Creating dictionaries for Apache OpenOffice and maintaining them through web services

Andrea Pescetti
pescetti@apache.org
Andrea Pescetti

- VP, Apache OpenOffice
- Unaffiliated volunteer
- Dictionary packager
- Day job: web developer
Getting Started
OpenOffice Language Support

$ svn ls https://svn.apache.org/repos/asf/openoffice/trunk/extras/l10n/source/ | grep -c /
112
Writing Aids: An Overview

- Spell checker
- Thesaurus
- Hyphenation Patterns
- Grammar Checker
Spell Checker

- Engine: **Hunspell**, integrated in OpenOffice.
- Hunspell dictionaries available for 100+ languages.
- [http://hunspell.sf.net](http://hunspell.sf.net)
Thesaurus

- Engine: integrated.
- OpenOffice-specific format.
- Must start from scratch.
- linguocomponent.openoffice.org
Hyphenation Patterns

- Engine: Hyphen, from Hunspell.
- Integrated in OpenOffice.
- Format: tool-specific.
- But you can convert TeX patterns: http://ctan.org/
Grammar Checker

- Available only as API.
- Options as extensions: LanguageTool, LightProof, CoGr00 and more.
- Format: tool-dependent.
Licensing Issues
Mere Aggregation

- Crazy variety of licenses.
- Many incompatible with AL2.
- But bundling is allowed: "mere aggregation", LEGAL-117
Extensions (OXT)

- Writing Aids are now extensions (XML+data+ZIP)
- Hosted anywhere, bundled at build time.
- Reinforces “mere aggregation”.
Choose your license

- **AL2**: Apache License, free and permissive, GPLv3 compatible.
- **LGPLv3/GPLv3**: can be used through mere aggregation.
- **AGPLv3**: untested so far, but likely mere aggregation too.
(Don't) Meet Apache Legal

- Extensions are externally hosted
- extensions.openoffice.org
- considered external too.
- No paperwork needed!
Distributed Management
Use a repository

- Make sources available in an online repository.
- Use version control.
- Expose a web-based change tracking interface.
Spell Checker

- One file in text format.
- Human readable, except rules.
- Good for collaborative editing.
Spell Checker: example

abbagliando/D
abbagliante/STUq
abbagliare/ALKhlTXI
abbagliata/QTU
abbagliato/EyT
abbaglio/OTq
abbaiaimento/OTq
abbaiaando/D
Thesaurus

• One file in text format.
• A generated index.
• Human readable.
• Good for collaborative editing.
Thesaurus: example

abiezione|1
(s.f.)|degradazione|vergogna|viltà
abile|3
(agg. Esperto, capace)|adatto|atto|idoneo
(agg.)|consumato|esperto|pratico
(agg.)|competente|efficiente|valido|virtuoso
Hyphenation

- One text file.
- Format: less readable than Perl!
- Changes very rarely.
- Fix bugs upstream, in TeX.
Hyphenation: example

g2n
2g1p
g2r
2g1s2
2g1t
Grammar checker

- LanguageTool: rules in XML.
- Fix upstream, in LanguageTool.
- Collaboration possible.
Grammar checker: example

<!-- Punto esclamativo ripetuto (!) -->
<rule>
  <pattern>
    <token>!</token>
    <token>!</token>
  </pattern>
  <message>Se ne mette uno solo: <suggestion>!</suggestion>.</message>
  <example type="incorrect">Fammi parlare<marker>!!</marker></example>
  <example type="correct">Fammi parlare<marker>!</marker></example>
</rule>
Packaging

- Generation of the OXT extension is scriptable.
- Post-commit hook possible.
- Keep generated OXT files in the same repository.
Collaboration possible in every component.

A script to package the extension.

A release manager to make stable versions available.
Going 2.0

- Native-lang community: best people to improve N-L tools.
- Motivated users, interested in improving OpenOffice.
- Issue: providing efficient infrastructure.
Web-based interface

• An idea from O0oCon 2010.
• Report missing or erroneous words from within OpenOffice.
• Easy to setup as web service.
• Notifications: e-mail to maintainers, suggestions in DB.
Web based interface: example

Term: *
altresì

Action to take: *
- Add to dictionary
- Remove from dictionary
- Fix wrong hyphenation

SUBMIT
Expose web services

- Direct usage of the web application via browser.
- Access available through web services too.
- Suitable for applications or macros.
Web services in OXT

• Embed a macro in the OXT dictionary package.
• Right-click on a word:
  • Nominate for inclusion in dictionary
  • Nominate for removal from dictionary
  • Report wrong hyphenation
Thesaurus maintenance

- Vithesaurus: free online tool for collaboratively creating and maintaining a thesaurus.
- In use (German) at http://www.openthesaurus.de
- https://github.com/danielnaber
Handling Duplicates

- Millions of users can lead to duplicate reports.
- But it's a plus: use frequency for ranking.
Handling Wrong Reports

- Annoying: users make some wrong suggestions and repeat them!
- The web application supports “motivated blacklisting”: repeated wrong submissions are handled and a message can be shown to the user.
Thanks for attention

Andrea Pescetti
pescetti@apache.org

www.openoffice.org

Image credits: Flickr, PLIO Archives.