

# Cómo citar datos de GBIF\*

## Libro blanco

### 1. Introducción

GBIF integra millones de registros, procedentes de cientos de fuentes (recursos) y proveedores distintos, bastante heterogéneos. A través del "Acuerdo de Uso de los Datos de GBIF" [Anexo I de este documento], se requiere de los usuarios de los datos el reconocimiento de los esfuerzos de aquéllos que hicieron posible la disponibilidad de esos datos. La disposición libre de los datos implica una cadena de reconocimiento de la propiedad intelectual (IPR) en la que cada parte contribuye a su manera, y debe recibir el reconocimiento adecuado. Aquellos que contribuyen con sus datos deben recibir el correspondiente crédito científico por ello.

En general, las publicaciones incluyen citas para que puedan ser usadas como referencias a otras fuentes de información. Deben hacer más fácil el acceso a esos recursos, la comprobación de los hechos y la reproducción de materiales y experimentos. Las citas a GBIF no son diferentes. Sin embargo, el hecho de que los juegos de datos servidos en Internet crezcan con frecuencia hace surgir nuevos desafíos técnicos: los registros pueden crecer, y los proveedores de los datos pueden retirar sus datos en cualquier momento.

Por muchas razones -entre las que están incluidas las referencias y citas- debe de poder identificarse cada registro individual y cada juego de datos. GBIF tiene gran interés en los Identificadores Globales Únicos (GUIDs), comparables a los números de acceso de GenBank, y en su capacidad de utilizarse en la bibliografía de una manera similar. Sin embargo, todavía se está trabajando en una solución satisfactoria basada en GUIDs en el momento de escribir estas líneas. En este texto se propone un mecanismo que daría a las citas una referencia local única en el marco del portal de datos de GBIF.

Este documento define el formato y la estructura a usar para citar datos provenientes de GBIF. Está respaldado por dos guías cortas para datos de localización (ocurrencia) de especies y de nombres, respectivamente. También se discuten en este texto algunos temas técnicos que deben tenerse en cuenta.

Los objetivos de este diseño incluyen los siguientes:

1. La consistencia en el modo de citar los datos, independientemente de las circunstancias que rodean la selección de los datos.
2. Evitar cambiar el formato que los proveedores han hecho público (demasiadas conjeturas).
3. Independencia de qué esquema de identificadores únicos se adopte en el futuro por GBIF.
4. Compatibilidad con los interfaces de los motores de recopilación de datos del protocolo de la iniciativa *Open Archives*.
5. Compatibilidad con los acuerdos existentes para compartir datos en GBIF.

### 2. Formato de las referencias

Las referencias bibliográficas en el ámbito científico tienen el formato: Autor/es, Año, Título, Publicación, Editorial. Este modelo podría valer también para los datos, pero reconstruir citas de este modo desde fuentes de datos muy heterogéneas terminaría posiblemente en fracaso. Se propone, así pues, un método más sencillo.

La correspondencia entre la forma clásica de citar descrita arriba y los componentes del modelo de información de GBIF es la siguiente:

---

\* Este documento es una traducción realizada por la Unidad de Coordinación de GBIF en España desde el **original** en inglés, aparecido en el portal de **GBIF Internacional** ([www.gbif.org](http://www.gbif.org)), publicado por **Hannu Saarenmaa** el 6 de mayo del 2005

- El portal de datos de GBIF [www.gbif.net](http://www.gbif.net) no es semánticamente un "autor", sino más bien un "editor" o un "recolector" (compilador). Estos cargos podrían ir en primer lugar en las citas tradicionales.
- El proveedor de datos original correspondería claramente al agente editorial.
- Los nombres de los recursos originales podrían corresponder a los títulos, aunque no exactamente. El título sería más bien: "Los registros 0000, 0001, ..., de <nombre del recurso>".

## 2.1. Registros individuales

Para los datos de localización (ocurrencia), el formato es el siguiente:

<Referencia a GBIF>. <Fecha y Hora>. <Referencia al proveedor>, <Referencia al recurso>, <Referencia al registro>.

En los casos en los que estén implicados varios registros de manera individual, el último elemento puede repetirse.

<Referencia a GBIF>. <Fecha y Hora>. <Referencia al proveedor>, <Referencia al recurso>, (Registros: <Referencia al registro>, <Referencia al registro>,...).

Para datos referentes a nombres, el formato es similar, pero incluyendo el nombre. También se cita al taxónomo que hizo la revisión, cuando se conoce.

<Referencia a GBIF>. <Fecha y Hora>. <Referencia al proveedor>, <Referencia al recurso>, <Nombre del taxónomo>.

## 2.2. Datos de un único recurso

La totalidad de los datos de un recurso se citan como el caso anterior, pero sin especificar ningún registro.

<Referencia a GBIF>. <Fecha y Hora>. <Referencia al proveedor>, <Referencia al recurso>

## 2.3. Todos los datos de un proveedor

La totalidad de los datos de un proveedor se citan como en el caso anterior, pero sin especificar ni recursos ni registros.

<Referencia a GBIF>. <Fecha y Hora>. <Referencia al proveedor>

## 2.4. Juego de datos de varios recursos y varios proveedores

Al contrario que los casos anteriores, que podrían extraerse directamente desde el proveedor, éste es el resultado de una consulta integrada al portal de datos de GBIF. El resultado de su almacenamiento podría ser parecido a esto (en HTML o XML):

<Referencia a GBIF>. <Fecha y Hora>.  
 <Referencia al proveedor>, <Referencia al recurso>, (Registros: <Referencia al registro>, <Referencia al registro>,...).  
 <Referencia al proveedor>, <Referencia al recurso>, (Registros: <Referencia al registro>, <Referencia al registro>,...).

Este tipo de cita podría resultar bastante larga y no ser publicable. En estos casos, y en aquellos en los que el juego de registros exacto debe estar disponible durante un periodo de tiempo más largo, sería deseable que la consulta y el resultado puedan ser citados como una entidad única. Una referencia de este tipo sería simplemente:

<Referencia a GBIF>. <Fecha y Hora>. Juego de datos archivado <Identificador de GBIF>.

### 3. Elementos individuales

La cuestión es entonces: ¿qué necesita incluirse en <Referencia a GBIF>, <Fecha y Hora>, <Referencia al proveedor>, <Referencia al recurso>, <Referencia al registro> e <Identificador de GBIF>?. Los elementos <Nombre> y <Nombre del taxónomo> son literales y no es necesario explicarlos en detalle.

#### 3.1. Referencia a GBIF

Ésta es una descripción simple e invariable del hecho de que se accedió a los datos a través de GBIF, como "Portal de datos de GBIF, www.gbif.net".

#### 3.2. Referencia a GBIF

Esto sería simplemente una reseña a la fecha en formato ISO 8601 en la que se realizó la consulta, con o sin hora del día, que excepto para registros individuales, sería el valor actual del campo *DateLastModified* (Fecha de última modificación). Por ejemplo:

2005-03-31T21:57:00Z

#### 3.3. Referencia al proveedor

Es el nombre del proveedor tal cual aparece en DiGIR o BioCAsE:

Ejemplo: Australian Antarctic Data Centre

#### 3.4. Referencia al recurso

El nombre del recurso puede incluirse en la mayor parte de los casos en el título principal. Típicamente, es el nombre de una colección o base de datos.

Ejemplo: Elephant Seal Sightings, Heard Island

Hay que resaltar aquí que los metadatos del proveedor y del recurso contienen los nombres de sus depositarios que podrían utilizarse como autores. Si se necesita definir un "autor", serían los contactos "administrativos" del recurso si están especificados. Si no hay contactos "administrativos", se deben usar los contactos de "Otros" (*Other*), o a falta de otra referencia, los contactos "técnicos". Estos nombres pueden incluir el primer o el segundo apellido, y en la mayoría de los casos también pueden adaptarse a cualquier formato requerido. Sin embargo, estos contactos no corresponden semánticamente a los autores sino más bien a los editores o recolectores (*compilers*) de la información. Así pues, no se les incluye normalmente en las referencias. Los autores se incluyen únicamente en las citas de datos de nombres.

#### 3.5. Referencia a los registros

Hay elementos en los estándares de datos que están diseñados para garantizar las referencias únicas. Para Darwin Core esto quiere decir que construimos la cita a partir de los elementos "Código de la Institución" (*InstitutionCode*), "Código de la Colección" (*CollectionCode*) y "Número de catálogo" (*CatalogNumber*). Como en el caso de (Registros: Institución A, Colección B, Números de catálogo ABC, DEF, GHI, JKL; Institución C, Colección D, Números de catálogo MNO, PQR)

Ejemplo: InstitutionCode AADC, CollectionCode Ellie-Heard, CatalogNumber 1000

Como en general, la institución y la colección se identifican respectivamente como el proveedor de datos y el recurso, sólo se necesitaría suministrar el número de catálogo como referencia del registro. Para los datos de nombres, el propio nombre podría darse como título.

Si el número de registros es demasiado grande para la publicación objetivo y/o no es necesaria la identificación individual de los registros, puede mencionarse sólo el número total de registros. Si se incluyen la totalidad de los registros de un proveedor o un recurso, puede omitirse la referencia a los registros individuales.

### 3.6. Identificador de GBIF

En esta sección se proponen temas relacionados con el almacenamiento y las referencias a juegos de datos archivados. El hecho de archivar datos procedentes de GBIF es un tema controvertido, ya que podría privar a los proveedores de su derecho a retirar los datos. Esto podría llegar a ser muy problemático en casos en los que se comparta por accidente datos "sensibles" o de especies en peligro. Así pues, es necesario señalar que aún no se ha tomado ninguna decisión en la creación de semejante tipo de mecanismo de almacenamiento.

Sin embargo, ya se están utilizando grandes juegos de datos para su análisis, y existe por ello la necesidad de almacenarlos, documentarlos y citarlos. Estas referencias pueden llegar a ser bastante considerables, de modo que no pueden publicarse usando los mecanismos descritos con anterioridad. De cualquier modo, incluso utilizando los parámetros de consulta originales, el juego de datos resultante no tiene porqué ser idéntico al obtenido en el momento en el que se realizó la consulta. Podría almacenarse el juego de datos de muchas maneras, pero la solución más sencilla pasa por almacenar las cadenas XML de consulta y de resultado.

Los juegos de datos almacenados se citarían usando el denominado <Identificador de GBIF>. Para poder hacerlo compatible con la iniciativa *Open Archive*, el formato del <Identificador de GBIF> debe respetar a la sintaxis URI (*Uniform Resource Identifier*). Es lo que denominamos GBIF\_URI. Hay que aclarar que este no se trata de un identificador global único, sino de uno local.

El GBIF\_URI debe ser corto y simple, pero debe ser capaz de producir y llevar a cabo la misma consulta de nuevo (incluso aunque sea una consulta que produzca distintos resultados en el futuro). También debe ser resistente a cambios futuros porque los usuarios lo incluirán en publicaciones escritas, y deberá devolver siempre algo comprensible cuando los usuarios realicen la consulta, aunque "comprensible" quiera decir un claro mensaje de error). Obviamente, estos URIs no podrán ser creados para cada página vista en el portal de datos de GBIF, pero debe ser posible generarlo bajo petición específica: por ej., un botón que el usuario pueda presionar o generar una petición XML. Ese evento crearía y almacenaría entonces una URI persistente en una base de datos y la devolvería al usuario o interesado.

Un GBIF\_URI basado en este modelo podría parecerse a los de estos ejemplos:

<http://www.gbif.net/record/1234567890>  
<http://www.gbif.net/resource/123456>  
<http://www.gbif.net/dataset/12345>

## 4. Ejemplos completos

Aquí tienen algunos ejemplos combinados de citas estáticas:

Portal de datos de GBIF, [www.gbif.net](http://www.gbif.net). 31-03-2005. Museum of Vertebrate Zoology, Terrestrial Vertebrate Specimens, Números de registro 20045, 25678, 31098; University of Washington Burke Museum, 120 registros.

Portal de datos de GBIF, [www.gbif.net](http://www.gbif.net). 31-03-2005. Catalogue of Life Partnership, Integrated Taxonomic Information System.

Portal de datos de GBIF, [www.gbif.net](http://www.gbif.net). 31-03-2005. Field Museum of Natural History, 10 registros; Museum of Vertebrate Zoology, 204 registros; Royal Ontario Museum, 1 registro; University of Washington Burke Museum, 36 registros; University of Turku, WWF Peru, 10 registros.

Las citas estáticas dadas como ejemplo pueden ser tediosas de construir manualmente, y de modo que sería mejor que fuesen generadas por alguna herramienta al efecto.

## 5. Recomendación para el campo *Citation* de los metadatos de DiGIR

Los metadatos de DiGIR incluyen un campo denominado *Citation* (referencia, cita) para los recursos. Usar el campo *Citation* sería una buena alternativa para manejar las referencias, pero requiere un cierto trabajo de homogeneización. En el momento de escribir esto, el uso del campo es muy irregular entre los proveedores y los recursos. Trataremos aquí el mejor uso que se podría dar a ese campo para mitigar este problema.

Primero: con el campo *Citation* no pueden utilizarse directamente los formatos descritos con anterioridad, ya que el usuario podría acceder directamente al proveedor de datos sin pasar por el portal de datos de GBIF. Segundo: son los responsables de los proveedores y los recursos los que conocen perfectamente cuál es la función de las distintas personas que producen los datos.

Así pues, el formato adecuado para una referencia construida a partir de esta información es mucho más parecida a una cita tradicional con autores, año, título y editor. Por ej: "Smith, A., Turner, B., 2003-2005. Institución X, Colección Y, Base de datos del Taxon Y". Si toda esta información está disponible -con una etiqueta que indique que tiene un formato correcto- el portal de datos de GBIF podría enviarla y podrían distribuirse citas construidas de esta manera.

Los servicios de comprobación y validación de GBIF podrían incluir una revisión del texto de este campo como parte del proceso de ayuda a la conexión a los nuevos proveedores, y del mismo modo, los proveedores existentes deberían ser orientados sobre el tema como parte del proceso normal de soporte.

## 6. Debate

No hay muchos ejemplos de otros portales sobre cómo pueden manejarse las citas de datos primarios. La mayoría del resto de los portales sólo citan al propio portal. Creemos que un modelo de referencia como ésta no cumpliría los requerimientos de los acuerdos de uso e incorporación de datos de GBIF. Los acuerdos establecen claramente la necesidad de reconocer los esfuerzos de los proveedores de datos. Los proveedores de datos son a menudo tan sólo cuerpos legales que publican los datos, pero no son dueños de los datos en el mismo sentido que los gestores de recursos (=bases de datos, colecciones). Quizá debería revisarse ese punto de los acuerdos.

El propósito de estas referencias es capacitar al lector/consumidor para conseguir la fuente de la información en cuestión. En el caso de bases de datos activas, esto incluye varios desafíos. No puede esperarse de los proveedores de datos de GBIF que mantengan los datos accesibles de manera indefinida. Todo lo contrario: pueden retirarlos en cualquier momento. Así pues, las referencias estáticas tal y como se presentan más arriba no aseguran la recuperación de los datos.

Las referencias dinámicas como el <Identificador de GBIF> podrían permitir, teóricamente, el acceso al material original almacenado bajo esa referencia. Sin embargo, la garantía de persistencia de dichas referencias y del material tras éstas debe planearse con cuidado. Cuestiones como los datos "sensibles" o la autoridad de los proveedores de datos deben aclararse y debe buscarse el consenso en ellas. Estos requerimientos entran claramente en conflicto y puede que requieran la revisión del acuerdo de provisión de los datos de GBIF. Este documento no asume que el servicio esté ya en funcionamiento, o que vaya estarlo en algún momento en GBIF. Este tipo de servicios podrían darse desde servicios externos de almacenamiento.

En otras comunidades podemos encontrar ejemplos de referencias directas a fuentes de información. En particular, la publicación electrónica de artículos científicos ha planteado la cuestión de cómo identificar los contenidos electrónicos. El protocolo de recopilación de datos de la iniciativa *Open Archive* (OAI-PMH) es un estándar para recuperar metadatos de bibliotecas (repositorios) de documentos digitales (Lagoze & al. 2004). Sería interesante añadir un interfaz XML en el portal de datos de GBIF que implemente un almacén de citas OAI-PMH, ya que posibilitaría el manejo de los juegos de datos del mismo modo que las publicaciones, y así allanar el camino para recibir el correspondiente reconocimiento científico por publicar datos.

## Referencias

Lagoze, C., Sompel, H. van de, Nelson, M. & Warner, S. 2004. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2004/10/12T15:31:00Z  
<http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

---

## **Anexo 1. Extracto del *Acuerdo de uso de los datos de GBIF*.**

3. Con el fin de posibilitar la atribución del uso de los datos a los dueños de los mismos, el identificador de propiedad de los datos debe mantenerse con todos y cada uno de los registros.
  4. Los usuarios deben reconocer públicamente, tanto el uso de los datos, como a los proveedores de datos cuya información sobre biodiversidad hayan usado. Los proveedores de los datos podrán requerir atribuciones adicionales para colecciones específicas dentro de sus instituciones.
- 

*Versión 0.2, Borrador, Hannu Saarenmaa 2005-02-10*

*Versión 0.3, Borrador, Hannu Saarenmaa 2005-03-08, based on input by Donald Hobern*

*Versión 0.4, Borrador, Hannu Saarenmaa 2005-03-29, based on Open Archives Initiative materials*

*Versión 0.5, Borrador, Hannu Saarenmaa 2005-03-31, based on comments by Donald Hobern, Jim Edwards, Per de Bjørn, Meredith Lane*

*Versión 0.7, Borrador, Hannu Saarenmaa 2005-04-01, based on comments from the staff*

*Versión 0.8, Borrador, Hannu Saarenmaa 2005-04-08, based on comments from the staff*

*Versión 0.9, Borrador, Hannu Saarenmaa 2005-04-08, grammatical corrections by Meredith Lane*

*Versión 0.11, Borrador, Hannu Saarenmaa 2005-04-13, comments by Jim Edwards*

# How to cite GBIF data

## White paper

### 1. Introduction

GBIF integrates millions of data records from hundreds of rather heterogeneous different sources (resources) and providers. Users of that data are required by [GBIF Data Use Agreement \[Annex 1\]](#) to recognise the efforts of those who make the data available. Making data available involves a value chain of IPR where each party has contributed something, and should be acknowledged as appropriate. Those who make data available should get scientific credit for doing so.

In general, citations are meant to be used in publications as references to other information resources. They should facilitate accessing these resources, checking the facts, and reproducing materials and experiments. Citations of GBIF data are no different. However, technical challenges arise from the fact that datasets served on Internet often grow, records can change, and the data providers can withdraw their data at any time.

For many reasons, including citations, individual data records and arbitrary datasets should be possible to identify. GBIF is interested in globally unique identifiers (GUIDs), comparable to GenBank's accession numbers, that would be used in biodiversity literature in similar fashion. However, a solution for the GUIDs is still being worked on and is not available at this writing. A mechanism is discussed here that would give citations a local unique reference within the GBIF data portal.

This document defines the formats and structures for citing GBIF data. It is supported by two short guidelines for occurrence and names data, respectively. The paper also discusses some technical issues that must be considered.

Goals in this design include the following:

1. Consistency in the form of citation regardless of the circumstances behind the selection of the data.
2. Avoidance of reformatting what the providers have made available (too much guesswork).
3. Independence of what particular unique identifier scheme may be adopted later by GBIF.
4. Compatibility with machine interfaces of [Open Archives Initiative Protocol for Metadata Harvesting](#).
5. Compatibility if the existing data sharing agreements of GBIF.

### 2. Format of the citations

Scientific citations normally take the form of Author(s), Year, Title, Reference, Publisher. This might work for data as well, but reconstructing such a citation of very heterogeneous data sources is probably going to fail. Therefore a simplified form is sought for below.

Mapping between the above classic form of citations and the components of the GBIF information model is as follows:

- GBIF Data Portal [www.gbif.net](http://www.gbif.net) is not semantically an "author", but an "editor" or "compiler". Such entities can be in first position in traditional references.
- The sourced Data Provider is clearly the publisher.
- Names of the Resources might match titles, but not quite. Title is rather a phrase like "Data records 0000, 0001, ..., from <resource name>".

#### 2.1. Individual record

For occurrence data the form is like this.

<GBIF citation>. <Datetime>. <Provider citation>, <Resource citation>, <Record citation>.

In cases where several records are concerned individually, the last element can be repeated.

<GBIF citation>. <Datetime>. <Provider citation>, <Resource citation> (Records: <Record citation>, <Record citation>, ...).

For names data, the form is similar but includes the name. Also the taxonomist who made the revision is recognised when that is known.

<GBIF citation>. <Datetime>. <Name>. <Provider citation>, <Resource citation>, <Taxonomist name>.

## 2.2. Data from a single resource

All data from a resource is like above, but without specifying any records.

<GBIF citation>. <Datetime>. <Provider citation>, <Resource citation>.

## 2.3. Data from an entire provider

All data from a provider is like above, but without specifying any resources or records.

<GBIF citation>. <Datetime>. <Provider citation>.

## 2.4. Set of records from many resources and many providers (dataset)

Unlike the above, which also could be retrieved independently from a provider, this is a result of an integrative query to GBIF data portal. The result of the storage would be like this (in HTML or XML)

<GBIF citation>. <Datetime>.  
<Provider citation>, <Resource citation> (, Records: <Record citation>, <Record citation>, ...);  
<Provider citation>, <Resource citation> (, Records: <Record citation>, <Record citation>, ...);  
...

Such a citation can get quite long and may not necessarily be publishable. In those cases, and where the exact dataset must be available over a longer period, it would be desirable that the query and the result be stored and referenced as one entity. Such a reference would simply be

<GBIF citation>. <Datetime>. Archived dataset <GBIF identifier>.

## 3. Individual elements

The question then is what needs to be included in <GBIF citation>, <Datetime>, <Provider citation>, <Resource citation>, <Record citation>, and <GBIF identifier>. The elements <Name> and <Taxonomist name> are as written and not elaborated further below.

### 3.1. GBIF citation

This is a simple static description of the fact that these data were accessed through GBIF, like "GBIF Data Portal, www.gbif.net".

### 3.2. Datetime

This would be simply a timestamp in ISO 8601 format when the query was issued, with or without time of day, except for individual records the current value of the DateLastModified field. For example:

2005-03-31T21:57:00Z

### 3.3. Provider citation

This is the name of the provider as retrieved by DiGIR or BioCAsE:

Example: [Australian Antarctic Data Centre](#)

### 3.4. Resource citation

Resource name can in most case be included as the main Title. This is typically the name of a collection or database.

Example: [Elephant Seal Sightings, Heard Island](#)

We should note here that data provider and resource metadata does contain names of their custodians that possibly could be used as authors. If an Author identity was attainable, it would be the resource "administrative" contacts where these are specified. If there are no "administrative" contacts, "other" contacts would be used, and default to "technical" contacts otherwise. The names here can have either surname first or last, and probably could be formatted correctly in most cases. However, these contacts semantically do not correspond to Authors but rather to Editors or similar compilers. Therefore, these are not included in the citation. Authors are included only in citations of names data.

### 3.5. Record citation

There are elements in the data standards which are intended to guarantee uniqueness. For Darwin Core this will mean that we construct the citation from the InstitutionCode, CollectionCode and CatalogNumber elements, like (Records: Institution A, Collection B, Catalogue numbers ABC, DEF, GHI, JKL; Institution C, Collection D, Catalogue numbers MNO, PQR)

Example: [InstitutionCode AADC, CollectionCode Ellie-Heard, CatalogNumber 1000](#)

As the institution and collection are normally identified as the data provider and resource, respectively, only the catalog number needs to be given as the record citation. For names data, the name itself would be given as title.

If the number of records is large for the publication targeted and/or the individual identification of the records is not necessary, only the number of records may be mentioned. If the all records from a provider or resource are included, the record citation can be omitted.

### 3.6. GBIF identifier

In this section we discuss the issues related to storing and citing archived datasets. Archiving GBIF data is a controversial issue as it potentially removes from data providers their capability to withdraw data. This would be very problematic in cases where sensitive data on endangered species was accidentally shared. Therefore we must note that that no decision on building such an archiving mechanism has been made.

However, large arbitrarily constructed datasets are being used for analysis and there is a need to store, document, and cite them. Such citations can become very large and unpublishable using the other mechanisms discussed above. Even using the original query parameters, the resulting dataset is not likely to be identical to what was obtained at the time when the query was issued. Archiving the dataset can be done in many ways, but storing the incoming and resulting XML stream may be the simplest solution.

Archived datasets would be referenced using the <GBIF identifier>. In order to be compatible with the Open Archives Initiative, the format of the <GBIF identifier> must correspond to that of the URI (Uniform Resource Identifier) syntax. We call it GBIF\_URI. It must be made clear that this is not a globally/universally unique identifier but a local one.

GBIF\_URI must be simple and short, but should be able to produce to perform the same request

again (even if it is a query which may return different results in the future). It must also be future-proof because users will be publishing them in printed publications, so it must always return something sensible when users request it, even if "sensible" means a clear error message). These URIs cannot probably be rationally created for each page view of the GBIF data portal, but should be possible to generate using a specific request, i.e., a button that the user can push or a XML request generate. That event would then create and store in a database a persistent URI and return it to the user or requester.

A persistent GBIF\_URI based on that model might look like these examples:

<http://www.gbif.net/record/1234567890>  
<http://www.gbif.net/resource/123456>  
<http://www.gbif.net/dataset/12345>

#### **4. Full examples**

Now we can give some combined examples of static citations:

GBIF Data Portal, [www.gbif.net](http://www.gbif.net). 2005-03-31. Museum of Vertebrate Zoology, Terrestrial Vertebrate Specimens, Record numbers 20045, 25678, 31098; University of Washington Burke Museum, 120 records.

GBIF Data Portal, [www.gbif.net](http://www.gbif.net). 2005-03-31. Catalogue of Life Partnership, Integrated Taxonomic Information System.

GBIF Data Portal, [www.gbif.net](http://www.gbif.net). 2005-03-31. Field Museum of Natural History, 10 records; Museum of Vertebrate Zoology, 204 records; Royal Ontario Museum, 1 record; University of Washington Burke Museum, 36 records; University of Turku, WWF Peru, 10 records.

The static citations given in examples can be tedious to construct manually, and therefore could best be generated by some appropriate tools.

#### **5. Recommendation for DiGIR Citation metadata field**

DiGIR metadata includes a Citation-field for the resources. Using the Citation-field would be good alternative way for handling citations, but it will take some standardisation work. At this writing the use of that field is very inconsistent across the providers and resources. To alleviate that problem, we discuss here what would be the best use of that field.

First, the above formats cannot be applied for the Citation-field as a user can access the data provider directly without going via the GBIF Data Portal. Second, the provider and resource owners know exactly what are the roles of various people in producing these data.

Therefore an appropriate form for this citation is much closer to a traditional citation with authors, year, title and publisher. E.g. "Smith, A., Turner, B., 2003-2005. Institution X, Collection Y, Taxon Y Database". If this is available, with a flag denoting well-formedness, GBIF data portal could forward it and the constructed citations could be dropped.

GBIF data validation services could include a review of this text as part of the process each new provider is helped to connect, and existing providers should be advised on this as part of the regular process of giving them feedback.

#### **6. Discussion**

There are not many examples on other portals how a citations of the primary data can be handled. Most of the other portals just cite to the portal itself. We think such a citation model would not fulfil the requirements of the GBIF Data Sharing and Data Use Agreements. The agreements are quite clear on the need to recognise the efforts of the data providers. Data providers are often just technical bodies who publish the data, but do not own the data the same way that the resource (=database, collection) custodians may. This point may have to be revisited

in the agreements.

The purpose of citations in general is to enable the reader/consumer to retrieve the source of information in question. In the situation of live databases, this poses challenges. It is not expected that GBIF data providers keep the data available indefinitely. Quite the contrary, they can withdraw it any time. Static references as presented above do not therefore always enable retrieval.

Dynamic reference like the <GBIF identifier> can potentially enable access to original material stored under such reference. However, guaranteeing the persistence of such references and the underlying material has to be planned carefully. Issues on sensitive data and data provider authority have to be clarified and agreed on. These are clearly conflicting requirements which may require revising the GBIF Data Sharing Agreement. This paper does not assume that such a service is yet in place, or will be built by GBIF. Such service could perhaps be offered by external archiving services.

In other communities there are examples of direct references to information sources. In particular electronic publishing of scientific articles has touched the issue how to identify electronic content. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a standard for retrieving metadata from digital document repositories (Lagoze & al. 2004). Adding an XML interface onto GBIF Data Portal that implements an OAI-PMH repository of citations is attractive as it could enable handling datasets the same way as publications, and hence pave the way for getting scientific merit for publishing data.

## References

Lagoze, C., Sompel, H. van de, Nelson, M. & Warner, S. 2004. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2004/10/12T15:31:00Z  
<http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

---

## Annex 1. Excerpt from **GBIF Data Use Agreement**.

3. In order to make attribution of use for owners of the data possible, the identifier of ownership of data must be retained with every data record.
  4. Users must publicly acknowledge, in conjunction with the use of the data, the data providers whose biodiversity data they have used. Data providers may require additional attribution of specific collections within their institution.
- 

*Version 0.2, Draft, Hannu Saarenmaa 2005-02-10*

*Version 0.3, Draft, Hannu Saarenmaa 2005-03-08, based on input by Donald Hobern*

*Version 0.4, Draft, Hannu Saarenmaa 2005-03-29, based on Open Archives Initiative materials*

*Version 0.5, Draft, Hannu Saarenmaa 2005-03-31, based on comments by Donald Hobern, Jim Edwards, Per de Bjorn, Meredith Lane*

*Version 0.7, Draft, Hannu Saarenmaa 2005-04-01, based on comments from the staff*

*Version 0.8, Draft, Hannu Saarenmaa 2005-04-08, based on comments from the staff*

*Version 0.9, Draft, Hannu Saarenmaa 2005-04-08, grammatical corrections by Meredith Lane*

*Version 0.11, Draft, Hannu Saarenmaa 2005-04-13, comments by Jim Edwards*