GBIF
www.gbif.org

Data Citation Principles Workshop
May 16 – 17, 2011 IQSS at Harvard University

Deep Data Citation Mechanism and Service for Scientific Data: Defining Framework for Biodiversity Data Publishers

Vishwas Chavan
Global Biodiversity Information Facility (GBIF) Secretariat

May 16, 2011

INTERNATIONAL YEAR OF FORESTS • 2011

Outlines

1. About GBIF
2. GBIF Data Publishing Framework
3. Data Citation
4. Data Citation formulations
5. Waterfall Model of Data Citation

About GBIF

GBIF: Vision and Mission

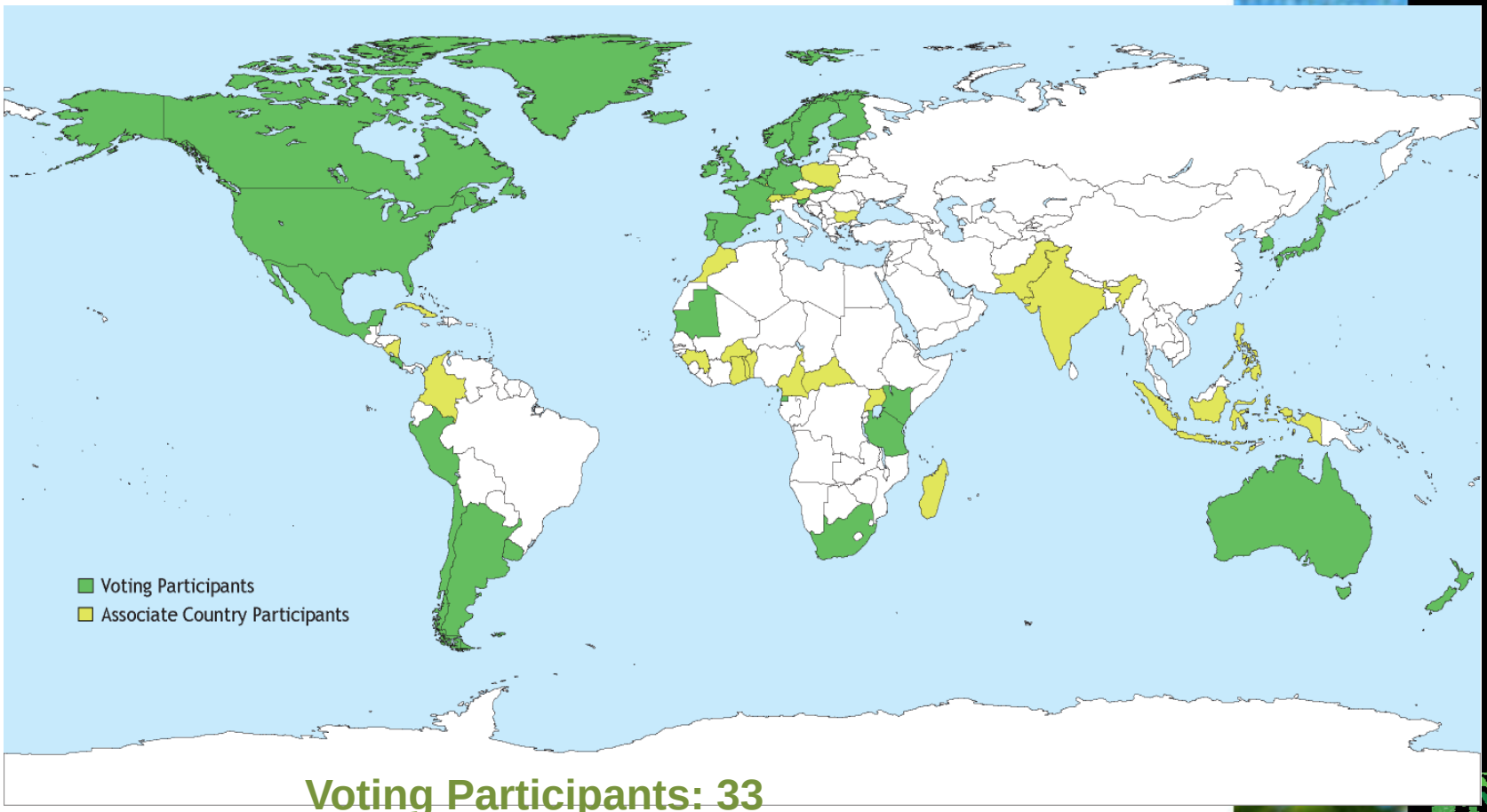
Vision:

A world in which biodiversity information is freely and universally available for science, society, and a sustainable future

Mission:

To be the foremost global resource for biodiversity information, and engender smart solutions for environmental and human well-being

GBIF Country Participants 2011



Voting Participants: 33
Associate Participants: 23
Associate Participating Organisations: 46

Growth in Data Records

GBIF
www.gbif.org

Data Publishing Framework

Why should I publish data?

- Recognition
- Opportunities
- Investment

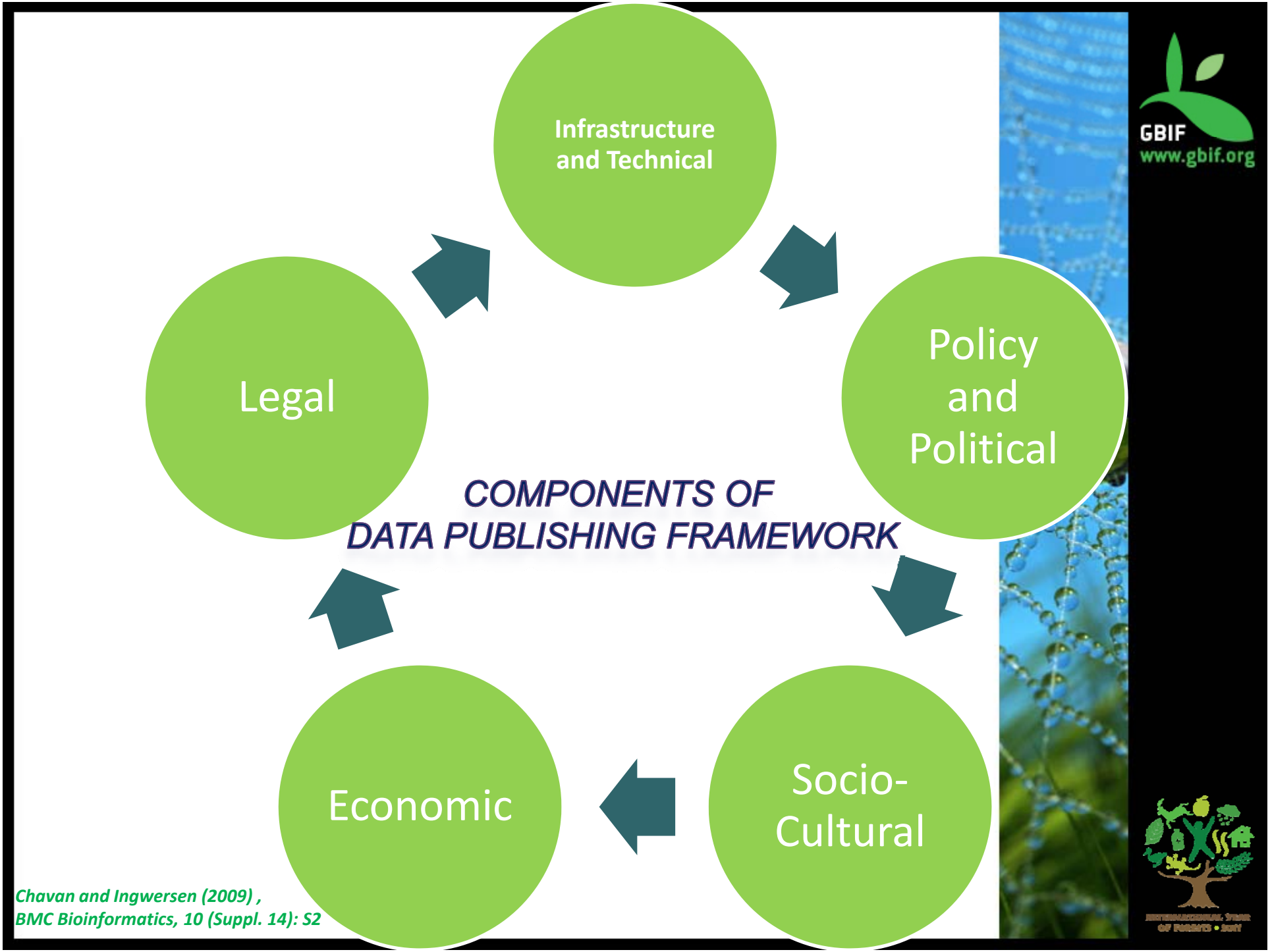
What is there for me?

GBIF
www.gbif.org

Data Publishing Framework

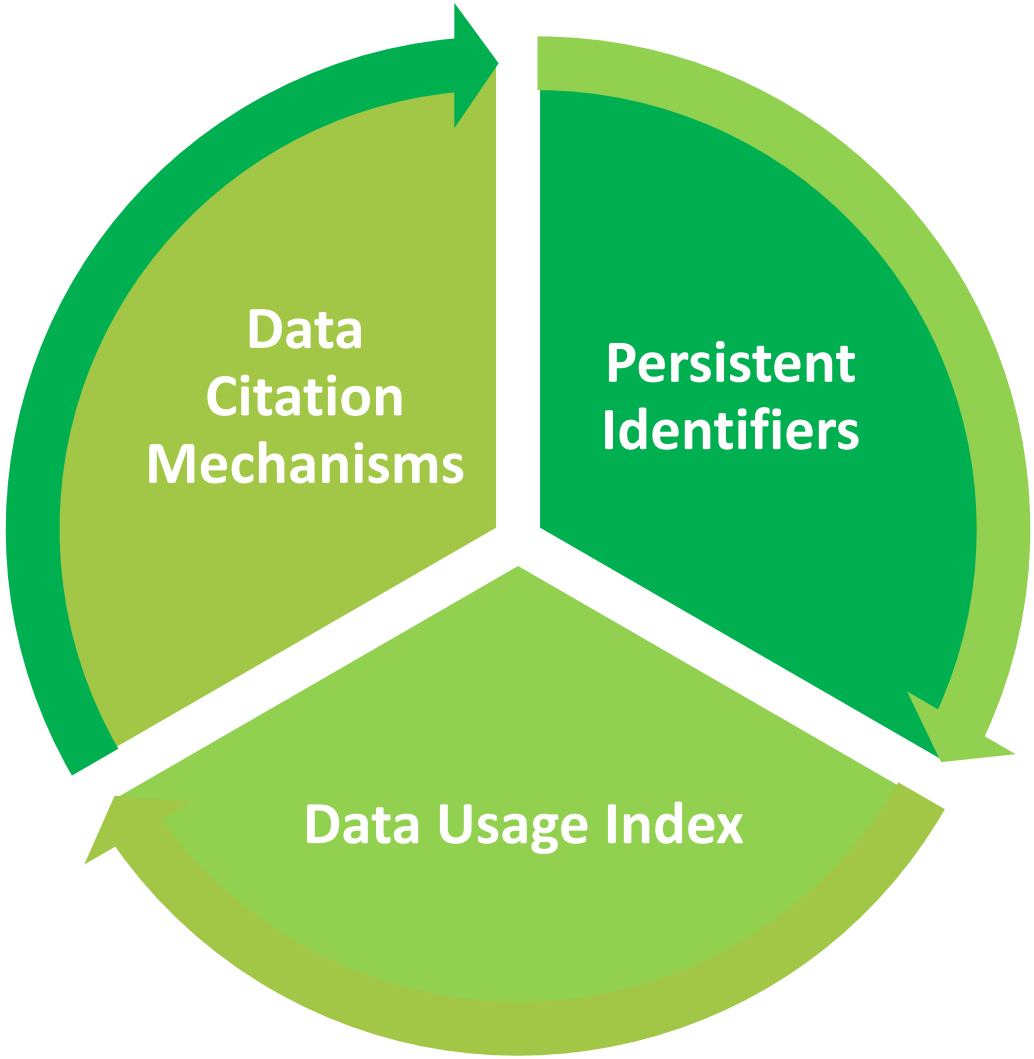
- ✓ **Cultural change towards 'free and open access' to biodiversity data**
- ✓ **Addresses social, technical, and policy concerns**
- ✓ **Answer 'What is there for me?' for ALL**

The Data Publishing Framework is defined as environment conducive to enabling free and open access to the world's primary biodiversity data. The core purpose of the framework is to overcome socio-political, technical-infrastructural, policy-political, legal and economic investment barriers or impediments affecting the discovery and publishing of data



Chavan and Ingwersen (2009) ,
BMC Bioinformatics, 10 (Suppl. 14): S2

DPF: Core Technical Components



Chavan and Ingwersen (2009) , BMC Bioinformatics, 10 (Suppl. 14): S2

Data Citation

Data Citations today: Example

Source: GBIF Data Portal, data.gbif.net
Search string: Panthera tigris
Search results: 696 records, from 37 datasets, published by 31 Data Publishers
Date: Thursday, 4 November 2010, Time: 10.03.30

Existing Data Citation style

Please cite this data as follows:

(accessed through GBIF data portal, Mammal specimens, http://data.gbif.org/datasets/resource/559)
(accessed through GBIF data portal, Vertebrate specimens, http://data.gbif.org/datasets/resource/541)
(accessed through GBIF data portal, Natural History Museum Rotterdam, http://data.gbif.org/datasets/resource/693)
(accessed through GBIF data portal, Database Schema for UC Davis Wildlife museum, http://data.gbif.org/datasets/resource/736)
(accessed through GBIF data portal, UNSM Vertebrate Specimens, http://data.gbif.org/datasets/resource/812)
.....
.....
.....

Un-answered facts

- What was the search string?
- How many records were retrieved?
- How many Data Publishers contributed to the data?
- When search was carried out?
- Who is the original contributor of the data?
- Who played what role from collection to publishing?
- How can I retrieve the same result?

Data Citation: What is needed?

- ✓ **Deep data citation mechanism**
  - ✓ **Recognise ALL with their roles**
  - ✓ **Multilayer citation – producer, publisher, aggregator, curator**
  - ✓ **Cascading Citations - citations within citations**
- ✓ **Data Citation Service**
  - ✓ **Resolve citation any time**
  - ✓ **Discover the underlined data**

GBIF
www.gbif.org

Data Citation: Challenges

- ✓ Dealing with dynamic streaming data?
- ✓ Resolving to human or machine interpretable description of object?
- ✓ Need for registry of name spaces?
- ✓ Can metadata standards support multiple GUIDs?
- ✓ Failure to enforce data citation as mandatory step in Publishing cycle

GBIF
www.gbif.org

Waterfall model of data citation

Data Citation formulations

Types of Publishers

- ✓ **Publisher (individual)**
- ✓ **Publisher (group of individuals)**
- ✓ **Institution or Research Group or Consortium**

Release / Update frequency

- ✓ **One time release**
- ✓ **Frequent updates**

GBIF
www.gbif.org

Data Citation formulations.....

Publisher (individual) one time data release

Publisher (YEAR), <Title of the data resource>, <total nos. of records>, published <modes of publishing>, <Primary access point>, released on<release date>, <Persistent Identifier>.

Rumble KJ (1998). Cephalopods of North America. 10023 records, published online, http://www.rumblejk.org/CephNA/, released on 31/12/1998, doi:10.4000/iisc.0.00.36.

GBIF
www.gbif.org

Data Citation formulations.....

Publisher (group of individuals) frequent updates

Publisher 1, ... and Publisher n <YEAR). <Title of the data resource>, <total nos. of records>, published <modes of publishing>, <Primary access point>, first released on<release date>, <current version no. or last updated/released on (date)>, <Persistent Identifier>.

Remsen D, Bello J, Sheldon S, Raymond M, and AJK Arino (2005 -). Fishes of the Cape Cod Region, MA,USA. 70089 records published online, http://www.remsen.net/capecodfishes/, first released on 17/05/2005, last updated on 10/10/2010, doi: 11.3389/mbl.1.11.131.

GBIF
www.gbif.org

Data Citation formulations.....

Institute/Research Group/ Consortium – frequent updates

<Publisher as Institution / Research Group / Consortium> <YEAR (Year first published / released -)>, <Title of the data resource>, <total nos. of records>, <Contributed by contributor 1(role), contributor 2 (role)..... contributor n(role)>, <published (modes of publishing)>, <Primary access point>,<Version no., or last updated/released on (date)>, <Persistent Identifier>.

Smithsonian National Museum of Natural History (2002 -), Museum Collection Records: Mammals. 579257 records. Contributed by Helgen KM (Principal Investigator, cutrator, author), Gordon LK (manager, author, curator), Peurach SC (author, manager), Potter CW (manager, author), Carleton MD (curator), Maldonado JE (author, developer), Wilson DE (curator, author), Thorington Jr RW (curator, author, validator), Ludwig CA (manager, developer, author), Lunde DP (author). Published online, http://collections.nmnh.si.edu/search/mammals/, first released on 12/02/2002, last updated on 15/09/2010, doi:17.3377/smi.8.57.965.

GBIF
www.gbif.org

Waterfall model for Data Citation

- ✓ Cascading citations
- ✓ Citations within citations
- ✓ Recognising roles in data management life cycle
- ✓ Three types of citations
  - ✓ Publisher determined citations
  - ✓ User driven citations
  - ✓ Composite citations
- ✓ Use of Persistent Identifiers (PI)
  - ✓ Persistent Identifiers for each citation types
  - ✓ Support multiple types of PIs – Handles, ARK, PURL, URN, LSID, DoI etc.
- ✓ Data Citation service
  - ✓ Registration service: assign PI to citations
  - ✓ Resolver service: resolve citations from PIs

GBIF
www.gbif.org

Waterfall model for Data Citation: exemplification

Source: GBIF Data Portal, data.gbif.net
Search string: Panthera tigris
Search results: 696 records, from 37 datasets, published by 31 Data Publishers
Date: Thursday, 4 November 2010, Time: 10.03.30

User Search

http://data.gbif.net (2010). user doi:09.1111/gbif.9.11.444.

User driven citation using Persistent Identifier

- resolve to full composite citation and/or
- snapshot of resultant data can be accessed

http://data.gbif.net (2010). Search string:Panthera tigris, 696 records, contributed by 37 data resources, user doi: 09.1111/gbif.9.11.444, accessed on 04/11/2010, 10:03:30. (data resources: doi: 09.1111/lsu.9.11.559, unr:lsid:msu.org:observation:541, http://nhmr.nl/ark:/1205/693xz693, http://hdl.loc.gov/ucd/736, http://purl.unsm.org/unsm/812, urn:gnhm:0-486-1047, ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ...).

Medium sized composite citation

resolve to full composite citation including detailed Publisher determined citation in cascading manner

GBIF
www.gbif.org

Waterfall model for Data Citation: exemplification

Full length composite citation

http://data.gbif.net (2010). Search string:Panthera tigris, 696 records, contributed by 37 data resources, user doi: 09.1111/gbif.9.11.444, accessed on 04/11/2010, 10:03

User driven citation

1. Louisian State University (2007), Museum of Natural Science: Collection of Mammal, 36000 records. Contributed by Patterson DN (**Institutional dataset, onetime release, doi** Sandeep PK (author, curator), Fieldman LN (author, developer), Remsen D (curator, validator), published online http://www.museum.lsu.edu/MNS/mammcoll.hml, released on October 2007, doi:09.1111/lsu.9.11.559.
2. Michigan State University (2001 -), MSU Vertebrate Collection, 76523 records. Contributed by Cook DK (Principal Investigator, author, cu **Institutional dataset, frequent update, lsid** developer), Lane MP (manager, author, curator)....., Morris JH (curator), published online http://musuem.msu.edu/ResearchandCollections/DVNH, first released on 01/10/2001, last updated on 18/01/2010, urn:lsid:msu.org:observation:541.
3. Cursada PK, Bello J, and AJK Moelicker (2006), Natural History Museum Rotterdam: Mammal collection, 1123 records, published online, http://www.nlbif.nl/nhmr_mc/, released on 7 July 2006, http://nhmr.nl/ark:/1205/693xz693 **Multiple authors, frequent update, ARK**

.....

.....

Single author, frequent update, handel

.....

37. Rumble KJ (1998 -). Vertebarte collection of Rumble 1960-1999. 786 records, published online, http://www.sbnature.org/rumble_collection/, first released on 13/09/1998, last updated on 27/01/2010, http://hdl.oclc.gov/sbnature/5678.

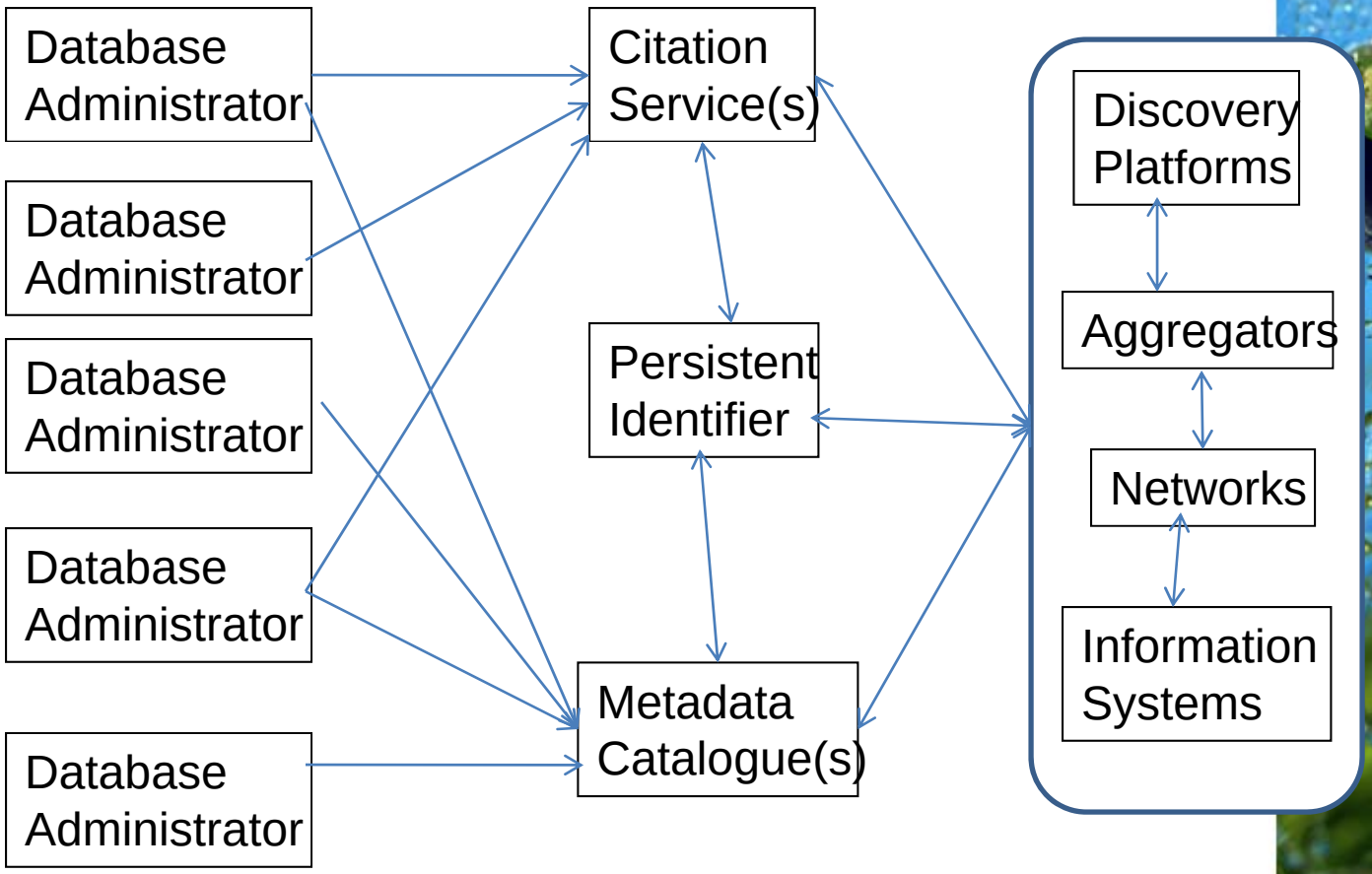
Waterfall model for Data Citation: How will this happen?

- ✓ **Publisher determined citations**
  - ✓ **Detailed citation as part of metadata document, and/or**
  - ✓ **Register citation at 'Citation Service'**
  - ✓ **Persistent Identifier is assigned to metadata document and/or citations alone**

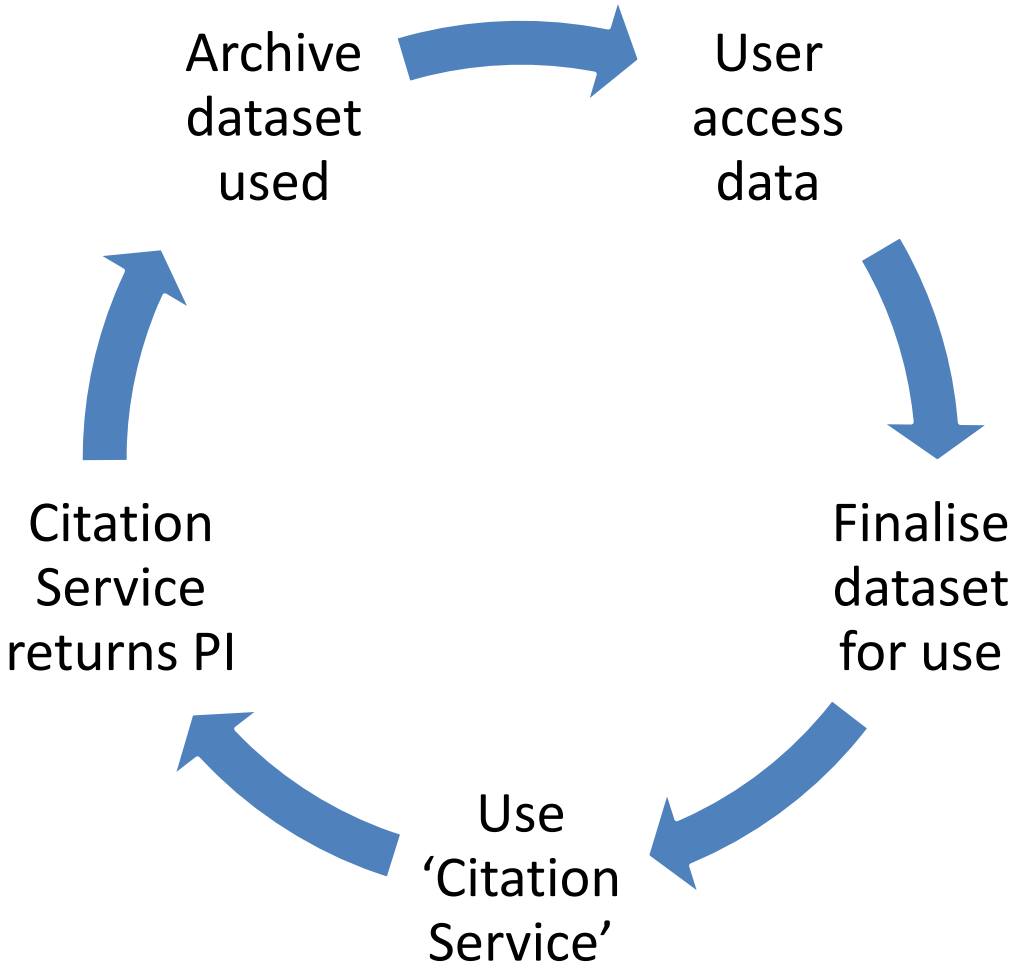
- ✓ **User driven citations**
  - ✓ **Search data through Publisher access point**
    - ✓ **Single dataset – Search result together with Publisher determined citation**
    - ✓ **Multiple datasets – Search result together with all datasets**
  - ✓ **User write 'user driven' string of citation**
  - ✓ **User register citation with 'Citation Service'**
  - ✓ **User archive snapshot of search linked to 'user driven citation'**

GBIF
www.gbif.org

Process for Publisher determined citations



Process for User driven citations



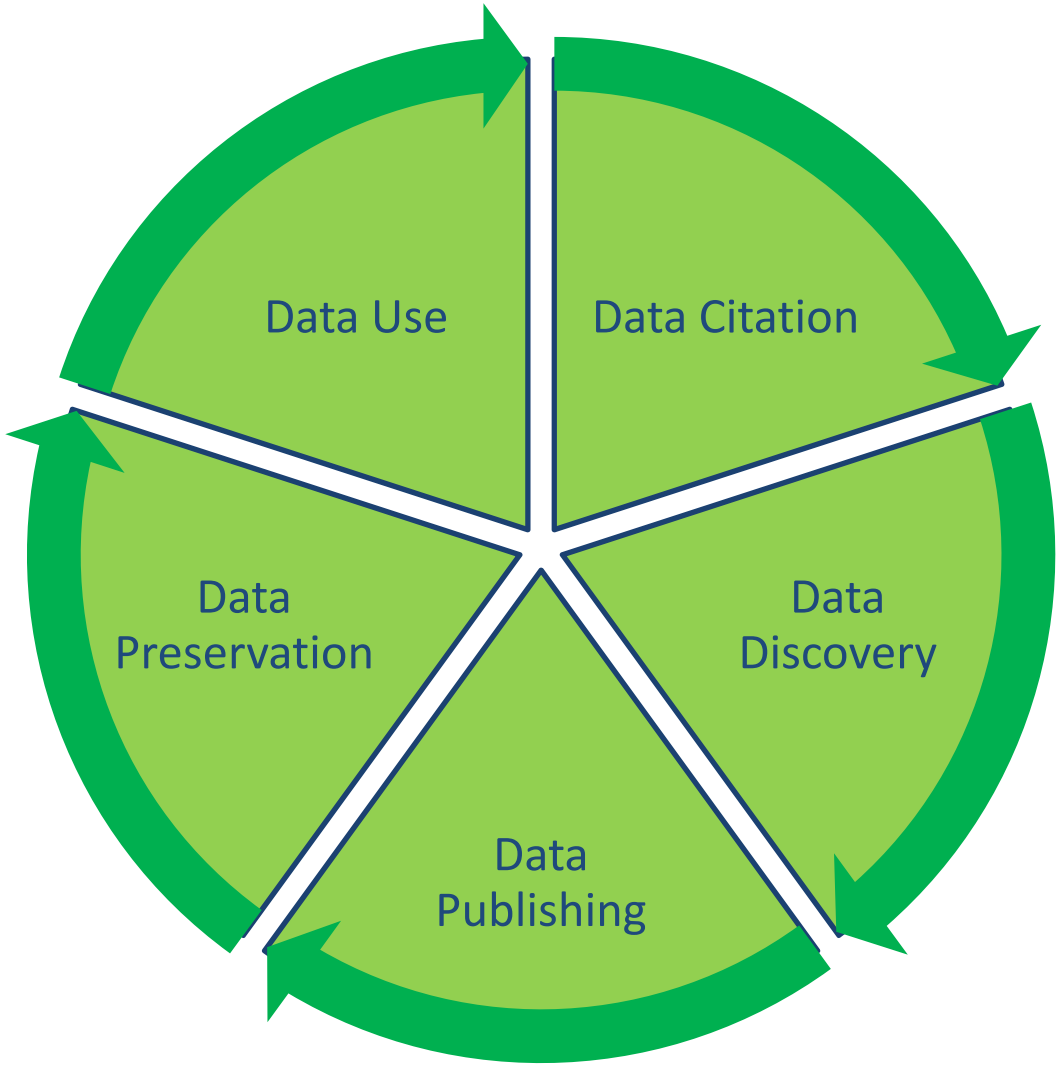
GBIF
www.gbif.org

Wish List for Data Citation

- ✓ Best practice guide for data citation
- ✓ Persistent identifiers to datasets
- ✓ Credit to all players from data producers to publishers, aggregators etc.
- ✓ All levels of granularity and combinations
- ✓ With or without annotations
- ✓ Link between traditional literature and data
- ✓ Coordinated citation support for ALL
- ✓ Research metrics for datasets

GBIF
www.gbif.org

Impact of Data Citation



GBIF
www.gbif.org

Data Publishing = Scholarly Publishing!

Email: vchavan@gbif.org