# n2kanalysis: a framework for automated and reproducible statistics from long-term ecological monitoring

Thierry Onkelinx

Flanders
State of the Art

## Case study long-term monitoring

NATURA 2000

- Natura 2000 (n2k) is a network of core sites
  - breeding and resting sites for rare and threatened species
  - some rare natural habitat types
  - 18% of EU's land territory and 6% of its marine territory
  - member states must report every 6 year on status and trend over the last 24 year
- **goal** of `n2kanalysis`: generate **automated** ***reproducible*** and ***traceable*** statistics

## Reproducible research

**benefits**

- repeating the analysis
  - on *same* data yields the *same* results
  - on *new* data yields *comparable* results
- allows for inspection
  - **what** is analysed and **how**
  - useful when *doubts* arise with third parties

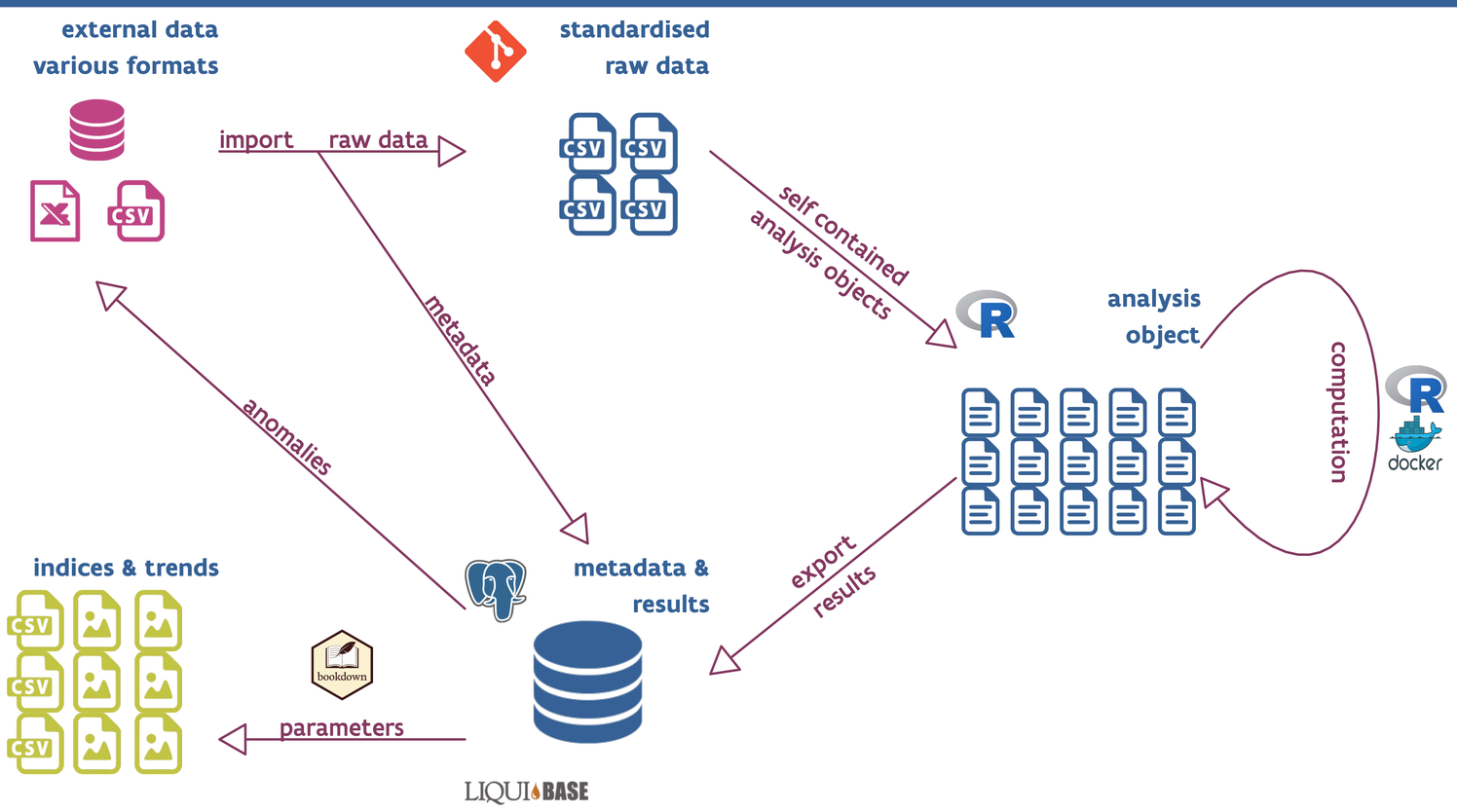
**downsides**

- a bit harder to manage
- requires more storage
- stands or falls on version control
  - code
  - data
  - software environment

## Traceable research

- retrace any parameter estimate to a specific analysis
  - including data, metadata and environment
- solution: add file & status **fingerprints** when communicating results
  - **fingerprint**: SHA1 hashes
  - **file** fingerprint: based on components which should never change in a specific analysis
    - stable metadata (model type, species group, location group, …)
    - input data
    - doubles as file name for the analysis object
  - **status** fingerprint: based on file fingerprint + changing components
    - metadata (status, used software, …)
    - fitted model

## Schematic data flow




## Data under version control

- `analysis data`
  - private git repo
  - even relative large monitoring schemes are doable
    - 100 species, 1200 sites, 27 years, 6 month
  - stable ordering of rows and columns is required
- `Results`
  - PostgreSQL database
    - only append data
    - `https://github.com/inbo/n2kresult`

## Environment under version control

- analyses run on virtual machine with Docker
- Docker image contains a fixed version of all required dependencies
- multiple versions of Docker image
  - keep old versions for older analyses
  - create new version when more recent software is required

## Analysis object as cornerstone

- S4 object `n2kModel`
- **metadata**
  - model type
  - species group
  - location group
  - import date
  - time range
  - seed
  - file & status fingerprint
  - used software + version
- **input** for analysis
  - `data.frame`
  - parent analysis when analysis depends on output of other analysis
- **fitted** model

## Anomalies?

- extreme values according to the model
  - high (low) fitted values while low (high) observed values
  - extreme hyper parameters (e.g. random intercept with large $\sigma$)
  - unstable imputations in case the analysis is based on multiple imputation
- might be due to
  - typo in data
  - correct but strange observation
  - wrong model
- inspect only $n = 10, 20, \ldots$ extreme values
  - tackle the most influential errors first
  - redo the analysis after fixing problems in the data

## Pro tip: use nominal validation status

- ***unchecked***: *default* status for all records
- ***updated***: records which have been *altered*
- ***good***: scrutinized records which are *correct* and *suitable* for the original goal
- ***abnormal***: scrutinized records which are *correct* but *not suitable* for the original goal
- ***rejected***: scrutinized records which *cannot be trusted*
- ***anomaly***: records which have not been scrutinized and flagged by an analysis as *anomaly*

## Custom R packages under version control. Available at https://github.com/inbo

**`n2kanalysis`**

- generic package, used for every monitoring
- defines **S4 classes** + validation
- **fits** the analysis objects
- extracts *model parameters* and **anomalies**

**`'xyz'analysis`**

- `'xyz'` stands for a specific monitoring scheme
- each monitoring scheme has its own package
- defines **import** of raw data and metadata
- defines how the **analysis objects** are created

**auxiliary packages**

- `n2khelper` contains generic auxiliary functions
- `n2kupdate`: export from R to database

RESEARCH INSTITUTE
NATURE AND FOREST

poster
https://doi.org/10.21436/inbop.14901626

Thierry Onkelinx
https://orcid.org/0000-0001-8804-4216

thierry.onkelinx@inbo.be
www.inbo.be