Quality **A**ssured (QA) data

Towards DOI quality of data generated at the UFZ

Mark Frenzel (Ecologist) & Thomas Schnicke (IT)

DataCite / Helmholtz Open Science Workshop
Leipzig, 12.01.2016

HELMHOLTZ
ZENTRUM FÜR
UMWELTFORSCHUNG
UFZ

QA + DOI: Best practice elements and goals

AIM: **Re-use** of data (open data)

- Recognition of **importance** of data management (attitude)
  - Top down
  - Bottom up
- **Documentation** and **selection** of **important** data sets
- **Meta data standard** for data description
- **Thesauri** (controlled vocabulary)
- Defined **workflows**
- Persistent storage / **hardware**
- Magic tools: **software** solutions

QA + DOI: Best practice elements and goals

Data publication: **DOI** for data sets

⇨**Linking data** to publications and people

↺ Feedback on attitude of data providers

Smiling data creators

Smiling users

HELMHOLTZ
ZENTRUM FÜR
UMWELTFORSCHUNG
UFZ

The UFZ situation

- >20 years of data production (exponential growth rates!)
- **Heterogeneity** of data (person-generated, device-generated; measurement, observation, modelling, data bases)
- **Documentation (meta data)** of data sets by data creators not standardized or missing
- **Long term availability** for scientific data not ensured
- **Quality control** work flows and tools partly not well-established

⇨DOI for datasets initiative started 3 years ago

(priority condition set by head of UFZ: **quality control** of data)

Challenge for Science & IT: development of **workflows**, **tools** and **attitudes!**

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

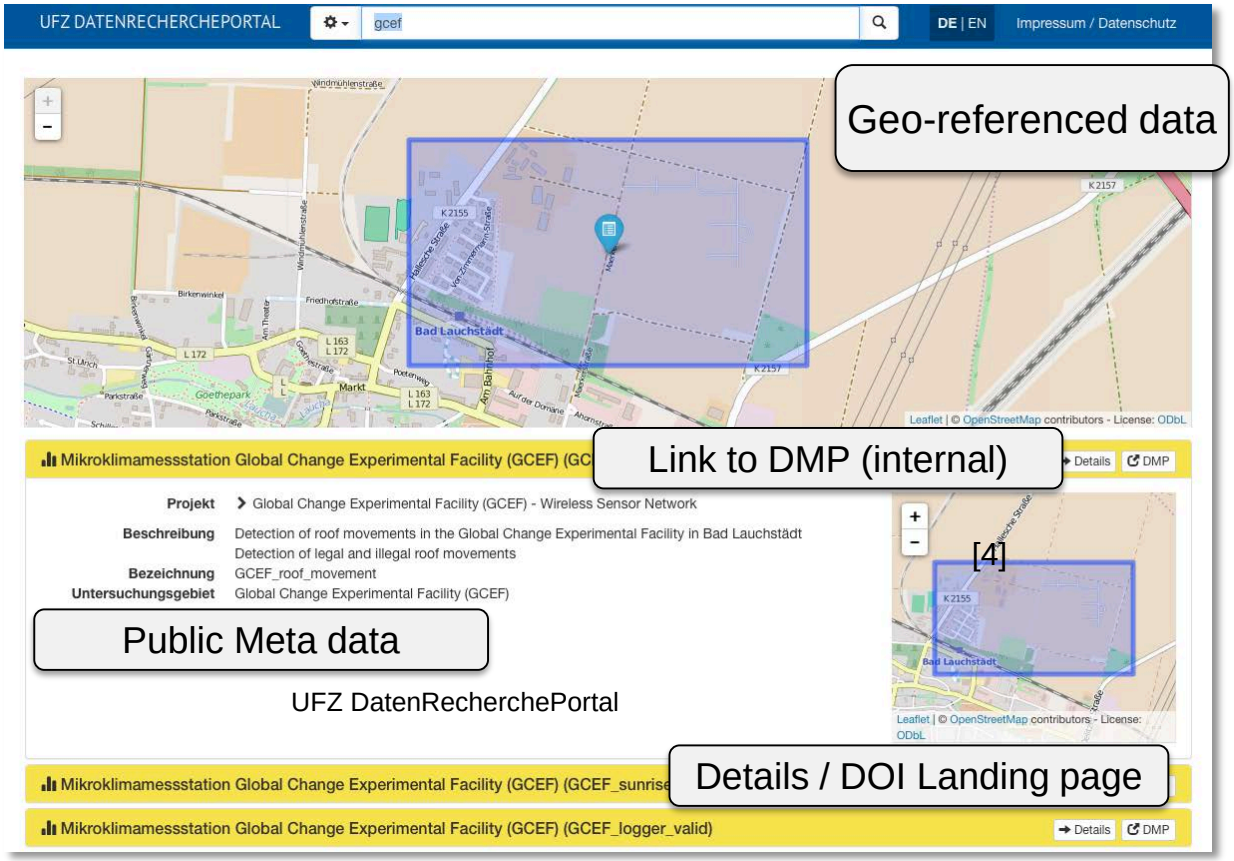
UFZ approach: IT tools (I)

Managing all kinds of data ⇨ UFZ **D**ata **M**anagement **P**ortal (DMP)

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

UFZ approach: IT tools (II)

Exploring all kinds of UFZ data (internal and external access) ⇨ UFZ **D**ata **R**esearch **P**ortal (DRP)



- Keyword- and Map-based search
- Data access with configured priviliges
- Details page = DOI Landing page

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

Different data – different requirements (I)

The case of UFZ logger data

Mostly sensors, e.g. weather stations (“Big“ Data) => TERENO

Responsibility taken by IT Department

- Database, interfaces
- Workflow, ...

- Level 0 = Raw data
- Level 1 = Transformation to target unit (e.g. V in °C)
- Level 2 = Quality controlled (**check + correction ⇨ missing data, outliers, ...**)

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

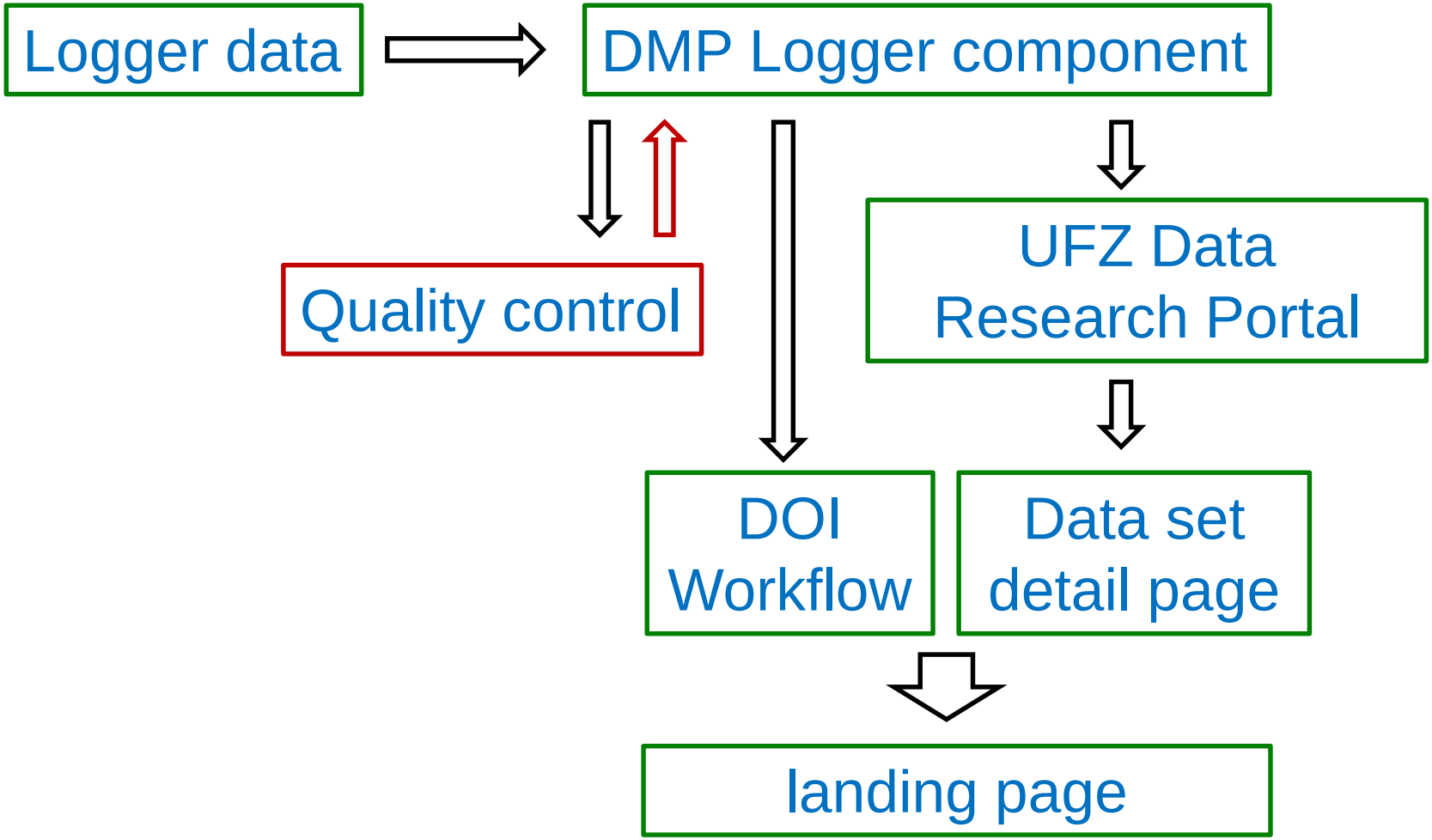
UFZ quality control of logger data (⇨ DOI)

Crucial for use and publication of data!

- Evaluation of **software** products
  - Searching the magic tool
    - Generic, allowing high degree of adaptation to user needs, user-defined algorythms
  - UFZ-decision for training on
    - **ORIGIN** (data analysis and graphics software)
    - other software solutions, i.e. **Matlab**, **R** and others on demand

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

Logger data to landing page at the UFZ



HELMHOLTZ
ZENTRUM FÜR
UMWELTFORSCHUNG
UFZ

Different data – different requirements (II)

Example biodiversity data (mostly person-generated data)

Issues

- Heterogeneity of data
- Logic of ecologists related to data (different from IT people)
  - Ecology: data based on spreadsheets ⇨ Data base?!

- Plausibility tests
  - Expert knowledge
  - Software (e.g. occurrence of species A at location B plausible?)
- Technical consistency
  - Correct data types
  - Correct cell entries

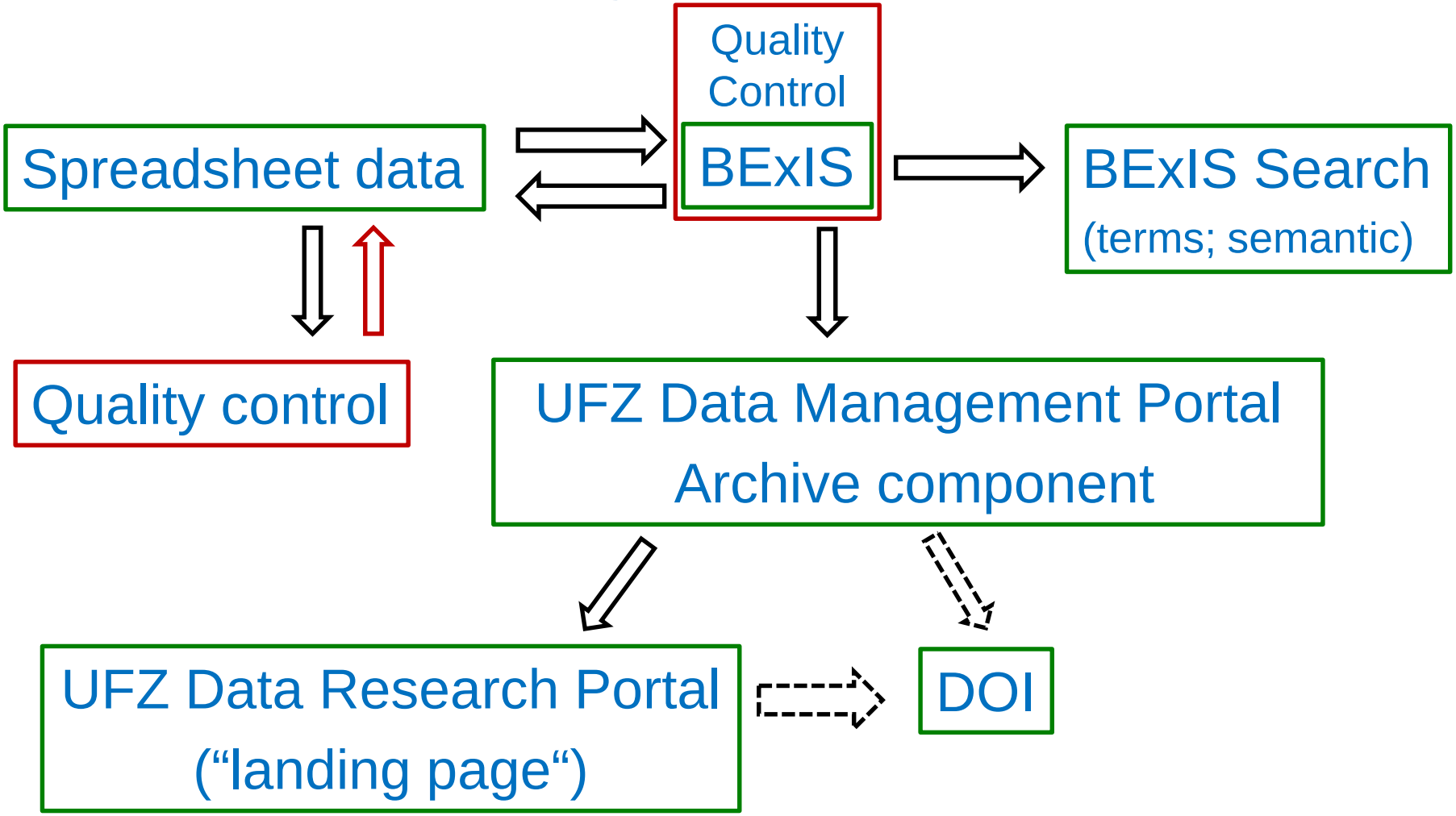
UFZ quality control of biodiversity data: BExIS

BExIS: **B**iodiversity **Ex**ploratories **I**nformation **S**ystem

- **DFG**-Project: generic open source information system for biodiversity data (funded until 2017; http://fusion.cs.uni-jena.de/bexis)
- Tool for **spreadsheet**-based biodiversity data (Excel import)
- **Features**: Import // Table-to-database // Export // Consistency check // Retrieval // Metadata // Admission rights
- Test case for UFZ DOI-ready data sets

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

Workflow biodiversity data UFZ



HELMHOLTZ
ZENTRUM FÜR
UMWELTFORSCHUNG
UFZ

Conclusions (UFZ)

- **Tools for quality control** of data are at hand
- **Technical (IT) environment** is DOI-ready
- **Promotion** within UFZ (scientists) needs to start
- DOI technical issues to be solved
  - DOI-relevant data
  - Granularity
  - Versioning, ...
- **Editorial team**?
- "**DOI-Agent**" issue still open (Helmholtz solution?)

⇨ Finally we should aim at **Open data!**

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ