Critter

an integrated data warehouse solution
from spreadsheet to database to web portal

ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG
Functional Ecology | Ecosystem Functions

Bridging the Pit of Doom for Arctic Benthic Data

Jan M. Holstein[1,2], Paul Kloss[1,2], Annette Breckwoldt[1], Tom Brey[1,2], Dieter Piepenburg[1,2,3]

[1] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research Bremerhaven, Germany, [2] Helmholtz Institute for Functional Marine Biodiversity (HIFMB) at University Oldenburg,
[3] Institute of Polar Ecology of Kiel University, Wischhofstr. 1-3, 24148, Kiel, Germany

About A data warehouse for benthic biodiversity data is currently being build. It acts both as an internal workbench for ecologists to **explore**, **curate,** and **visualize** data, as well as a backbone for **complex analysis**, **story telling** and **stakeholder services**. It provides a solution to transfer spreadsheet data into a database.

Background Contemporary machine learning methods are more data hungry than ever and, in principle, biodiversity data are abundant, but transitioning from spreadsheet data management to well implemented databases is challenging. The transition costs in terms of data tidying and meta data research are significant, and data are at risk of being lost. Particularly, for historic data, which are mostly conceptualized within the context of a usually narrow defined research question, these transition costs create a **pit of doom** in which many datasets are currently being lost.

Tools are needed for…

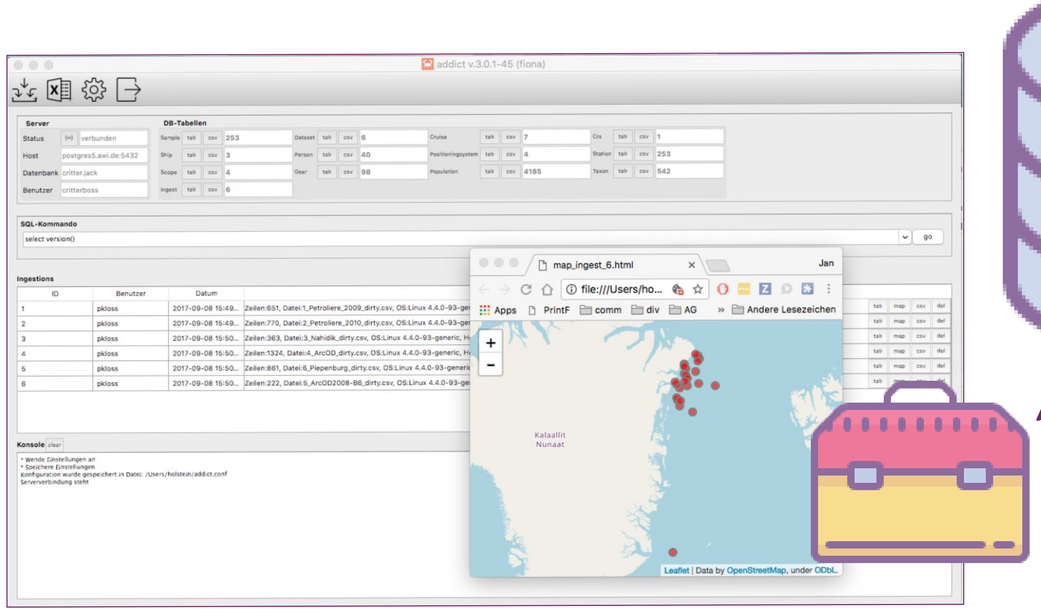
- **empowering scientists**
  to transform, ingest and curate their data thus enabling them to refurbish the data “one last time”**.**
- **data ingestion**
  to get spreadsheet data into the database.
- **quality control**
  to check data for plausibility.
- **data tidying**
  so data honour well defined standards on which subsequent services and methods can rely on.
- **data curation**
  to help curating the data and keep track of data improvements and enrichments.

The database delivers…

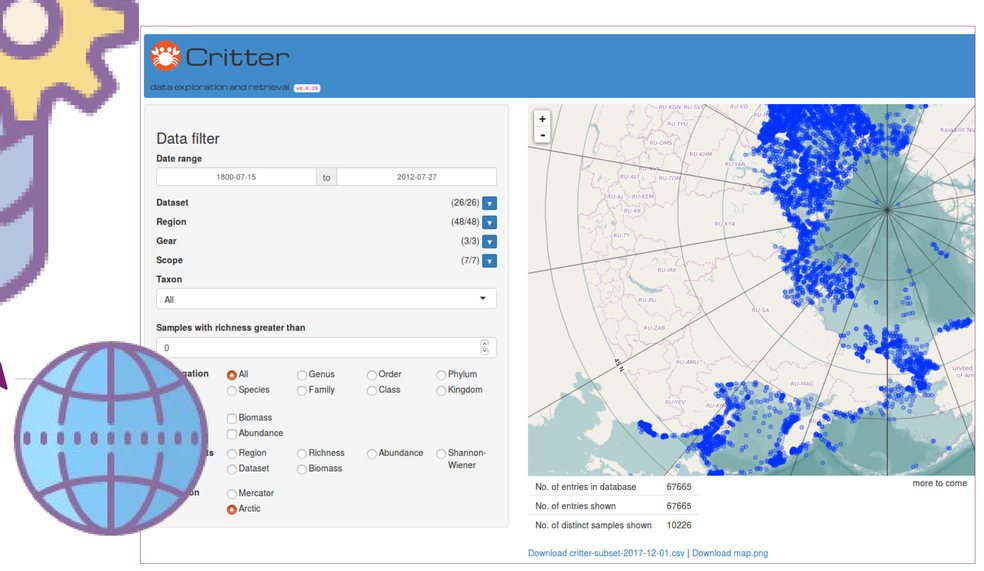
- **sophisticated data interrogation**
  through data coherence and consistency.
- **high data quality**
  by enforcing documented standards and eliminating hazards from localization, encoding, and pseudo-intelligent import functionality.
- **accessibility & security of data**
  because of professional storage and interfaces.
- **reproducible science**
  through persistent identification.
- **data science premises**
  through data coherence, reproducibility, and accessibility.

The Web portal permits…

- **easy access for anyone**
  with a web browser and an internet connection.
- **easy data exploration**
  by using the powerful R language and interactive frameworks.
- **visual data exploration**
  initially with generic display formats and statistics.
- **downloading subsets of data**
  e.g. for ecological modelling.
- **building your own service**
  for specific use cases and story telling through a well defined interface,
  for instance REST API.



data flow



Screenshot of ingest tool. It produces spreadsheet templates and import spreadsheets into the database. Automatic and manual consistency checks enable you to resolve data issues before data enter the database. On top general statistics for important tables are shown. The mid section organizes data into datasets. Data can be exported, reviewed, edited, and plotted.

Data models are the core of the information system. They control meaning and relations of data. They live in a database management system, in this case PostgreSQL 9.6.x with PostGIS extension to handle geoobjects.The toolkit is used to ingest, curate, but also export data. The webtool only extracts data for purposes like visualization and download.

Screenshot of the web-based visualisation and retrieval tool (early prototype). You can filter the data via various constraints (left), review the geographic distribution of the result (right) and inspect diverse statistics of the data (not shown) before you download the data for further processing.

Outlook

Currently, we have over 70,000 species information on occurrence, abundance or biomass for over 7000 samples all across the Arctic. We are in the process of gathering information on how users want the information presented and how they want to interact with the information system. Implementing a user-led design philosophy seems crucial but requires competences outside of the skill set of ecologists, warehouse architects and data scientists.
Next step would be expanding to North Sea data and thinking about a modular approach to plug-in methods in order to foster reproducible research.

Jan M Holstein
jan.holstein@awi.de

BREMERHAVEN
Am Handelshafen 12
27570 Bremerhaven
Telefon 0049 471 4831 3421
www.awi.de

Conservation of Arctic Flora and Fauna

ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG
Functional Ecology | Ecosystem Functions

HELMHOLTZ GEMEINSCHAFT