
SketchEngine + DPS + Integrations

- Background & Objectives
- How?
- Demo
- Summary
- Credits

Meeting March 2011

For more information contact: Holger Hvelplund
e-mail: hvelplund@idm.fr

Background

- **Increase SEO by increasing cross references**

Number of cross references in existing dictionaries are low. Cross references are important for SEO and for number of pages users visit in a session.

- **Increase footprint**

Footprint in free online dictionary products from free players is high. Wiktionary.org contains 2.2 mio entries. Wikipedia.org contains 3.5 mio entries. According to recent blog Wordnik contains 9 mio. entries. Answers.com, thefreedictionary.com, dictionary.com are also rapidly increasing the footprint of their services.

- **Quality of the service**

Traffic analysis of existing web sites gives indicators for improvement to the service. For example traffic to the “spell-checker” page

How?

- Use Sketch Engine + DPS + Integrations to:
 - Increase number of cross references in dictionaries by adding auto-generated wordsketch and thesaurus section to each dictionary entry
 - Increase footprint of dictionaries by adding auto-generated entries consisting of:
 - “Good” example sentence section
 - Wordsketch section with collocates and “good” example sentences associated to each collocate
 - Thesaurus section (where users in example sentences can see how the two words are similar and how they are different)

Increase footprint: Identify new words

- Analyzing “spellchecker” traffic
- SketchEngine
 - News wire corpus (currently used for identification of neologisms)
 - Building subject specific corpora – for example using content from journals and subject oriented books
 - Using WebBootCaT for building additional subject oriented corpora
- Wiktionary.org (English)
- Wikipedia.org (English)
- WordNet
- Google Insight (www.google.com/insights/search/#)
- Other free online resources

How?

Layered Information Architecture

- Layer #1: Repository of publications
- Layer #2: Sketch Engine: Subject oriented corpora
- Layer #3: DPS Sketch Entry database
- Layer #4: DPS Dictionary database

How?

Demo

- DPS project #1: WordNet dictionary
- DPS project #2: Automatically generated sketch entries
- Integration process for:
 - using Sketch Engine API for retrieving and converting, storing wordsketch and thesaurus reports into dictionary entries in DPS project;
 - merging WordNet dictionary and automatically generated sketch entries into online dictionary.

Demo

- Dictionary entries
 - <http://skedps.cw.idm.fr/dictionary/british/234>
 - <http://skedps.cw.idm.fr/dictionary/british/330>
- SEO optimized dictionary entries
 - <http://skedps.cw.idm.fr/dictionary/british/11>
 - <http://skedps.cw.idm.fr/dictionary/british/22>
 - <http://skedps.cw.idm.fr/dictionary/british/42>
 - <http://skedps.cw.idm.fr/dictionary/british/54>
 - <http://skedps.cw.idm.fr/dictionary/british/167>
- New auto-generated entries
 - <http://skedps.cw.idm.fr/dictionary/british/43470>
 - <http://skedps.cw.idm.fr/dictionary/british/43491>
 - <http://skedps.cw.idm.fr/dictionary/british/43552>
 - <http://skedps.cw.idm.fr/dictionary/british/43556>
 - <http://skedps.cw.idm.fr/dictionary/british/43568>

Comments to the demo

- Demo is using incoherent and misaligned content:
 - WordNet dictionary
 - Wiktionary.com for identifying new words
 - General corpus for finding relevant information for new words
- New words identified in Wiktionary are typically rare and old fashioned/use words

Summary

- **Objective #1:**

When dictionaries do not contain entries for words that users look up dictionaries should present in an easily accessible form the best information lexicographers have at hand for producing a good dictionary entry.

- **Objective #2:**

Any word that appear in a book that is published should be available in the dictionary

- **Objective #3:**

Auto-generated glossary feature: offer to other publishers/content providers that their content can be supplemented with a glossary containing “definitions” of all words in the content.

Ideas for what could come next?

- Action 1: How to filter away "noise" in auto-generated content (SEO, and new entries).
- Action 2: How to make clear distinction between auto-generated and editorially checked content (including cross references).
- Action 3: How to improve performance of the interaction between DPS and SkE.
- Action 4: Improve quality of auto-generated content (SEO and new entries) by building subject specific corpora with publisher's own content - for example with contents of journals - and use the subject specific corpora for:
 - identifying collocates and new words; and
 - collecting information (i.e. example sentences) associated to collocates and new words.
- Action 5: Auto-detect dictionary entries that mostly need editing by automatically computing "alignment factor" of dictionary entries to so-called auto-generated "corpus entries".
- Action 6: Create collaborative environment where efficiency of dictionary production can be improved by triangle consisting of: (a) end-users (directly or indirectly); (b) lexicographers (directly or indirectly); and (c) statistically tools and technology like SkE - DPS generated content.

Summary

Credits

- Credits
 - Karim El-gargati, Mikaël Lebrun, Allan Ørsnes
 - Adam Kilgarriff, Vincent Lannoy, Philippe Climent
- More information:
 - See IDM's web site: <http://www.idm.fr>
- Questions:
 - E-mail to hvelplund@idm.fr