Minutes DK-visit and DINA TC meeting 9th of June 2015 Stockholm

SE: Markus Skyttner, Ingimar Erlingsson, Markus Englund, Kevin Holston, Lisa Sundström and Ida Li

DK: Nikolaj Scharff, Martin Stein, Thomas Stjernegaard Jeppsen, David Konrad

Python scripts for import/export, csv.files
Target group: advanced database administrators, "data wranglers" working with migrations from other legacy database systems to Specify database. This script does not include import/export of the trees Taxonomy, Geography and Storage, they need to be loaded separately. Future use: The scripts could also be used to load data into the DINA virtual appliances.

Importing data through the Specify Workbench has limitations when dealing with datasets larger than 1 million records. With these python scripts the export/import only takes app 20 minutes. Automatically command line to import the data, which is very close to a turn-key system. Another issue with using the Workbench is that is not importing loans/accessions. This method uses an intermediate standardized format, DINA-archive format for importing and exporting, which covers database constraints with respect to datatypes and field content constraints while the Python scripts address constraints on table relationships and key values

The Python tool loads data directly into the database from the csv files mirroring table structure and relationships. Row content must be validated beforehand with respect to local database configuration. Tree-related records, such as taxonomy, are uploaded in an initial stage via the Workbench, and file processing includes cross-referencing csv data with these Specify records. Use requires some command-line interaction with these loading scripts and preferences although packaging of these Python scripts enables one to conduct most of the processing through a user-friendly interface (i.e., as a browser webpage).

Introduction "DINA Data Tool"
Data cleaning tool for curators, database managers are using the Specify workbench to load data.

1. Data entry from physical collection objects, hand written lists etc.
2. Data import from csv files.

DK approach is to get sensible csv.files out of very poor quality datasets with long strings of text. They look incredible simple, but there is a lot of complexity hidden in the long strings. The effort is to normalize these datasets and generate the complexity into Specify to improve the quality of the datasets. The DK-team is also migrating several legacy databases such as Filemaker, Paradox etc.

Workbench pros and cons: two step handling, a sand-box in front of the Specify for the last step for validating data. Depends on Specify 6, tedious to enter tree based data and also very limited validation.

The DINA Data Tool is built on java script, both back-end and front-end. The tool can also upload csv files directly to a Specify database, populating the tree-related tables with records (e.g., taxonomic and geographic names or storage units). The resulting records, and the field mapping set up by the user, are stored external to the Specify. After records are uploaded directly to the database, they can be visualized immediately via the web client. Visualization of trees in Specify requires recalculation of nodes (i.e., "update tree" in the Specify Java client).

The DINA data tool provides a functionality that enables more than one person to work on one single workbench dataset.

Discussion on forms in Specify:
Markus S: heavily usage of forms with a high level of freedom. We have talked about that it would be nice to get away from the complexity of the xml-files. Thomas: it would be nice to have a more standardized, a more simple form for basic information. Markus: is there is path to get rid of the xml.files? They are currently linked between Specify 6 and Specify 7, but you can just build your own interface on top of the web API if you want to.

Martin vision: a set of required fields (set by administrator) free to customize your interface. Markus: used to be a very good idea, but there could actually be better with less flexibility. Martin: start with a very simple interface.

DINA Data Tool demo for SE-team and a small selected end user group: Sabine Stöhr, Monika Myrdal, Johannes Lundgren, Ove Johansson, Jörgen Langhof, Tobias Malm.

Thomas: We will add functionality for search and replace. A list of 50 000 and 3 000 is misspelt, how can you do a search and change. This is a long term goal, but does not exist right now. The vision is to create a tool that can help curators to clean data. Preview function: Mark all the names that are not matched in the databases. In the future this will also be a tool that can also match external resources.