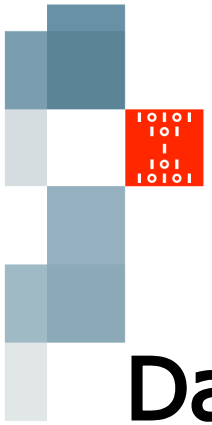


netarchive.dk

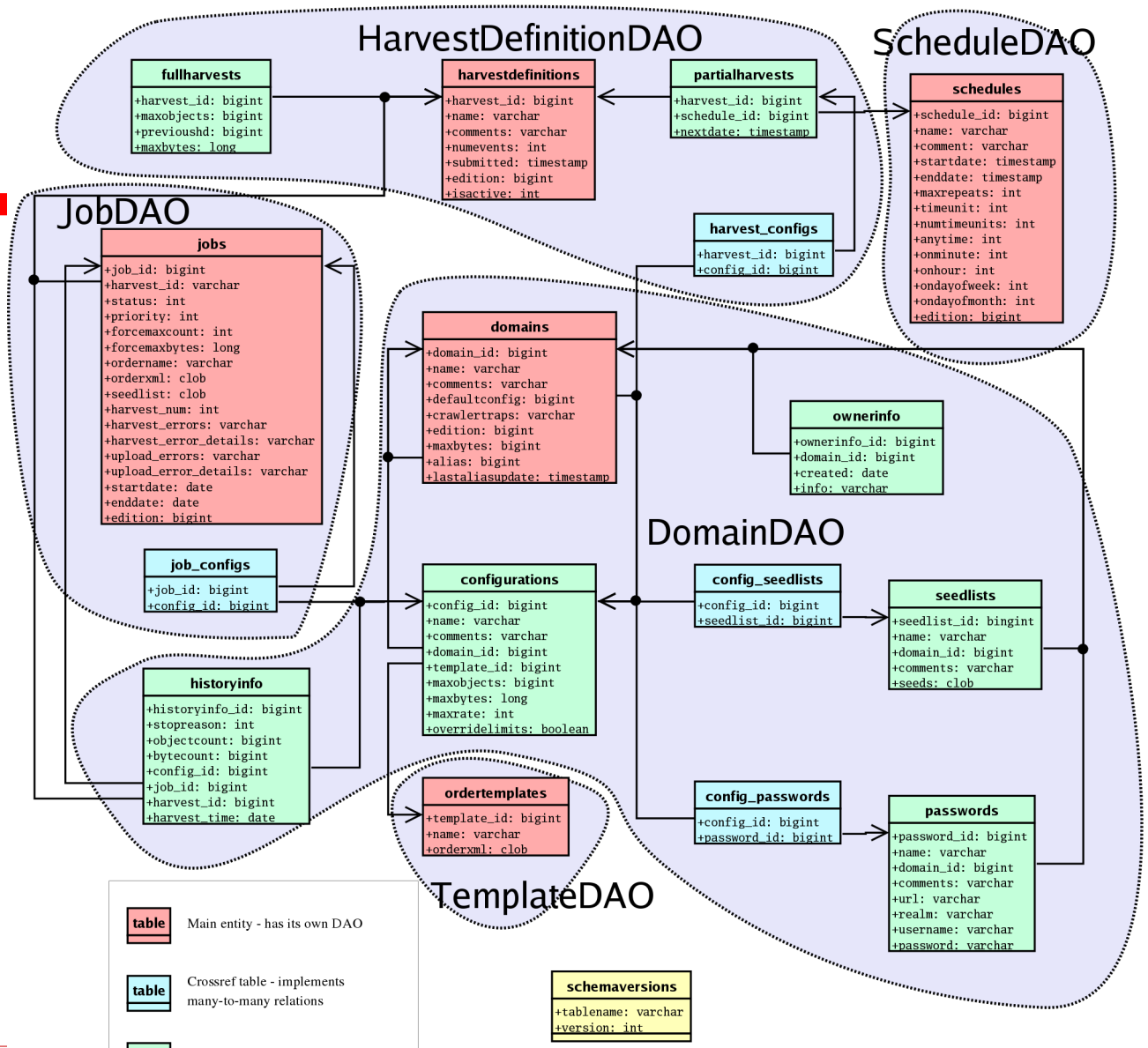
Database

Lars Clausen
Netarkivet

netarchive.

DAOs

Physical layout of harvest definition database



Scheduling

- Snapshot harvest hard on scheduler
- Tweaks have improved performance
- Complex algorithm partly to blame
  - Tries to make jobs with similar-size domains
  - Guesses based on prior harvests
  - Allows for sudden jumps
- Likely to become a plug-in

Derby vs. MySQL vs. ...

- We currently have plugins for MySQL and Derby (embedded and server)
- No tests yet of relative speed
- SQL tweaking has solved major bottlenecks
- Code should work with most SQL DBs
- Indexes may work differently in other DBs

DBSpecifics plug-in

- `static DBSpecifics getInstance();`
- `String getDriverClassName();`
- `void dropJobConfigsTmpTable(Connection c, String tableName);`
- `String getJobConfigsTmpTable(Connection c);`
- `void backupDatabase(File backupDir);`
- `void shutdownDatabase();`
- `void updateTable(String tableName, int toVersion);`

Creation scripts

- Must be a way to create the DB
- Typically an SQL script
- Initial DB must include one template
  - Setting: defaultOrderXml

Updating existing DBs

- Not currently in the code
- Plug-in has stub for update
- Updates on a table-by-table basis
- Version of each table kept in DB

Scheduling

- Scheduler is pressed at start of snapshot
- Tweaks have improved performance
- Complex algorithm partly to blame
  - Tries to make jobs with similar-size domains
  - Guesses based on prior harvests
  - Allows for sudden jumps
- Likely to become a plug-in