Onomastics

Version 1.2, updated 2002-12-15
Author Andrej Arn, Sam Blum

What does onomastics mean?
Onomastics is the science of names, naming conventions, and the origin and history of names.

See also: http://dmoz.org/Science/Social_Sciences/Language_and_Linguistics/Onomastics/

And what can it do for us?
Andrej was working for a company in 2002, that had to 'clean out' there client DB that contained many 1000's of addresses. It turned out that 15-20% where not fully correct. E.g. Use of wrong gender (Mr./ Mrs.), first and last name switched, duplicate client entries (Tom Edison / Thomas Edison).

Whenever you have a DB with personal information, you want a good quality of your data. It can be a phone book, a member list, client list, whatever.

Think of what happens if your company sends the same letter to a client twice, one to "Bill Clinton" and one to "William Clinton". Or when Julia receives a letter to "**Mr.** Julia Roberts".

Besides the extra costs, the client will have a bad impression of your organization.

Things like that happen way too often when entering a form:

- Someone selected the wrong gender radio button.
- Someone adds "Rick Miller" as new client because the existing record is listed as "Richard Miller".

Using the ***BlueShoes*** onomastic plugin can reduce these kind of errors before they happen.

Problem areas:

- typo: simth instead of smith.
- foreign characters: mueller or müller, juerg or jürg.
  note that some systems use juerg instead of jürg cause of character sets, and some ppl need to be written that way.
  If your name is written "juerg müller" in your passport, you can't go and open a bank account as "jürg müller".
- nicknames: peggy for margaret, bill for william.
- short forms: b. gates for william gates (ouch!).
- localization:
  andreas from germany, andré from france, andrei/andrej from russia and andrea from italy is the same name, just translations. moved to the us, and all become andrew

because those americans can't make a difference.
mr. hundertwasser may also become hundredwater or so. with italians 'andrea' we have another problem, cause in germany andrea exists too, but is feminin.

- name families: adelheide, adelheidi.
  especially nowadays it's cool to invent new names, or at least new spellings. the americans might be different, having michael as #1 baby name since 1964 *shakes head*.
  in this example, heidi may be a different family member and a nickname, both is possible.
- titles and qualifiers: dr. prof. ph.d. nat. ...
  may be written as "dr. smith" or "smith, dr."
- prefixes: Di Caprio, D'Agostino, Al Afif, Mc Donald, ...
- married: "peggy smith" becomes "margaret johnson". bingo.
  she may aswell become "peggy johnson-smith" or so, which leaves us a chance to still find her.
- write form: john peter, johnpeter, john-peter
- localization again:
  once french was the world language #1, it was used to translate names from russian etc into 'european'. a real example is a person that was named Lejnine and now became Lezhnin (with passwort change and everything).
- sex operation: mister thomas smith becomes miss emelie smith.
- with spanish names and fulltext indexing we have to be more strict because everyone is called jose maria fernandez garcia gonzalez rodriguez. You'll find a match of those in nearly every name.

The package offers functionality to find duplicates in your db, calculate the similarity of 2 name pairs, find out the gender of a first name, check if a name pair (“peter johnson”) is in the wrong order (“johnson peter”) etc.

It uses databases with thousands of names, plus additional information (name prefixes, titles etc) to do so.

Example usage of the onomastics package is demonstrated here:
http://www.blueshoes.org/en/examples/onomastics/

To manage the data we use a gui with vml, here’s a screenshot: