TDWG 2011 Annual Conference
Comparison of performance in different data models for taxonomic databases
Overview

• The Edaphobase project
• Database structures to be compared
• Methods of comparison
• Results
EDAPHOBASE

- Connecting information from
  - collections
  - publications
  - projects (observation data)
- Data about distribution and ecology, not specifically taxonomy
- Soil organism groups: Collembola, Oribatida, Gamasina, Chilopoda, Diplopoda, Lumbricidae, Enchytraeidae, Nematoda
- Joint partners: SMN Görlitz, SMN Karlsruhe, ECT GmbH Flörsheim
- GBIF information system soil zoology
Database structures being compared: Generic model

Additional:
- couple of Views
- couple of Rules/ Functions
Database structures being compared: Relational model

Institutiontype

Institution

Taxon

TaxonRank

Author

Document

Sourcetype

Additional:
• no Views
• couple of Rules/ Functions
Database structures being compared: Hybrid model
Methods of comparison I

• Building up the test environment
• Implementation of generators
  – Fixed string generator – always the same string
  – SELECT generator – value by chance from a set of values (even by database request)
  – Append generator – ordered list of other generators
  – Pattern generator – value by chance according to a pattern
Methods of comparison II

- Configuration of generators by XML

- Framework of tests
  - Insert of data
  - Saving runtime data
  - Repeated insert for more data ... Saving ...
  - Time measurement
  - Visualisation of runtime data by JFreeChart

- Dropping the test environment
Test I

- Runtime Environment:
  - OS: Ubuntu Lucid 64 Bit (Kernel 2.6.32)
  - CPU: Intel Core i7 Q840
  - RAM: 4GB DDR3 Dual Chanel
  - HDD: 500GB 7200 rpm
  - Database: PostgreSQL v9.0 with standard configuration
Test II

- Test Environment
  - insert cycles: 1000
  - datasets per cycle: 135
    - institution: 15; document: 30; author: 40; taxon: 50
  - meantime of 10 executions per query and cycle
  - queries with random id-selection
Comparison of execution time at 100,000 datasets for the different schemas

duration in ms

- generic schema (height abstraction)
- generic schema (low level)
- relational schema (selective list)
- relational schema

- scientific taxon name
- taxa of parent publication authors
- publications of parent taxon author
Acknowlegdements

• to the project leaders: W. Xylander, D. Russell
• to the partners: ECT GmbH Flörsheim (J. Römbke), SMN Karlsruhe (H. Höfer)
• to the programmers: A. Franzke, S. Lesch
Further Acknowledgements

- to the funding: BMBF (support code 01LI0901A)
- to the program committee
- to the audience for attention