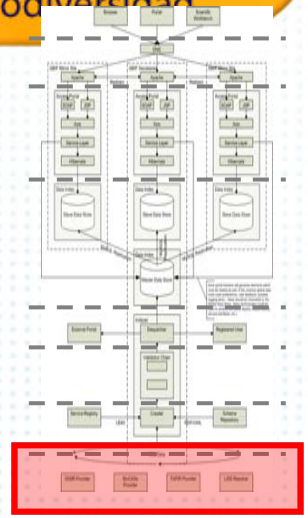
Portal architecture (new version)



- **The following slides indicate some of the function planned for the new data portal, following the sequence of layers in the architecture diagram**
- **Items planned for release by the end of 2006 are shown in black**
- **Possible extensions (in 2007 and beyond) are shown in grey**
- **Tentative dates for early (beta) access to versions of functions are included where appropriate**
  - **NOTE – dates are generally for FIRST access to versions of each function**
  - **NOTE – in many cases the dates relate to TEST versions**
  - **NOTE – many of these features already exist in some form in the prototype data portal – the dates refer to access to the new implementations**
- **User surveys are in place to refine requirements and to prioritise functions**
- **For further information, or to contribute to discussion on these features, please visit the Data Portal Design wiki:**
  - **http://wiki.gbif.org/dadiwiki/**

Fuente: www.gbif.org

Data resources

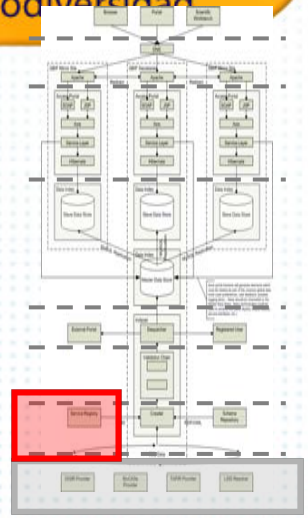


The new implementation should offer much greater flexibility for indexing new resource types

- **Taxon Occurrence (c. Aug 2006)**
  - **DiGIR with Darwin Core (v. 1.2, v. 2.0, MaNIS, OBIS)**
  - **BioCASe with ABCD (v. 1.20, v. 1.48, v. 2.06)**
  - **TAPIR with Darwin Core v. 2.0**
  - **TAPIR with ABCD v. 2.06 (including IPGRI)**
- **Nomenclature and Taxonomy (c. Aug 2006)**
  - **Taxon API (TAPIR) with Taxon Concept Schema (TCS)**
- **Possible Future Resources (for evaluation and prioritisation)**
  - **Taxon API (TAPIR) with TCS for regional checklists**
  - **Taxon API (TAPIR) with TCS for statutory lists (red lists, CITES, etc.)**
  - **SPICE with Species 2000 Common Data Model for taxonomic data**
  - **TDWG Structured Descriptive Data (SDD) documents**
  - **SDD query via TAPIR or LSID+RDF**
  - **Collection metadata (TDWG Natural Collections Description data)**
  - **Ecological data sets with EML metadata**
  - **Links to digital biodiversity literature**
  - **Sets of web links for species pages, images, sequences, etc. classified by taxon**
  - **Tab-delimited taxon-based data (possibly mapped to GBIF ontology?)**
  - **RDF data using LSIDs and (TDWG? or GBIF?) ontology**

Fuente: www.gbif.org

Service registry



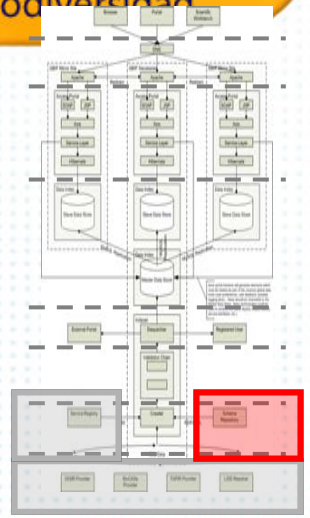
Revised UDDI registry with improved metadata handling

- **Updated UDDI registry for all data resources (Jun 2006)**
  - **Categorisation by higher taxa, provider country, countries in data, basis of record (specimen vs. observation, etc.)**
  - **Linkage to RDF metadata for each resource from index database**
  - **Possible use of proxies within Data Portal to map between data models**
- **Standardised provider data use agreements (Dec 2006)**
  - **Increased provider control over how data are managed by the portal**
  - **Possible use of standard Creative Commons licences**
- **Provider `console´ interface (c. Dec. 2006)**
  - **Password-controlled screens for each provider**
  - **Configure periods when resources can be indexed**
  - **Request immediate re-index**
  - **View status of indexing for each resource**
  - **View reports of possible issues detected during indexing**
  - **View usage reports**

Fuente: www.gbif.org

See: http://wiki.gbif.org/gbifwiki/wikka.php?wakka=ServiceRegistry

Schema repository



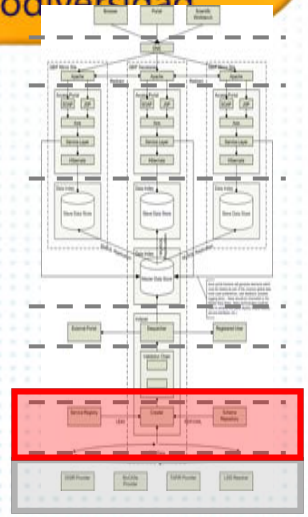
This component should be developed in conjunction with ongoing architecture work in TDWG

During 2006, a `stub´ implementation will be created offering only as much function as is required

- **Ontology of biodiversity data classes**
  - Classes such as TaxonConcept, Specimen, CharacterState, etc.
  - Include properties of interest to data portal (during indexing or data display) for each class
- **Tools for handling biodiversity data**
  - Annotation of class properties with representations in different data models, national language labels, etc.
  - Association of properties with methods to transform between alternate representations (e.g. {day, month, year} ⇔ {date})
  - Association of properties with methods to validate content
  - Generation of handlers for different data models (for use in indexer and elsewhere)
  - Web service interfaces to expose concepts and logic used by data portal

Fuente: www.gbif.org

Resource crawler



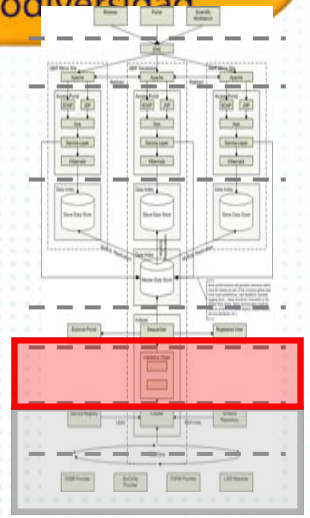
Discover data updates and retrieve data for indexing

- **Scheduled execution (Aug 2006)**
  - **Automated selection of resources to index**
  - **Multi-level indexing (refresh metadata, index updates, index all)**
  - **Provider control over schedule for indexing their resources**
- **Strategy-driven approach (Aug 2006)**
  - **Use `probe´ requests and past experience in indexing the resource to identify capabilities of provider software**
  - **Identify strategy to find records to index (all, new, updated, previously inaccessible or invalid, etc.)**
  - **Develop `map´ of resource to track indexing progress, inaccessible or invalid records, etc. and to allow indexing to be paused**
  - **Retrieve records from resource according to map and submit them to the Validation Chain**
  - **Upon completion, store map and outcomes for reference in future indexing**

Fuente: www.gbif.org

See: http://wiki.gbif.org/gbifwiki/index.php/Data+Indexing

Interpretation and validation



Validation Chain handles records from Crawler

- **Configurable workflow model (Aug 2006)**
  - Handling for different classes of data and different data formats
  - Easy to add additional steps and branches as needed
  - Simplifies development of logic e.g. to handle viral taxonomy
- **`Raw´ data parsed using Schema Repository**
  - Mapping of source data formats into forms compatible with Data Index
  - Extensible to handle additional requirements (e.g. hybrid formulae, linkout to gazetteer tools)
  - Aim is to normalise data from different sources as far as possible
- **Validation of content**
  - Annotate records with metadata to assist providers and users
  - Missing or unclear values for key fields (e.g. `X´ for basis of record, `? ?´ for scientific name
  - Incompatible values (e.g. latitude and longitude outside named country)
  - Taxa outside identified scope for resource (e.g. fungi included in resource identified as including data only for Plantae)
  - Understand and be able to report any potential issues with a record or resource and to report them to providers
  - Also determine which resources are particularly strong in different areas (e.g. fully georeferenced)

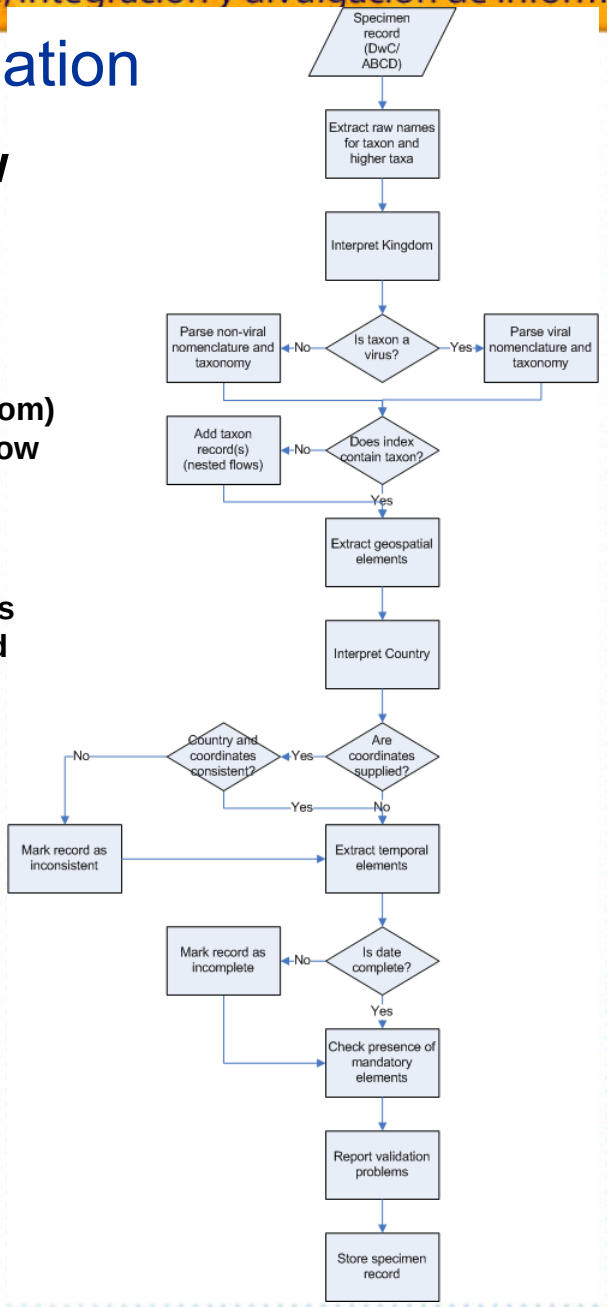
Fuente: www.gbif.org

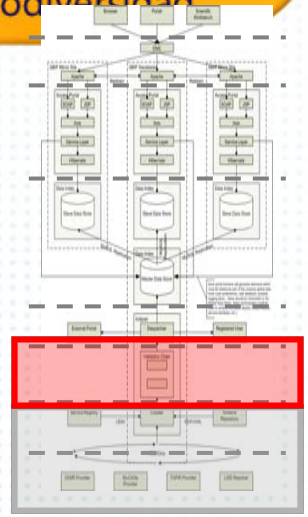
Interpretation and validation

Simple example workflow

Process specimen record

- → **Handle taxonomy**
  - → **Extract raw taxon names**
  - → **Parse name (according to kingdom)**
  - → **Optionally trigger nested workflow to add new taxon to index**
- → **Handle geography**
  - → **Extract geospatial elements**
  - → **Validate country and coordinates**
  - → **Annotate record if conflict found**
- → **Handle dates**
  - → **Extract temporal elements**
  - → **Validate temporal elements**
  - → **Annotate record if invalid or incomplete**
- → **Report outcomes**
- → **Store record**

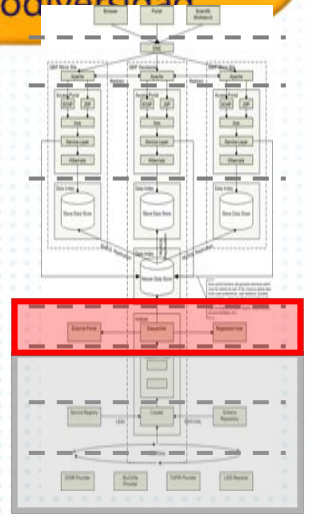




Fuente: www.gbif.org

See: http://wiki.gbif.org/datawiki/wiki.php?wakka=DataIndex

Data despatcher



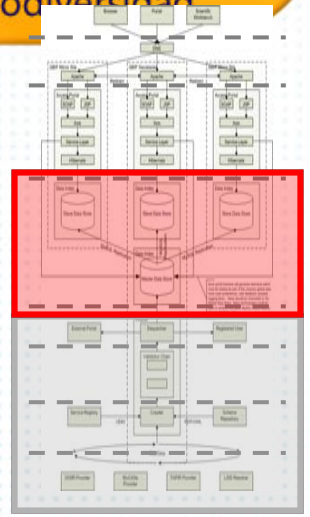
Decision point for parsed and validated data objects

- **Configurable pattern-based handling of objects**
  - **Decisions based on properties assigned to objects by Validation Chain**
  - Possible point to trigger notifications to users and external portals as new records are added e.g. for a taxon or region of interest
- **Store object in data index database (Aug 2006)**
  - **Handlers for all data classes included in index**
  - Readily extensible to support alternative database instances
- **Trigger feedback from Validation Chain to data providers**
  - Construct reports for data providers regarding indexer execution
  - Summary statistics (e.g. % records georeferenced, correspondence between taxonomy and any taxonomic authority identified in metadata)
  - Reports either mailed to data provider or made available through console interface

Fuente: www.gbif.org

See: http://wiki.gbif.org/datawiki/wiki.php/wakka=DataIndex

Synchronised data stores



Master database with readonly mirrors

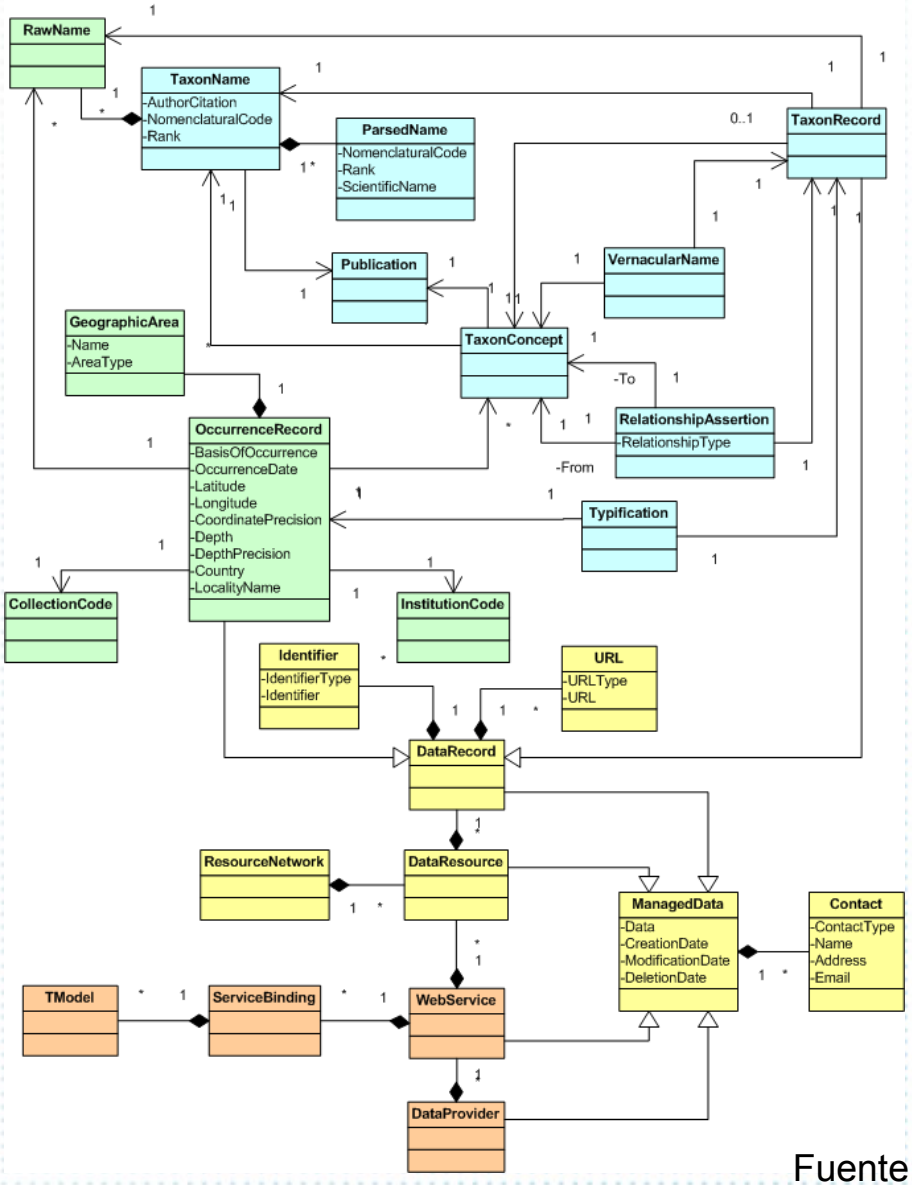
- **Master database managed by Despatcher**
  - Index records added, modified, or marked deleted according to latest versions found by Crawler
  - When records are removed by a provider, an empty record is retained in the database, including only the record identifier, metadata identifying the provider, and the dates during which the record was available
  - All index records in database assigned internal LSIDs to support rapid retrieval
- **Live portal instances using readonly mirror databases**
  - Separation of load between indexing and user requests
- **Comparative testing of MySQL and PostgreSQL**
  - During 2006, MySQL and PostgreSQL database instances will be used and tested in parallel
  - Database modeled using Hibernate as the persistence layer (easy deployment of alternative RDBMS solutions)
- **Centralised logging of information on data usage**
  - Centralised gathering and reporting of data usage through all mirror portals

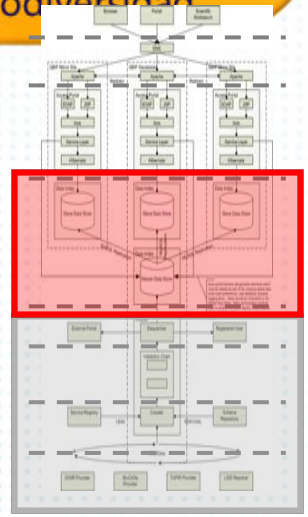
Fuente: www.gbif.org

Synchronised data stores

Logical model

- Orange: UDDI metadata
- Blue: data primarily from nomenclatural/taxonomic resources
- Green: data primarily from specimen/observation resources
- Yellow: metadata held for all classes of data

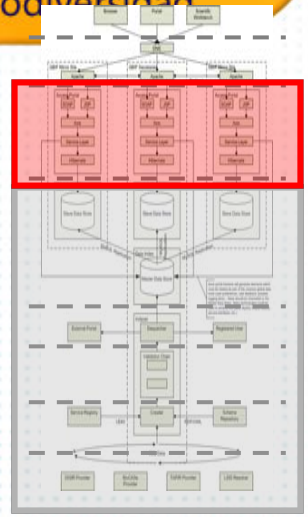




Fuente: www.gbif.org

See: http://wiki.gbif.org/datawiki/wiki.php?wakka=DataIndex

Web applications



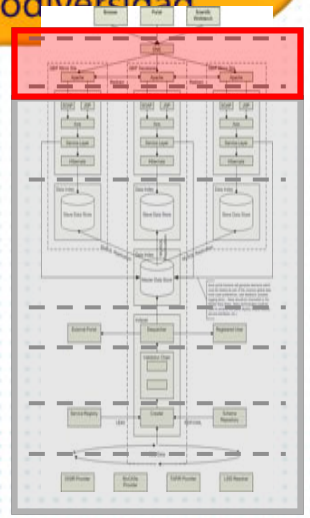
Application logic to access index data stores

- **Java service layer for data access**
  - Interface supporting all requests used by portal interface and web services
  - Interface implemented in a local version for use within an instance of the portal and in a client/server version which could be used by remote portals (e.g. regional or thematic portals) or analytic tools
- **Model-View-Controller (MVC) framework**
  - Web applications implemented using an MVC framework to simplify deployment of new interfaces and reuse of underlying components
- **Configurable HTML user interface (Jul 2006 onwards)**
  - User interface designed to simplify redeployment of components in other portals (using client/server version of the Java service layer)
  - User customisation for language, contact details and layout preferences
  - Registration for notification of new data by taxon and/or geographic region
  - RSS feeds
- **Web services**
  - Application interfaces as used to build HTML user interface
  - Search and access services using TDWG TAPIR protocol and standards
  - WFS access to taxon occurrence data
  - LSID resolution

Fuente: www.gbif.org

See: http://wiki.gbif.org/dadiwiki/wikka.php?wakka=HTMLUserPortal
See: http://wiki.gbif.org/dadiwiki/wikka.php?wakka=XMLDataServices

Mirrored access



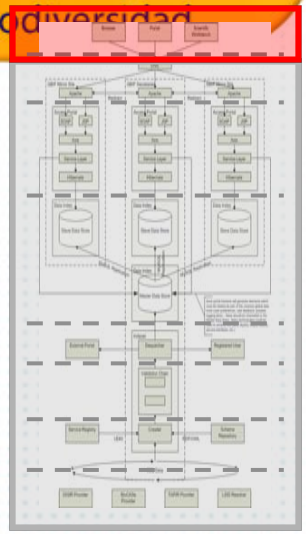
DNS-based use of mirror portals (Dec 2006)

- **http://www.gbif.net/ resolved to regional mirror**
  - **Apache-based redirection in case of local downtime maintenance**
- **Full HTML UI and web services from all mirrors**
  - **All portal instances based on shadows of the Master Data Store**
  - **Personalisation settings operate against all mirrors**
  - **Usage statistics managed centrally**

Fuente: www.gbif.org

See: http://wiki.gbif.org/portalwiki/bin/view/GBIF/GBIFDataPortalDeploymentModel

Clients



Support web browsers and client applications

- **HTML user interface (Jul 2006 onwards)**
  - **Browse and search capabilities**
  - **Download of tab-delimited or XML data sets**
  - **User feedback to data providers**
  - **User personalisation options**
- **Portal interface exposed as web services (Dec 2006)**
  - **WSDL and SOAP access to all data access functions used to construct the HTML UI**
  - **Client implementation of Java Services layer API based on web services**
- **Additional web services to support community standards**
  - **TAPIR with Darwin Core**
  - **TAPIR with ABCD**
  - **Taxon API with Taxon Concept Schema**
  - **Web Feature Service**
- **Encourage development of applications and toolkits**
  - **Expose data for analysis by external tools and workflow applications**

See: http://wiki.gbif.org/dadiwiki/wikka.php?wakka=HTMLUserPortal
See: http://wiki.gbif.org/dadiwiki/wikka.php?wakka=UserServices
See: http://wiki.gbif.org/dadiwiki/wikka.php?wakka=XMLDataServices

Fuente: www.gbif.org

Software para estandarizar datos

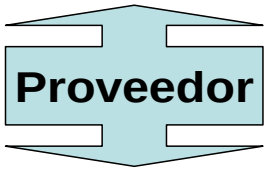
- Función → exponer la información entre diferentes proveedores, de una manera estándar.
- Interacción mediante mensajes XML
- Evita interacción de bajo nivel (SQL)
- Software → esquema conceptual + protocolo de comunicación

Software para estandarizar datos

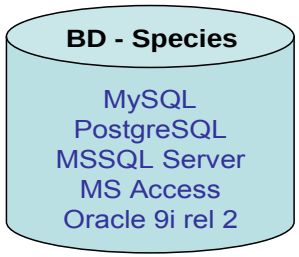
El cliente puede ser un Portal que integre datos de diferentes proveedores.

XML RQ/RP
Esquema XML

Comunicación Proveedor-Cliente se hace mediante XML



Driver - SQL

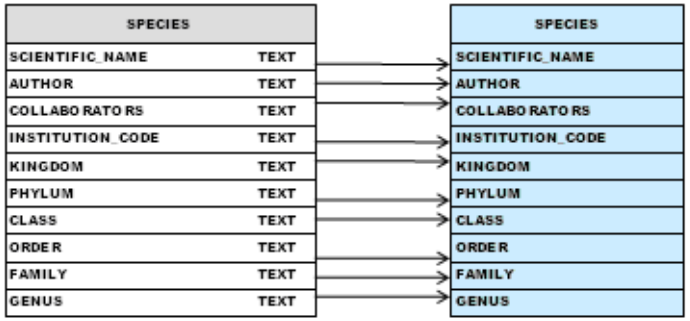


Comunicación SW Proveedor-BD se hace mediante instrucciones SQL

Asociación Proveedor → BD

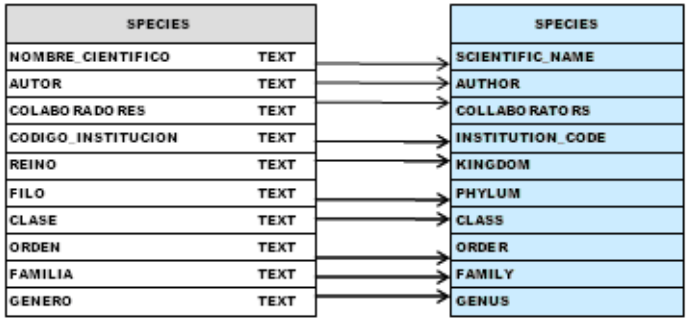
PROVEEDOR A

Mundo ideal: elementos con el mismo nombre que los elementos del esquema conceptual



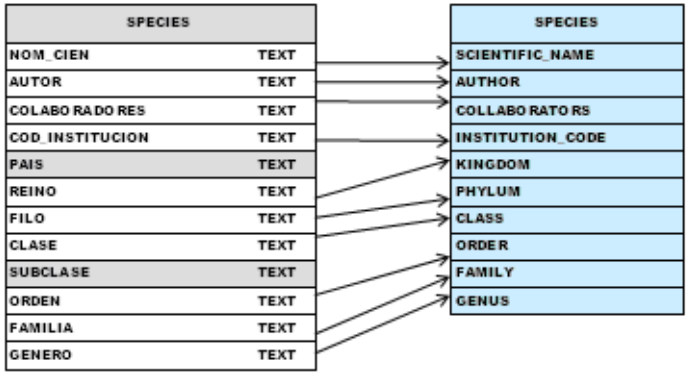
PROVEEDOR B

Elementos en un lenguaje diferente a los elementos del esquema conceptual



PROVEEDOR C

Menos elementos que los presentes en el esquema conceptual y con diferente nombre



Ejemplo Solicitud DiGIR/DwC

```
<request>
    <header>
        <version>1.0</version>
        <sendTime>2003-03-09T19:14:58-05:00</sendTime>
        <source>216.91.87.102</source>
        <destination resource="atta">http://osa.inbio.ac.cr/DiGIR/DiGIR.php</destination>
        <type>search</type>
    </header>
    <search>
        <filter>
            <like>
                <darwin:ScientificName>Inga vera%</darwin:ScientificName>
            </like>
        </filter>
        <records limit="10" start="0">
          <structure schemaLocation="http://digir.sourceforge.net/schema/conceptual/darwin/brief/2003/1.0/darwin2brief.xsd"/>
        </records>
        <count>true</count>
    </search>
</request>
```

Ejemplo Respuesta DiGIR/DwC

```
<response>
    <header>
        <version>$Revision: 1.10 $</version>
        <sendTime>19-02-2008 15:25:46-0600</sendTime>
        <source resource="atta">http://osa.inbio.ac.cr:80/digir/DiGIR.php</source>
        <destination>172.16.16.9</destination>
    </header>

    <content>
        <record>
            <darwin:DateLastModified>2007-03-12</darwin:DateLastModified>
            <darwin:InstitutionCode>INB</darwin:InstitutionCode>
            <darwin:CollectionCode>Plantae</darwin:CollectionCode>
            <darwin:CatalogNumber>1509251</darwin:CatalogNumber>
            <darwin:ScientificName>Inga vera</darwin:ScientificName>
        </record>
    </content>

    <diagnostics>
        <diagnostic code="STATUS_INTERVAL" severity="info">600</diagnostic>
        <diagnostic code="STATUS_DATA" severity="info">4,4,0</diagnostic>
        <diagnostic code="MATCH_COUNT" severity="info">81</diagnostic>
        <diagnostic code="RECORD_COUNT" severity="info">2</diagnostic>
        <diagnostic code="END_OF_RECORDS" severity="info">false</diagnostic>
    </diagnostics>
</response>
```

Posibles escenarios para un Portal

- Escenario sin estandarización de datos
  - Poco extensible
  - Difícil su mantenimiento
  - Mucho costo
- Escenario con estandarización de datos
  - Escenario óptimo
  - Mas extensible, mas fácil de mantener

Escenario sin estandarización de datos

consulta

PORTAL

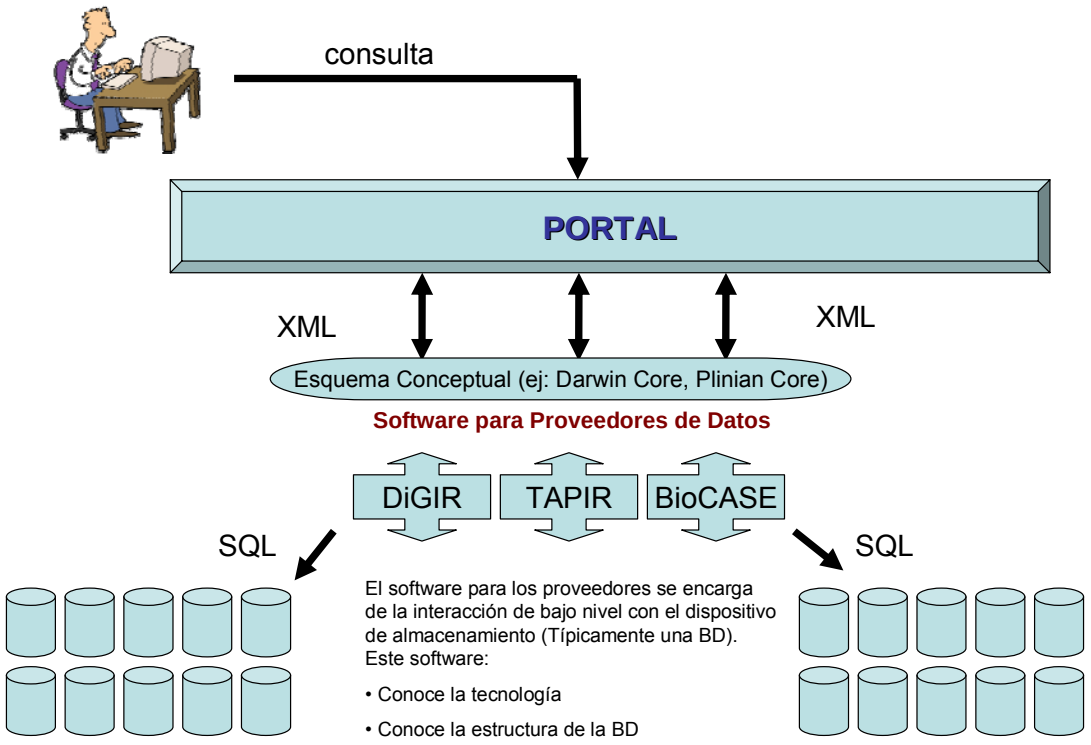
Para que el portal exponga los datos de los proveedores, debe conocer como comunicarse a cada proveedor individualmente.

- Conocer la tecnología (Motor de la BD)
- Conocer cómo los elementos a extraer están estructurados (diseño de la tablas dentro de la BD)

SQL

SQL

Escenario con estandarización de datos



Comparación Portal GBIF - IABIN

- La pagina principal
- La búsqueda por ocurrencias/especimenes
- La búsqueda por país
- Navegación por jerarquía taxonómica

Portales para acceso a información de biodiversidad

- INBio
  - http://atta.inbio.ac.cr/
- OTS – La Selva
  - http://sura.ots.ac.cr/local/florula3/index.htm
- ORNIS
  - http://olla.berkeley.edu/ornisnet/
- MaNIS
  - http://manisnet.org/
- GBIF
  - http://www.gbif.org/

Muchas Gracias