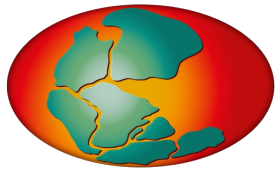# PANGAEA Collection Types: Facts and Challenges

Uwe Schindler

PANGAEA (MARUM, University of Bremen)

About

PANGAEA.
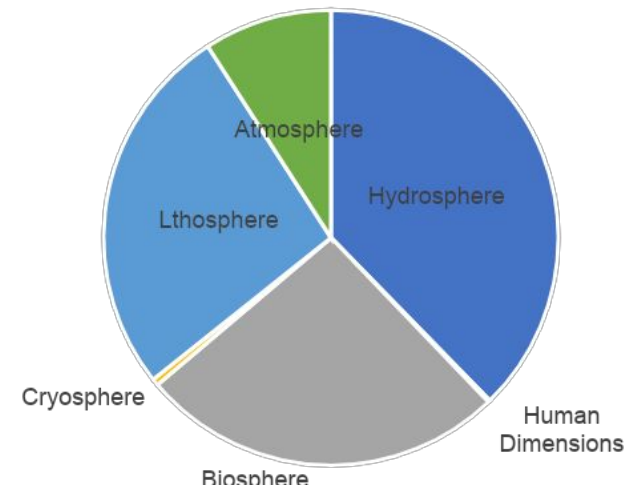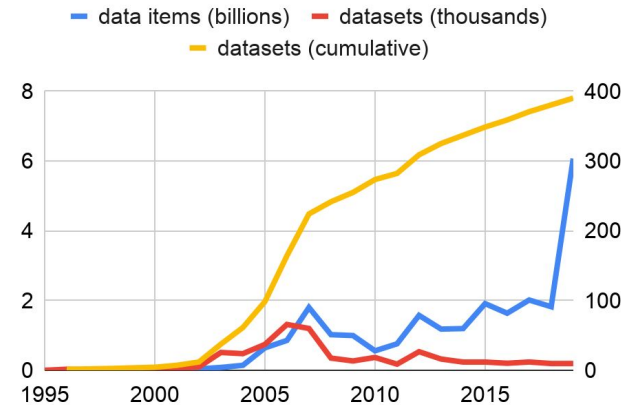Data Publisher for Earth & Environmental Science

growth

- **Integral part of science:** More than 250 European to international projects since 1995 (https://www.pangaea.de/projects)
- **highly heterogeneous & dynamic**
- **multidisciplinary**

Number of data sets	~402.000
Number of data items	>19 Billion
Data volume	<3 PB
Increase	~4% per year

>15.000 registered users
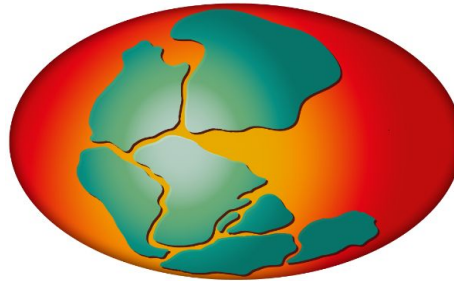12 – 15.000 unique sessions / month

PANGAEA – Data/Metadata Dissemination

Data set types

Publishable entities

- originality
  - requires more than trivial or mechanical intellectual effort
  - applies in particular for derived works
  - data production, selection/filtering, QA, processing, contextualization in general justify a publication. Example: compilations of published data sets (data products)
- threshold of originality
  - concept in copyright law that is used to assess whether a particular work can be copyrighted
- independant versus dependant publication
  - rules the way how the data are cited
- dependant
  - supplementary data (deprecated in PANGAEA)
  - single data entities as part of collections (PANGAEA publication series)

"Standard" data set

- Self-standing data set publication

- No dependencies (on other publications or datasets)

- *Optional:* relation to description of methodology or similar

- Full & simple citation with DOI name

- Single-file download possible

Example: https://doi.org/10.1594/PANGAEA.931365

Publication series

- Childs are dependent

- Parent dataset should be cited only, childs have "In:" citation

- for homogeneous data collections *(identical data templates)*

- Can be "open": New child datasets added when available

- ZIP-based download possible

Example: https://doi.org/10.1594/PANGAEA.877578

Bundled publication (1)

- Childs are dependent

- Parent dataset should be cited only, childs have "In:" citation

- Clear dependencies between child data sets:

  - it does not make sense to publish children separately because the usability is limited

  - comprehensiveness of the data sets is only ensured by supplying all data sets as a package

Example: https://doi.org/10.1594/PANGAEA.914589

Bundled publication (2)

- Data sets are typically heterogeneous

- Description of the collection, in particular title, authors, abstract. It is needed for the usability of the children!

- *Should never be an artificial construction!*

- ZIP-based download possible

Example: https://doi.org/10.1594/PANGAEA.914589

Editorial publication (1)

- Heterogeneous data collection

- Childs stay independent:

  - childs may be part of other datasets

  - no notion of collection in citation of child

- An editorial compilation of data sets for a specific purpose

Example: https://doi.org/10.1594/PANGAEA.887388

Editorial publication (2)

- The added value of such collection should be mentioned in its abstract

- DOI for citation with "(eds.)": editors ≠ individual authors of datasets

- No download as single entity possible

Example: https://doi.org/10.1594/PANGAEA.887388

Bibliography (1)

- A simple list of (heterogeneous) data set / collections titles, which have been compiled for e.g. a specific topic, but without any additional editorial processes

- Childs stay independent:

  - childs may be part of other datasets

  - childs may be other bundled publications, publication series,...

  - no notion of collection in citation of child

Example: https://doi.org/10.1594/PANGAEA.902603

Bibliography (2)

- DOI is used solely as persistent identifier, but bibliography should not be cited as replacement for the data sets in third party works!

- Citation only if the process of compilation is to be referred
  - e.g., "a collection of Arctic CLoud Observations was put together by Ehrlich A et al. in https://doi.org/10.1594/PANGAEA.902603"

- No download as single entity possible

Example: https://doi.org/10.1594/PANGAEA.902603

Feedback?

Thank You!