WHO'S ABSENT?

Is it You?
What’s Missing From ALL the Portals.

Using data we have to uncover data we don’t.

Sharon Grant
Janeen Jones
Kate Webbink
Pete Herbst
Rob Zschernitz
Rusty Russell

TDWG, 2017
T.B.D.

1. Can DwC together with the NCD be used to provide greater visibility of undigitized occurrences?

2. In addition, can these data serve as a means to further refine existing digitized occurrences?
Big Data Digitized.

At time of writing there are over:

- **849 million** occurrence records in the Global Biodiversity Information Facility (GBIF) portal ([gbif.org](http://gbif.org))
- **104 million** on the iDigBio site ([idigbio.org](http://idigbio.org))
- **71 million** in the Atlas of Living Australia ([ala.org.au](http://ala.org.au))
- **20 million** in VertNet ([vertnet.org](http://vertnet.org)).
Data Aggregators - where does all the data go?

GBIF, Global Biodiversity Information Facility

VeriNe

iDigBio, Integrated Digitized Biocollections

Atlas of Living Australia, ala.org.au

fossilworks

european biodiversity portal

India Biodiversity Portal
Data Aggregators - where does all the data go?

- GBIF: 849,747,902
- VeriNe: 20,082,238
- iDigBio: 104,495,464
- Atlas of Living Australia: 71,959,723
- Fossilworks: 1,338,423
- EU BON: 616,121,819
- India Biodiversity Portal: 1,302,446
Big Data Not Yet Digitized.

In 2000 there were an estimated 2.5 billion(ish) specimens in the world’s natural history collections.

(Bruce A. Stein, Lynn S. Kutner, Jonathan S. Adams - 2000 - Nature)

Today’s "Backlog" is:

2.5 billion(ish) - 850,000,000 = 1.65 billion(ish)
Big Data Not Yet Digitized.
What’s Out There That Can Help.

Records created during or after an acquisition to establish ownership (provenance) of objects to the institution.

These can be things like:
- accession records
- inventories
- invoices
- deed of gifts, etc.
Registrars, Archivists, Collections Managers

What We Have That Can Help.

Accession Card

Z No. 7147 - Invertebrates
Date: 11/4/1942
FIELD MUSEUM OF NATURAL HISTORY

Transaction: Gift
Credit to: H. S. Dyke
Address: Field Museum of Natural History - Insects

Recorded from: H. S. Dyke
Address: Field Museum of Natural History - Insects

Collections: H. S. Dyke

Date collected: March 6 April 1942
Locality: Chicago area

Description:
2 lots of land shells, containing 2 species. Value (actual): $5

Number of specimens: 14 (2 in lot)
Catalog number: Field 7147
Value (actual): $5

Acknowledgement: To
Notes:
<table>
<thead>
<tr>
<th>Accession No.</th>
<th>1793</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field Museum of Natural History—Anthropology</td>
<td></td>
</tr>
<tr>
<td>Received from</td>
<td>CHARDIN, Teilhard de</td>
</tr>
<tr>
<td>Collected by</td>
<td>same</td>
</tr>
<tr>
<td>Locality</td>
<td>China</td>
</tr>
<tr>
<td>People</td>
<td>early (Paleolithic)</td>
</tr>
<tr>
<td>Object</td>
<td>artifacts</td>
</tr>
<tr>
<td>Gift</td>
<td>X</td>
</tr>
<tr>
<td>Purchase</td>
<td>/</td>
</tr>
<tr>
<td>Catalog Number</td>
<td>201501-615</td>
</tr>
<tr>
<td>When Received</td>
<td>1927-28</td>
</tr>
<tr>
<td>Remarks</td>
<td></td>
</tr>
<tr>
<td>Price</td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>Data</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Collector</td>
<td>Chardin, Tielhard de</td>
</tr>
<tr>
<td>Locality</td>
<td>China</td>
</tr>
<tr>
<td>Age/Period</td>
<td>Early (Paleolithic)</td>
</tr>
<tr>
<td>Catalog numbers</td>
<td>201501-615</td>
</tr>
<tr>
<td>Description</td>
<td>artifacts</td>
</tr>
</tbody>
</table>
Data Entry and Research
## Lots more Information!

<table>
<thead>
<tr>
<th>Field</th>
<th>Data</th>
<th>DwC</th>
<th>NCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collector</td>
<td>Henry Field, P. Tielhard de Chardin</td>
<td>recordedBy</td>
<td></td>
</tr>
<tr>
<td>Locality</td>
<td>England, France, Germany, Switzerland, Greece, Belgium, Czechoslovakia, Gibraltar, Egypt, Palestine, China, Japan, Java, Australia, Africa.</td>
<td>verbatimLocality</td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>France, Germany, Switzerland, Greece, Belgium, Czechoslovakia, Egypt, Palestine, China, Japan, Australia</td>
<td>country</td>
<td></td>
</tr>
<tr>
<td>Continent</td>
<td>Europe, Africa, Asia, Australasia</td>
<td>continent</td>
<td></td>
</tr>
<tr>
<td>Age/Period</td>
<td>Magdalenian, Neanderthal, Chellean, Solutrean, Arugnacian, Acheulean, Austrian, Eneolithic, Mousterian, Pliocene, Oligocene, Quaternary, Sands, Azilian</td>
<td>E.g. earliestPeriodOrLowestSystem</td>
<td></td>
</tr>
<tr>
<td>Catalog numbers</td>
<td>201501-615</td>
<td>catalogNumber</td>
<td></td>
</tr>
<tr>
<td>Count of Objects</td>
<td>114</td>
<td>individualCount</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>flint and bone implements, casts of implements and casts of human remains.</td>
<td>description</td>
<td></td>
</tr>
</tbody>
</table>
Putting It Into Theory.

NCD & DwC & AcE
Natural Collections Descriptions (NCD)
Darwin Core (DwC)
Darwin Core (DwC) & extensions
Putting It Into Theory.

NCD & DwC
Putting It Into Theory.
Acquisition Data
Acquisition Extension for NCD (AcE)?
Proposed Fields

- RECORD LEVEL
- COLLECTION
- ACQUISITION
- LOCATION
- TAXON
- EVENT
- GEOLOGICAL CONTEXT
### Fields That Already Exist Somewhere

- **RECORD LEVEL**
  - institutionCode; collectionCode

- **COLLECTION**
  - description

- **TAXON**
  - taxonID | scientificNameID | scientificName | higherClassification | kingdom | phylum | class | order | family | genus | subgenus | specificEpithet | infraspecificEpithet | taxonRank | verbatimTaxonRank | vernacularName | nomenclaturalCode

- **EVENT**
  - eventDate | verbatimEventDate | habitat

- **LOCATION**
  - locationID | higherGeographyID | higherGeography | continent | waterBody | islandGroup | island | country | countryCode | stateProvince | county | municipality | locality | verbatimLocality

- **GEOLOGICAL CONTEXT**
  - geologicalContextID | earliestEonOrLowestEonothem | latestEonOrHighestEonothem | earliestEraOrLowestErathem | latestEraOrHighestErathem | earliestPeriodOrLowestSystem | latestPeriodOrHighestSystem | earliestEpochOrLowestSeries | latestEpochOrHighestSeries | earliestAgeOrLowestStage | latestAgeOrHighestStage | lowestBiostratigraphicZone | highestBiostratigraphicZone | lithostratigraphicTerms | group | formation | member | bed
New Fields?

- **LOCATION**
  - **NEW: region** - A spatial region or named place. The place the specimens/artifacts/objects originated roughly. This could be anything from a specific region (Porter County, Indiana), Ocean (eg Southern) to the world at large (Worldwide), because the verbatimLocality field is full of randomness.

- **ACQUISITION**
  - **NEW: basisOfAcquisition** - controlled vocabulary - Accession, Inventory, Invoices, Deed of Gift
  - **NEW: acquisitionID**: The institutions identifier (preferably unique) for the record.
  - **NEW: cataloguedCount** - The number of individuals with an assigned catalog number irrespective of whether they have digital occurrence records.
  - **NEW: digitizedCount** - The number of digital occurrences associated with this record at the time of the creation of upload Ace.
  - **NEW: occurrenceIDs** - GUIDs for digitized and published occurrence records. (eg triplet, im / GUID) (sensu DwC).
  - **NEW: georeferencedCount** - The number of digital occurrence record that have both latitude and longitude.
  - **NEW: associatedMediaCount** - The number of digital occurrences with media associated (sensu DwC).
  - **NEW: collectedBy**
Putting It Into Practice.

“We Want...”
Britons

"Wants You"

Join Your Country's Army!

God Save the King

Reproduced by permission of London Opinion

I Want You

For U.S. Army

Nearest Recruiting Station
Tell me what you do and don’t have digitized...

...and tell me in the format that I need it in.

By Collection Type
By Geographic Region
Over time
By Bio Region
Monthly estimates
By Research Project
Tell me what you do and don’t have digitized...

- Grant Applications
- Annual/Quarterly Board meetings
- Research inquiries
- Peer projects
- Data aggregators

...and tell me in the format that I need it in.

- By Collection Type
- By Geographic Region
- Over time
- By Bio Region
- Monthly estimates
- By Research Project

We Want Answers - NOW!
We Want Answers - NOW!

Tell me what you do and don’t have digitized...

- Grant Applications
- Annual/Quarterly Board meetings
- Research inquiries
- Peer projects
- Data aggregators

...and tell me in the format that I need it in.
We Want Answers - NOW!

Tell me what you do and don’t have digitized...

- Grant Applications
- Annual/Quarterly Board meetings
- Research inquiries
- Peer projects
- Data aggregators

...and tell me in the format that I need it in.

- By Collection Type
- By Geographic Region
- Over time
- By Bio Region
- Monthly estimates
- By Research Project
Tell me what you do and don’t have digitized...

Grant Applications
Annual/Quarterly Board meetings
Research inquiries
Peer projects
Data aggregators

...and tell me in the format that I need it in.

By Collection Type
By Geographic Region
Over time
By Bio Region
Monthly estimates
By Research Project
Already Online
Putting It Into Practice.
WHO'S ABSENT?

Is it You?
Putting It Into Practice.

Use Case
**Answers - How much backlog?**

Backlog Count = (# of items accessioned) - (# of items catalogued)

<table>
<thead>
<tr>
<th>Collection</th>
<th># of records</th>
<th># of backlog items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insects</td>
<td>390,466</td>
<td>8,559,281</td>
</tr>
<tr>
<td>Fossil Invertebrates</td>
<td>161</td>
<td>0</td>
</tr>
<tr>
<td>Anthropology</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Invertebrate Zoology</td>
<td>0</td>
<td>567</td>
</tr>
</tbody>
</table>
Answers - How “good” are your records?

RANKING

1. Digital accession record exists
2. Total Object (lots) > 0
   OR Total Items (specimens) > 0
3. Locality of Accession record Not Null
4. Catalogue # Not NULL
5. Reverse-attached catalogue records Not NULL
   (Total Count - Count of Attachments > 0)
6. Has Digital Catalogue record
7. Has Partial Data*
8. HasLatLon = Yes OR HasMultimedia = Yes
9. HasLatLon = Yes AND HasMultimedia = Yes
   AND Has **Full Data = Yes

* Partial Data (Has 3 or 4 of the following) ** Full Data (Has all 5 of the following)

TaxonRank = Family, Genus, species, subspecies or variety / State/Province Not NULL / Collector Not NULL / YearCollected Not NULL / DarCatalogNumber Not NULL
Summary

Searched by: Afrotropical

227,920 records
(164,518 FMNH records, 63,402 Naturalis records)

At least 82,870 backlogged items.