Angling for data: making biodiversity metadata more FAIR

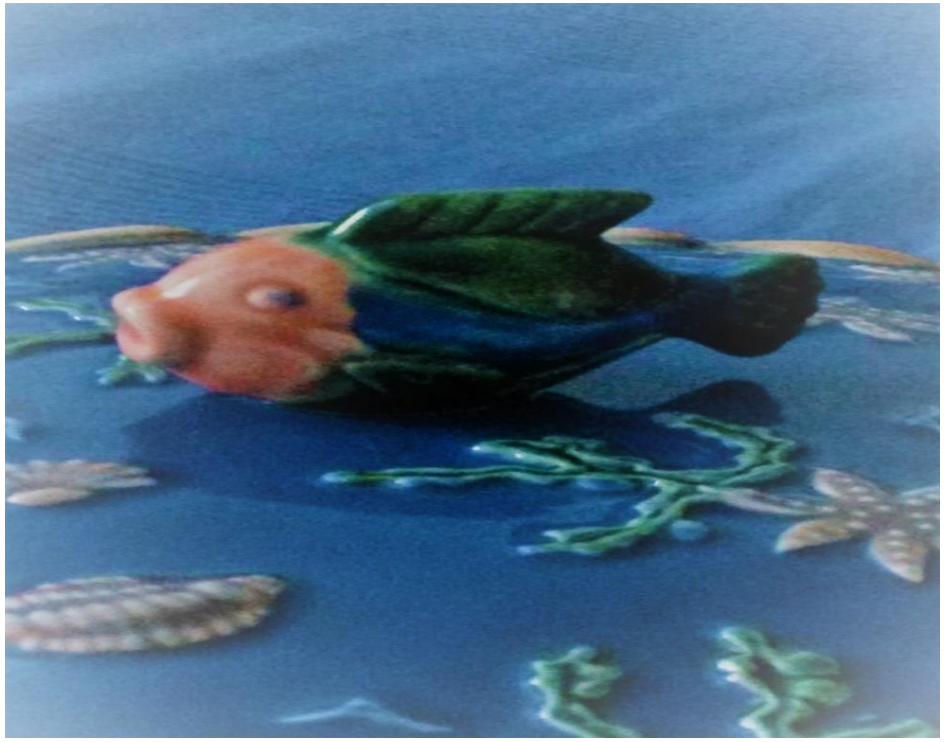
/ Joakim Philipson (Stockholm University Library)

@TDWG 2017, Ottawa, Oct. 3

Stockholm University

Go Fish!

- (Swedish, verb:) ”meta” = ’to angle’
- FAIR **meta**data will give all a larger lake to fish from, and a greater diversity of fish!



Stockholm University

Implementing the FAIR Principles?

Box 2 | The FAIR Guiding Principles

15 Criteria

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Data Archiving and Networked Services DANS

See: http://datafairport.org/fair-principles-living-document-menu and https://www.force11.org/group/fairgroup/fairprinciples

FAIR-principles (*Findable*, *Accessible*, *Interoperable*, *Re-usable*)
FAIR metrics tool from DANS (**https://goo.gl/749dmf** OSFair ws, 2017-09-08)

Stockholm University

What is FAIR? Explication attempts

FAIR = **Findable, Accessible, Interoperable, and Re-usable** metadata and identifiers:

•1. To be **Findable** any Data Object should be uniquely and persistently identifiable
•1.1. The same Data Object should be re-findable at any point in time, thus Data Objects should be persistent, with emphasis on their metadata
•1.2. **A Data Object should minimally contain basic *machine actionable* metadata that allows it to be *distinguished from other Data Objects***
•1.3. **Identifiers** for any concept used in Data Objects should therefore be **Unique** and **Persistent**
•2. Data is **Accessible** in that it can be always obtained by machines and humans
•3. Data Objects can be **Interoperable** only if:
•3.1. (Meta) data is ***machine-actionable*** ... **Metadata being machine readable is a *conditio sine qua non* for FAIRness.**
•3.3 (Meta) data within the Data Object should thus be both *syntactically parseable* and *semantically machine-accessible*
•4. For Data Objects to be **Re-usable** additional criteria are:
•4.1 Data Objects should be compliant with principles 1-3
•4.2 (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) ***linked* or *integrated, like-with-like***, with other data sources

Stockholm University

What has FAIR *metadata* to do with data quality?

Findable & Accessible

-> *potentially more eyeballs to scrutinize data, detect errors*

Wellcome Images [CC BY 4.0], via Wikimedia Commons

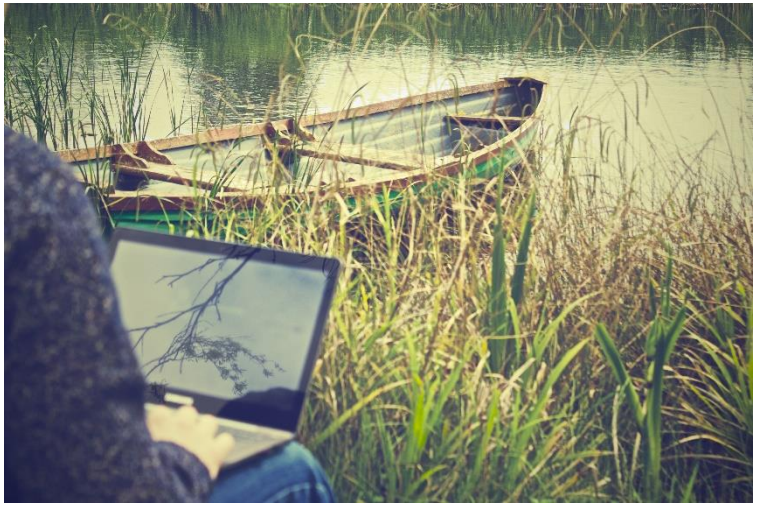


Photo: Dave Meier Stocksnap.io **CC0 1.0**

<- **Interoperable**

<- **Re-usable**

-> *more testing by different users*

Laura Russell (GBIF VertNet 2011): https://vimeo.com/album/1904479/video/40447148

Stockholm University

Wish list!

1. **Findable** – e.g. by schema.org markup of html-records (cf. Uniprot)

2. *Citation (*in several formats if applicable), easy to copy.

3. Good searchability options (indexing and filtering)

4. **Accessible** –"can be always obtained by machines and humans" ->
   a clear policy and/or demonstrated ability to always keep metadata records in the database, even for datasets that have for some reason been withdrawn.

5. Resolvable *Permanent* globally unique *Identifiers* - ***PIDs*** of standard formats (doi, hdl, …) – widely *used* -distributed (also making them more **Findable** or *'googlable'*)

6. **Interoperable** - "(Meta)data is ***machine-actionable***" ... ->
   a *metadata file in XML or JSON* for download (or harvesting via API or OAI-PMH) should be part of each dataset or package.

7. Preferably using *non-proprietary, sustainable fileformats*

8. **Re-usable** - "(Meta)data … automatically (or with minimal human effort) *linked* or *integrated, like-with-like*, with other data sources" ->
   Provide ***export*** *of metadata* to other formats (DC, DataCite, DDI, DIF, DWC, EML, RDF?)
   AND ***compliance*** with exportformat standards provided (**validation**, keeping promises)

9. **File overview and structure** (relations between files), with explicit statements of fileformats, file sizes, and possibly checksums included in metadata.

10. **Standardized licenses** (CC-BY et al., selection from drop-down menu)

Stockholm University

Wish list!

1. **Findable** – e.g. by *schema.org* markup of html-records (cf. Uniprot) -> a wider lake to fish from! (or, bigger chance to be caught):
   google it: fish egg lectin

From view-source:http://www.uniprot.org/uniprot/P68512 of db entry:

```
<main class="uniprot"
property="schema:about" resource="http://purl.uniprot.org/uniprot/P68512"
typeof="schema:CreativeWork" id="content"><section id="page-header">
<h2 class="page-title" property="schema:alternateName">UniProtKB - P68512 <span
property="schema:alternateName">(FEL_CYPCA)</span></h2>
…
<div id="content-protein" class="entry-overview-content">
<h1 property="schema:name">Fish-egg lectin</h1>
</div>
…
<a
href="http://dx.doi.org/10.1042/BJ20030413"
>Biochem. J. 376:433-440(2003)</a> [<a
property="schema:sameAs"
href="https://www.ncbi.nlm.nih.gov/pubmed/12956625"
>PubMed</a>] [<a property="schema:sameAs"
href="http://europepmc.org/abstract/MED/12956625"
```

Stockholm University

Wish list!

2. Findable – *Citation(s,* in several formats if applicable), easy to copy.

Examples from wikimedia v:

Bibliographic details for File:Sampling coral microbiome (27146437650).jpg

- Page name: File:Sampling coral microbiome (27146437650).jpg
- Author: Wikimedia Commons contributors
- Publisher: *Wikimedia Commons, the free media repository.*
- Date of last revision: 5 August 2016 01:10 UTC
- Date retrieved: 11 June 2017 12:54 UTC
- Permanent URL: https://commons.wikimedia.org/w/index.php?title=File:Sampling_coral_microbiome_(27146437650).jpg&oldid=203149246
- Page Version ID: 203149246

Citation styles for File:Sampling coral microbiome (27146437650).jpg

APA style
File:Sampling coral microbiome (27146437650).jpg. (2016, August 5). *Wikimedia Commons, the free media repository.* Retrieved 12:54, June 11, 2017 from https://commons.wikimedia.org/w/index.php?title=File:Sampling_coral_microbiome_(27146437650).jpg&oldid=203149246.

MLA style
"File:Sampling coral microbiome (27146437650).jpg." *Wikimedia Commons, the free media repository.* 5 Aug 2016, 01:10 UTC. 11 Jun 2017, 12:54 <https://commons.wikimedia.org/w/index.php?title=File:Sampling_coral_microbiome_(27146437650).jpg&oldid=203149246>.

MHRA style
Wikimedia Commons contributors, 'File:Sampling coral microbiome (27146437650).jpg', *Wikimedia Commons, the free media repository,* 5 August 2016, 01:10 UTC, <https://commons.wikimedia.org/w/index.php?title=File:Sampling_coral_microbiome_(27146437650).jpg&oldid=203149246> [accessed 11 June 2017]

FILTER APPLIED 18 SEPTEMBER 2017

Citation: GBIF Occurrence Download doi:10.15468/dl.muveie accessed via GBIF.org on 18 Sep 2017

File: 5 MB Darwin Core Archive For how long will GBIF store this data?

Involved Datasets: 137

Scientific name	Gadus morhua linnaeus, 1758

Zenodo: ->

Share

Cite as

Carozza, David A, Bianchi, Daniele, & Galbraith, Eric D. (2017). Metabolic impacts of climate change on marine fish communities and fisheries - Dataset and figure plot script [Data set]. Zenodo. http://doi.org/10.5281/zenodo.495487

Start typing a citation style...

- 3 Biotech
- 3D-Printed Materials and Systems
- 3D Printing in Medicine
- 3D Research
- 4OR
- AAPG Bulletin
- AAPS Open

But no citation help for single records or search results:

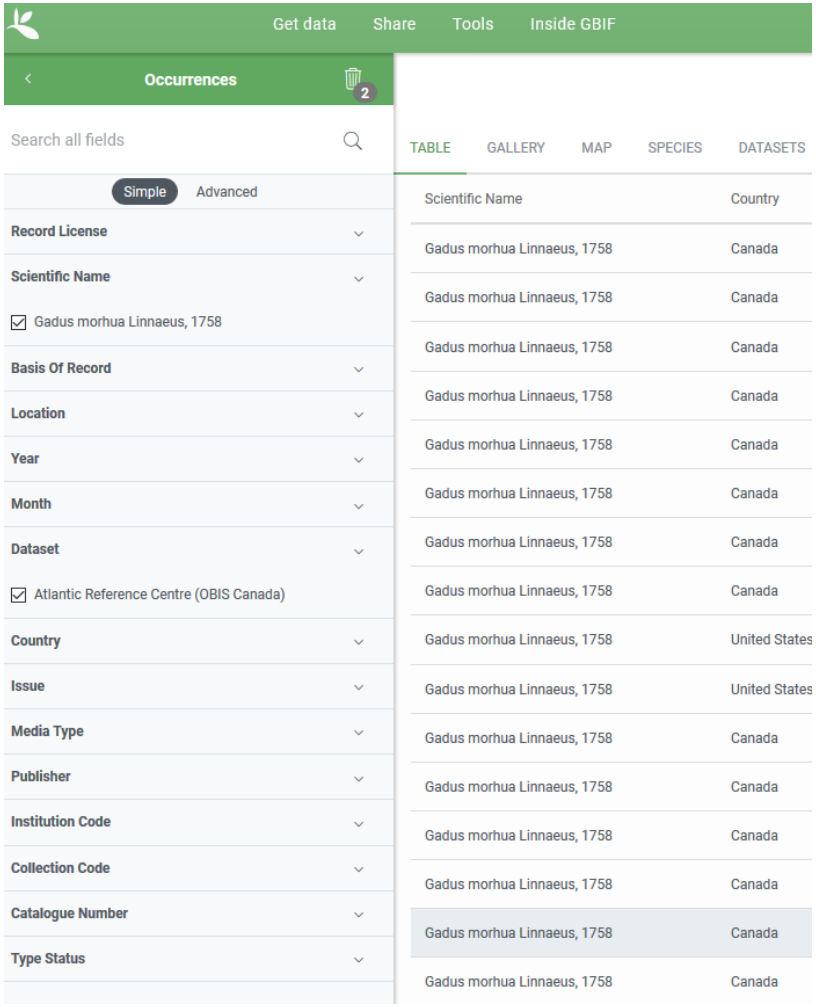
https://www.gbif.org/occurrence/17469157 – Atlantic Reference Centre, 2005-07-11

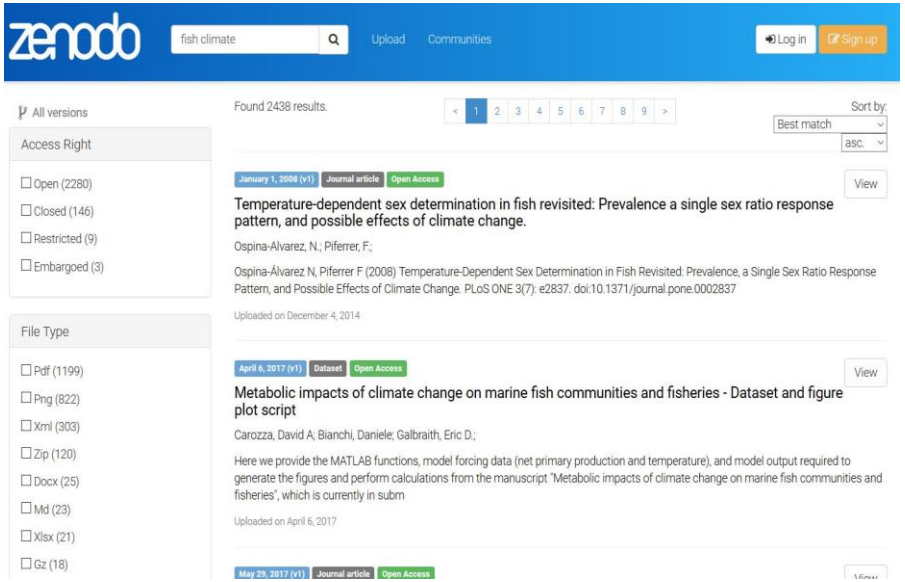
https://records.nbnatlas.org/occurrences/search?q=lsid:NBNSYS0000175392 - login req.

Stockholm University

Wish list!

3. Findable – *good searchability in database (with clear filtering options)*
Example: https://zenodo.org/search?page=1&size=20&q=fish%20climate





<- GBIF very good at this! Example:
https://www.gbif.org/occurrence/search?taxon_key=8084280

Fishbase fine-grained search-form, complex post-search-filtering:
http://fishbase.org/search.php
Search Genus+Species: *Gadus morhua* ->
http://fishbase.org/summary/Gadus-morhua.html ->

Stockholm University

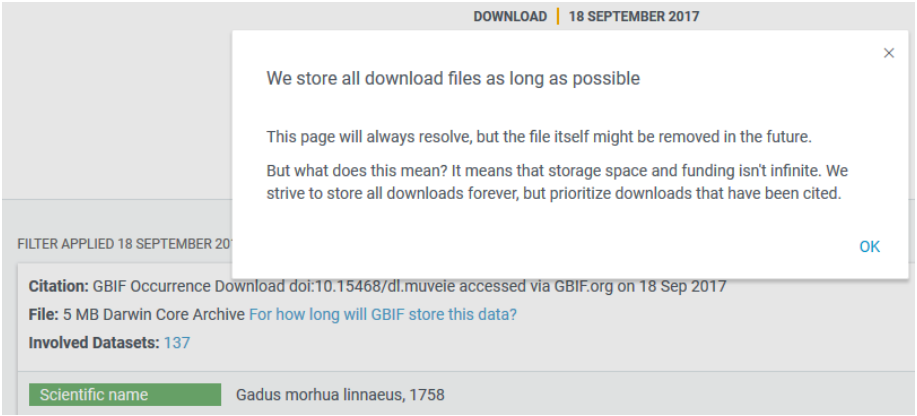
Wish list!

4. Accessible … data ”can be always obtained by machines and humans” ->
a clear policy and/or demonstrated ability to always keep metadata records in the database, even for datasets that have for some reason been withdrawn.

Few research data repositories live up to this yet. A recent study from TU Delft (https://doi.org/10.5281/zenodo.321423) showed only 1 of 37 repositories that complies:

”…To be compliant requires a clear policy statement (or various examples of data this has actually happened to) indicating that metadata is still available even if the data is removed. … only 3% of repositories are compliant (i.e. only one repository from the sample!).”

***GBIF* has something of such a declaration for downloads, promising permanent resolution of page, but does it mean that actual metadata will be preserved?**



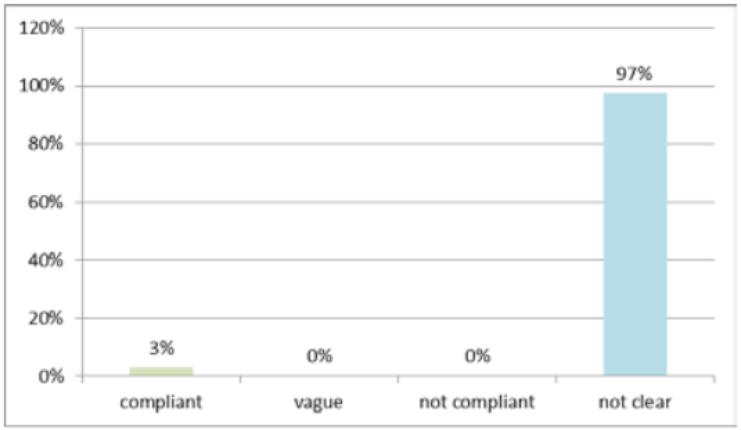


Figure 4: Compliance Rate for A4 metadata are accessible, even when the data are no longer available.

Stockholm University

Wish list!

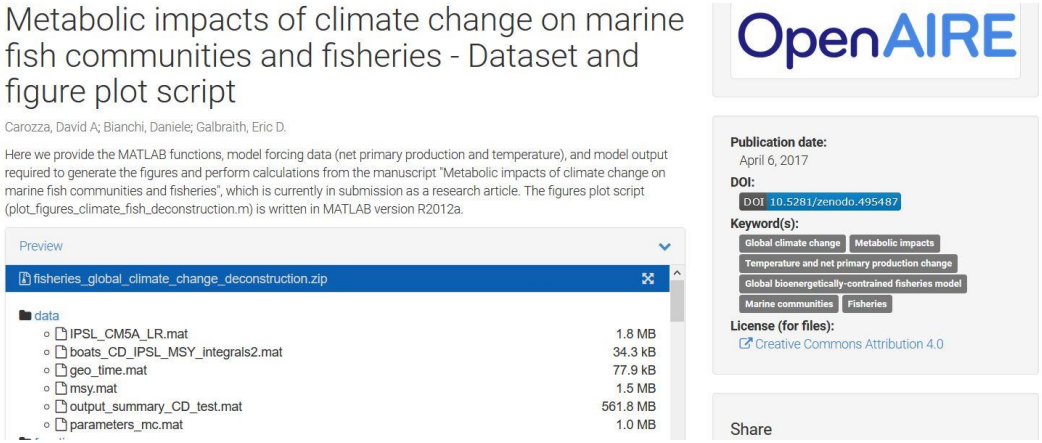
5. Accessible – Resolvable Permanent unique *Identifiers* - *PIDs* of standard formats (doi, hdl, …)

Examples: *https://zenodo.org/record/495487* = https://doi.org/10.5281/zenodo.495487

GBIF ***occurrence search*** download:DOI 10.15468/dl.muveie – 'googlable' as "10.15468/dl.muveie"

NBNatlas occurrence search: Googling: **lsid:NBNSYS0000175392** – 'googlable' and unique

But often no PIDs for single records or at least not clear what if anything could be used as a PID – record no.? apparent UUID as part of URL? etc.; not 'googlable'.



Figshare: doi:10.6084/m9.figshare.4234886.v4

Harvard Dataverse: hdl:1902.1/19629, V3

https://records.nbnatlas.org/occurrences/62be3121-cf9d-4283-ba7d-c53120fccca4

https://www.gbif.org/species/8084280

https://www.gbif.org/occurrence/17469157

Stockholm University

Wish list!

6. Interoperable - ”(Meta)data is ***machine-actionable*** ... **Metadata being machine readable is a *conditio sine qua non* for FAIRness**” ->
a *metadatafile in XML or JSON* for download (and / or harvesting via **API** or **OAI-PMH**) should be part of each dataset or package.
Example: https://www.ncdc.noaa.gov/paleo-search/study/15974

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Formerly the National Climatic Data Center (NCDC)... more about NCEI »

SEARCH RESULTS (9 STUDIES) ■ CLIMATE RECONSTRUCTIONS ONLY (1 STUDY)

- NOAA Caspian Sea-Level Changes During the Last Millennium (Naderi Beni, A.)
- NOAA Global Lake-Level Variations from 18,000 to 0 Years Ago: A Paleoclimatic Analysis (Street-Perrott, F.A.)
- NOAA LAKE STATUS RECORDS FROM EUROPE: DATA BASE DOCUMENTATION (Yu, G.)
- NOAA LAKE STATUS RECORDS FROM THE FORMER SOVIET UNION AND MONGOLIA: DOCUMENTATION OF THE SECOND

Caspian Sea-Level Changes During the Last Millennium

Originator:

Naderi Beni, A.; Lahijani, H.; Mousavi Harami, R.; Arpe, K.; Leroy, S.A.G.; Marriner, N.; Berberian, M.; Andrieu-Ponel, V.; Djamali, M.; Mahboubi, A.; Reimer, P.J.

Citation Information:

Naderi Beni, A., H. Lahijani, R. Mousavi Harami, K. Arpe, S.A.G. Leroy, N. Marriner, M. Berberian, V. Andrieu-Ponel, M. Djamali, A. Mahboubi, and P.J. Reimer. 2013. Caspian sea-level changes during the last millennium: historical and geological evidence from the south Caspian Sea. Climate of the Past, 9, 1645-1665. doi: 10.5194/cp-9-1645-2013

NOAA Study Page:

https://www.ncdc.noaa.gov/paleo/study/15974

JSON Metadata:

https://www.ncdc.noaa.gov/paleo-search/study/search.json?xmlId=13739

DIF Metadata:

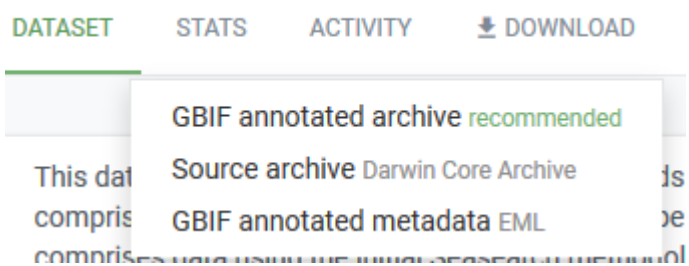
http://www1.ncdc.noaa.gov/pub/data/metadata/published/paleo/dif/xml/noaa-lakelevel-15974.xml

Download Data:

Data File	Caspian Sea Level

Lat: 50.2856, Lon: 43.8574

GBIF and biodiversity db providers do provide machine actionable (meta)data as DwCa, EML or CSV for download
+ tools for validation:



[Darwin Core Archive/EML validator]

Stockholm University

Wish list!

7. Interoperable – Preferably using *non-proprietary, sustainable fileformats*: e.g.: *.csv* instead of *.xls*

GBIF and biodiversity db providers seem to be good at this …

Resources for format control:

Sustainability of Digital Formats / Library of Congress:
https://www.loc.gov/preservation/digital/formats/intro/intro.shtml

PRONOM formatregister / National Archives, UK:
http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new

Stockholm University

Wish list!

8.(1) Re-usable - ”(Meta)data … automatically (or with minimal human effort) ***linked* or *integrated, like-with-like***, with other data sources" ->
Provide ***export of metadata*** to other formats – directly or via APIs, OAI-PMH (*DC, DataCite, DDI, DWC, EML, RDF …*)

Examples: https://zenodo.org/record/495487 ->

GBIF and biodiversity db providers –
no standard md export generally provided, except for domainspecific formats (ABCD, DWC, EML)? What about DataCite (DMK)?

GBIF species record – no export formats?
https://www.gbif.org/species/8084280
- more comprehensive than JSON via API:
http://api.gbif.org/v1/species/8084280

NBNatlas *species* records provided also as JSON and link-search to GBIF, EoL, BHL, PESI, ARKive:
https://species.nbnatlas.org/species/NBNSYS000017539

Catalogue of Life

http://www.catalogueoflife.org/col/search/all/key/Gadus+morhua – export CSV

Publication date:
April 6, 2017
DOI:
DOI 10.5281/zenodo.495487
Keyword(s):
Global climate change | Metabolic impacts | Temperature and net primary production change | Global bioenergetically-contrained fisheries model | Marine communities | Fisheries
License (for files):
Creative Commons Attribution 4.0

Share
Cite as
Carozza, David A, Bianchi, Daniele, & Galbraith, Eric D. (2017). Metabolic impacts of climate change on marine fish communities and fisheries - Dataset and figure plot script [Data set]. Zenodo. http://doi.org/10.5281/zenodo.495487
Start typing a citation style...

Export
BibTeX CSL DataCite Dublin Core JSON MARCXML Mendeley

Stockholm University

Wish list!

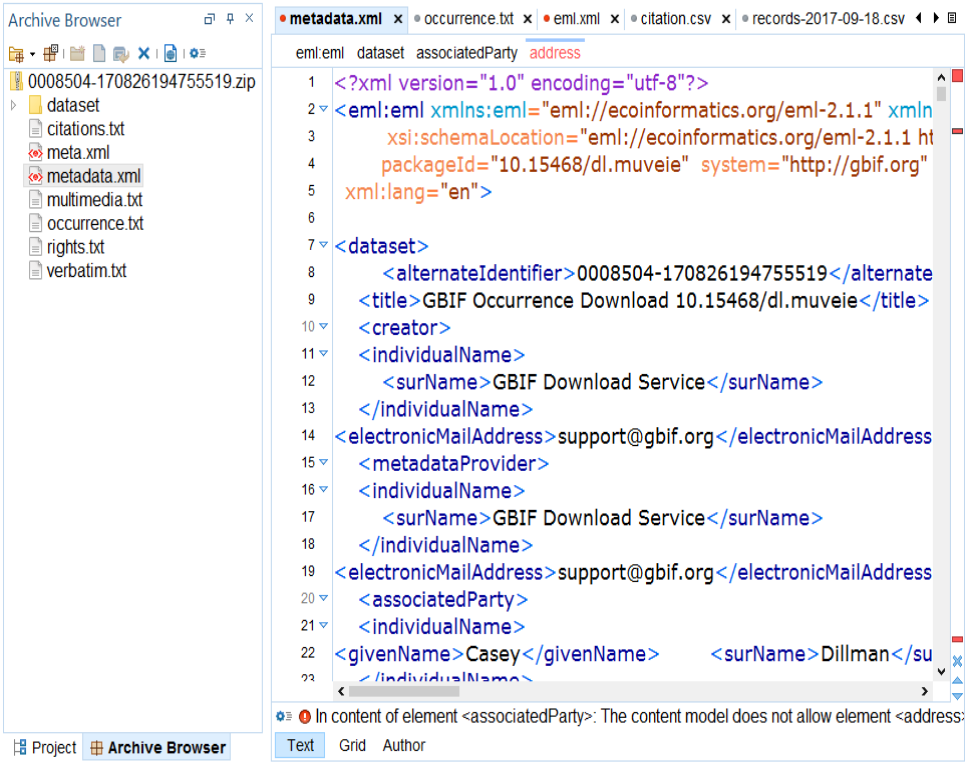
8.(2) Re-usable - *Compliance* with metadata standards offered (= ***validates*** OK for default metadata format of the db and export formats, incl. APIs / OAI-PMH); ”keeping promises”. Example: **Zenodo**

https://zenodo.org/oai2d?verb=ListRecords&metadataPrefix=datacite3&from=2017-04-04&until=2017-04-08 – missing namespace declarations for xsi and datacite; forbidden html-tags within elements, does not validate

GBIF and biodiversity db providers – not always given that metadata files comply even with “domestic formats” (DWC, EML)?

Example: *GBIF Occurrence Download 10.15468/dl.muveie* 0008504-170826194755519.zip/ *metadata.xml*

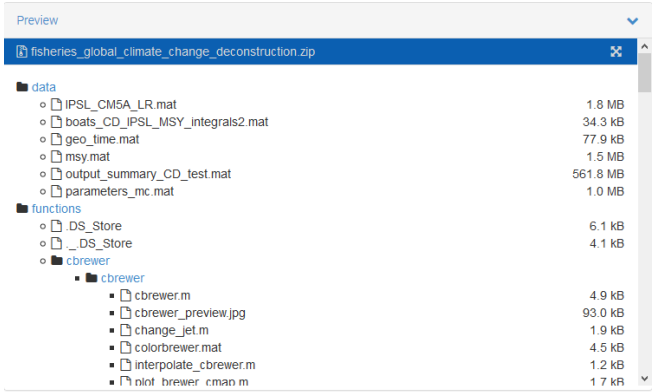
does not validate against *eml.xsd*



Wish list!

9. Re-usable - *File overview* and *structure* (relations between files), with explicit statements of fileformats, file sizes, and possibly checksums included in metadata.

Zenodo: preview and structure in html-page, but not ”machine actionable”;
not in metadata export formats(?)



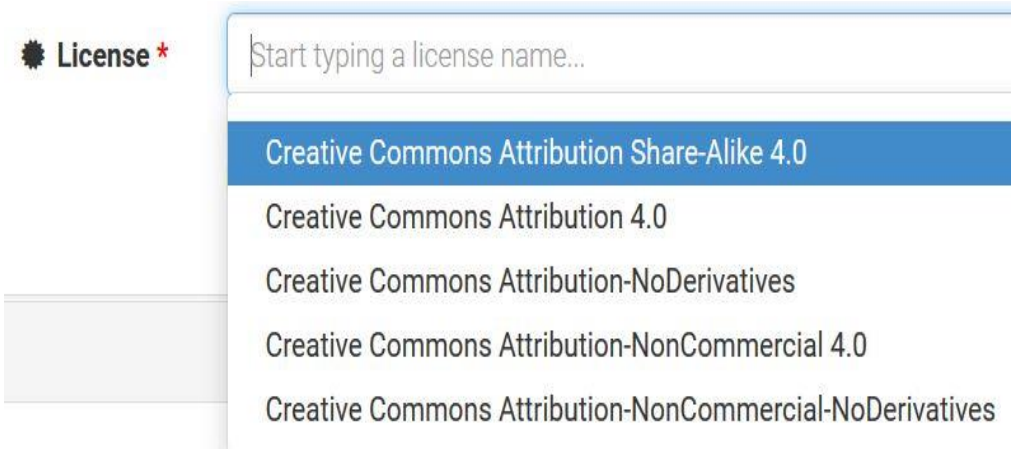
GBIF and biodiversity db providers: file structure incl. fileformats, filesize largely only through download (of zip-packages), sometimes requiring login (e.g. NBNatlas)

Stockholm University

Wish list!

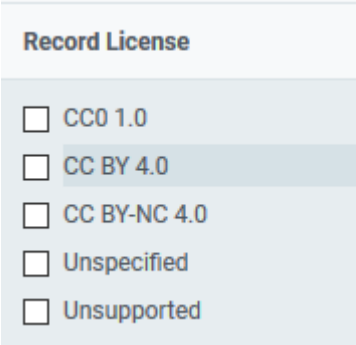
10. Re-usable - Standardized Licenses (URIs), drop-down menu at upload

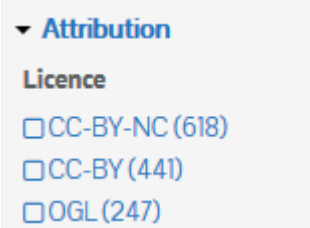
Example: Zenodo ->



Unaware how licensing works at upload to GBIF and biodiversity db providers?

Filtering option on License type in occurrence searches in GBIF and NBNatlas!





Stockholm University

Thank you!

joakim.philipson@su.se

@TDWG 2017, Ottawa, Oct. 3

Stockholm University