M5.11 - Review existing key construction software and workflow interaction with both Scratchpads and the CDM

Date: Monday, February 28, 2011
Milestone
Work package: WP5: Data
Reporter: Régine Vignes-Lebbe

Introduction

Identification of specimens and observations is a crucial task . Today methods are based on molecular sequences (barcoding), image or sonogram analysis, morphometry, or morphological data. For this last type of data, which is also the most often available information, we use identification keys. Since the first printed key published by Lamark (Flore Francoise 1778) taxonomists have developed various forms of « identification keys » : single-access, free-access and multi-entry keys (Hagedorn, 2010).
In this document we make a short review of existing key construction software for these three types of keys, and then we describe the architecture and workflow to offer such webservices linked to Scratchpads and the CDM.

Types of Key generator software

1- INPUT
All key generator software need an input with a knowledge representation of the descriptive data. These descriptive data can be attached to already identified specimens, or to named taxa. The majority of existing software offer keys from taxonomic descriptions, but few of them work on a set (sample) of specimen's descriptions (IKBS).
Case based reasoning systems, using specimen's descriptions, compute a similarity measure between the attributes of the unknown specimen and the description of each identified specimen ; then the unknown specimen is assigned according to the K nearest descriptions.

2- IDENTIFICATION METHODS

a) Expert systems

Some software for assigning an object to a class were named expert systems. They apply a set of rules (the reasoning is forward chaining or backward chaining). In taxonomy, the knowledge is not expressed as rules, but as descriptions. So the elaboration of such set of rules would need to compare all the pairs of taxonomic descriptions to deduce rules such as if a specimen is compatible with the premises, then it belongs to the taxon (or one of the taxa) of the conclusion of the rule.
We don't work on all the methods for identification, but we focus on key generator software.

b) Single access keys

From the descriptive data, some software compute a tree structure giving one single strategy to identify any unknown specimen: these single access keys are dichotomous or polytomous. In this case the key generator software has the task to compute the key, and another software has in charge the use of the key.

Key generator software implement recursive algorithms which construct one by one the different nodes (or leads) of the key. For each node the descriptor is chosen thanks to a scoring measure. The algorithm stops whenever the best feature do not discriminate taxa anymore or when there is only a single taxon left.

Due to the polymorphism of the taxa, a node does not necessarily creates a partition of the remaining taxa. Many different scores can be computed to optimize different criteria on the final topology of the key (or graph structure): minimising average path in the key , number of terminal nodes , total number of specimens …). The problem is that there is not one "key" in the absolute ; the criteria to optimize depend on the domain, and inside a domain, on the properties of the taxa to be distinguished.

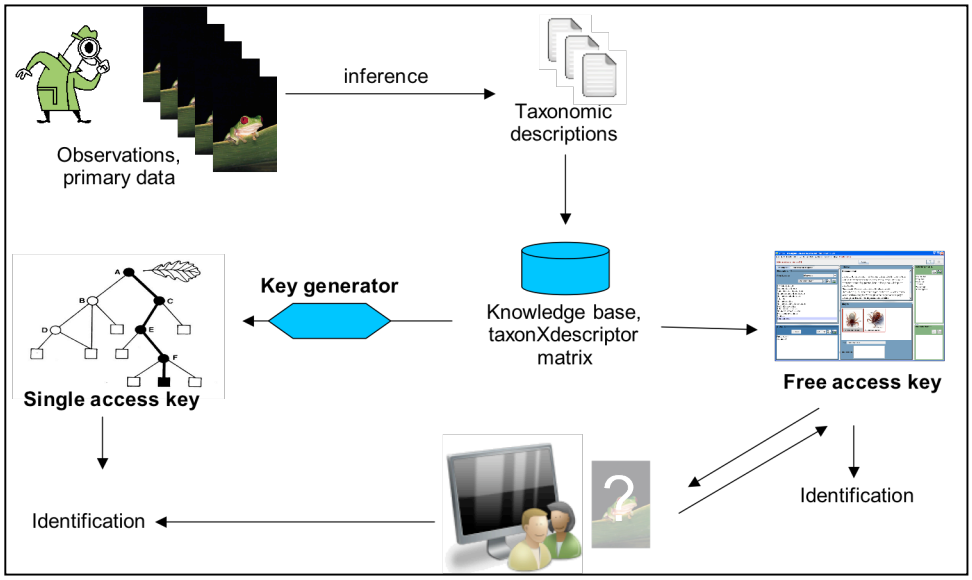
For more details on the criteria to built a key, see Gerard et al, 2010.

c) Free access keys (monothetic or multi access keys)

At the contrary, free-access keys run directly on the descriptive data without to compute another knowledge representation form. At each identification step, the user chooses one (monothetic access key) or several (polythetic or multi-entry key) descriptors and selects the attributes according to his specimen. This type of key is more flexible than single access keys ; they allow partial identification, they are adapted to various contexts without additional work (season, region, development stage, incomplete specimen, different taxonomic skill of the users etc.), and some software manage uncertainty.

For an overview and comparison of existing software see:
- EDIT website http://www.bdtracker.net/
- Delta website edited by M. Dallwitz (http://delta-intkey.com/)

The following schema describes the human and computerized steps from identified specimens to the identification (it means assignment in an already delimited taxon) of unknown specimens:



Key construction software and workflow interaction with both Scratchpads and the CDM

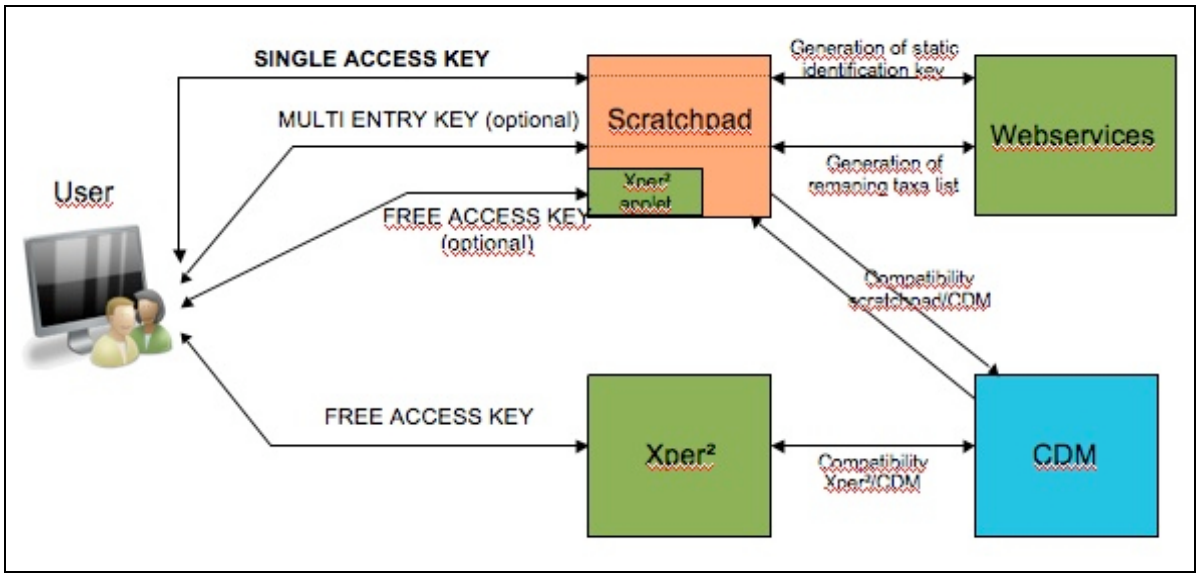
1-THE SINGLE ACCESS KEY GENERATOR IMPLEMENTED IN THE CDM LIBRARY
Different functions were previously coded in the CDM library in order to generate single access keys from descriptive data in the CDM store (EDIT).
The recursive algorithm looks for the most discriminative feature, thanks to a scoring measure, and creates step by step a key each node representing the set of taxa discriminated at one place in the key, and each branch representing the state(s) of the best descriptor at this point, leading to a new set of taxa. Thus the algorithm stops whenever the best feature do not discriminate taxa anymore or when there is only a single taxon left. The final output is stored in a CDM object PolytomousKey.
This library runs only on the CDM classes of EDIT plateform. During the first step of ViBRANT it will allow to test and to improve various options of the key generator before to code it as a web service accessible from the Scratchpads. We can list the principal functions of the key generator:

- Calculating the score, for qualitative and numerical data, it means the capability of a descriptor to distinguish the taxa.
- Creating branches of a node, including merging options
- Pruning the key
- Formatting the output

Other options could be implemented in a second step to take into account modifiers (to deal with the typicality of values in taxonomic descriptions), descriptor weight and item weight, multi descriptors and reticulation in the key.

2- THE FREE ACCESS KEY SOFTWARE Xper^2
Xper2 is one of the available software for free access keys. Its import-export in SDD and many other formats will allow to offer free access keys to the scratchpads.
For more information on Xper2 software see http://lis-upmc.snv.jussieu.fr/lis/

3- WORKFLOW INTERACTION
The following schema describes the workflow for the different types of identification keys.



Scratchpads and the CDM-EDIT plateform offer two different interfaces for collaborative work. To edit descriptive data, the Scratchpads include a matrix editor which will be improved during the project, and the EDIT platform integrates a tool dedicated to descriptive data (Xper2). In the two cases, the descriptive data are stored in the corresponding store and can be exported in the international exchange format TDWG-SDD (Structured Descriptive Data).

The key generator webservice loads the SDD file and the key parameters, then it generates the key; the scratchpads receives the single-access key file and can display the key in different formats.