A Data Quality Management Module for BExIS

Michael Owonibi, Roman Gerlach, Andreas Ostrowski, Eleonora Petzold
Friedrich-Schiller-Universität Jena

Birgitta König-Ries
Friedrich-Schiller-Universität Jena and
German Centre for Integrative Biodiversity Research (iDiv)
Halle-Jena-Leipzig
Background and Motivation
• DFG priority programme
• Aims at studying the
  –relationship between biodiversity of different taxa and levels
  –role of land-use for biodiversity of different taxa and levels
  –role of biodiversity for ecosystem processes
• Scientific infrastructure established in 2006
• Integration of contributing projects
  –26 projects commenced in early 2008
  –8 projects commenced in early 2009
  –~20 projects commenced in 2011
• Turnover of ~40% in 2011
• Currently 80 PIs & 300 members
German Centre for Integrative Biodiversity Research iDiv

- DFG Research Centre
- Founded in 2012
- Consortium of 3 universities and 8 other institutions
- 8 new professorships
- around 200 scientists
- Strong focus on synthesis

- Idiv-biodiversity.de
Projects’ background

– sharing of data among projects crucial for the success of the project
– importance of integrative data analyses and data re-use
  → common platform with fine grained access rights needed
– Wide variety of disciplines
– Little/no prior knowledge of what projects and thus what data to expect
  → very heterogeneous data
  → no fixed schema
The System

Resources
- Field book
- Mowing scheme
- Black boards
- Publication list
- Event Registration

Maps
- plot maps
  - interactive & editable
  - downloadable
- data maps
  - data retrieval
  - Google Maps visualised

Data
- diverse file formats & syntax
- upload & update via file
- direct online edit
- query, filter, merge online
- download diverse formats
- online R analysis
- version control
The System

- Central web portal with fine grained access rights
- Online since 09/2007
- Management of structured data
  - 1:n relational data sets, matrices
  - Normal data tables
- Management of unstructured data
  - Binary formats such as pdf or tiff, models
- XML-based storage technology
  - Metadata, primary data, configuration
• … mostly manual

• Raise awareness for data quality through teaching

• Provide structure for required metadata

• Manual checks for metadata completeness and correctness
• Random checks for data quality

• Automated checks for completeness of sensor data
(most common) Data Quality Problems

• Duplicates
• Start/End date in metadata ≠ start/end date in primary data
• wrong data type
• missing decimal separator
• wrong values due to typos
• wrong/missing sensor values
• wrong values due to errors (e.g., wrong species identification)
• separation of data into several columns e.g. date is written in 3 columns: year, month, day or species name divided into genus, species
• usage of acronyms and missing/incomplete definition of them
• ....
New BExIS

- Synthesis Work
- Semantic Search
- Analytics
- Spatial Querying
- Data Quality Management

UI
- UI Framework
- Integration
- Core Functions
- Modularity

Security

Data Access
- DB2
- PgS
- ...

External Tools
- Web Services
- Archiving
- Publishing
- Import/Export
What would we like to have?
Planned Data Quality Checks

• Numeric Fields
  – compute statistics and
    • find outliers
    • compare to similar datasets
  – compare value against ranges described in metadata
  – check for patterns and deviations
Planned Data Quality Checks

(2)

• Short String Fields
  – check for patterns and deviations
  – compare with species lists etc.
  – remove/replace special characters
  – check for typos
  – compare against domain values
  – check for terms not contained in vocabulary lists
Planned Data Quality Checks

(3)

- Date Fields
  - check for correct format
  - check for outliers
  - compare against domain range

- Project Based Checks
  - check for correct PlotIDs
  - check for adherence to data policy

- GIS based Checks
  - compare coordinates with info in meta data
Planned Data Quality Checks

(4)

– Dataset wide checks
  • check for almost blank columns/rows
  • check for blanks, 0, NA-valued and null-valued cells
  • compare columns against summary data in metadata
  • check for duplicated rows

– Data structure checks
  • check for incorrect typing
  • check for nonatomic values
  • check for mixture of data types within a column
Intended Data Quality Workflow
iDiv

Planned Quality Assurance Workflow

• At data structure design time
  – Specify DQM criteria on (groups of) variables, e.g.,
    • Integrity constraints, data types, patterns, dependencies,
      domain value ranges, …
  – Specify related datasets
  – System creates (downloadable) Excel-templates with
    (embedded) validations based on data-structure

• At creation time
  – Excel Sheet will help discover errors immediately

• At upload time
  – validation against the easier to check DQM criteria
• After upload, batch processing of
  – data profiling
  – data analytics and mining
  – send report to user
  – provide tools for (mass-)correction
Thanks …

• .... to DFG for Funding (SPP 1374, BExIS++, iDiv)

• .... the BExIS fathers Ernst-Detlef Schulze and Jens Nieschulze for starting the project

• ... the current and past BExIS, BExIS++, iDiv, AquaDiva teams

• .. and to you for your attention!
A Data Quality Management Module For BExIS

Michael Owonibi, Roman Gerlach, Andreas Ostrowski, Eleonora Petzold
Friedrich-Schiller-Universität Jena
Birgitta König-Ries
Friedrich-Schiller-Universität Jena
and
German Centre for Integrative Biodiversity Research (iDiv)
Halle-Jena-Leipzig