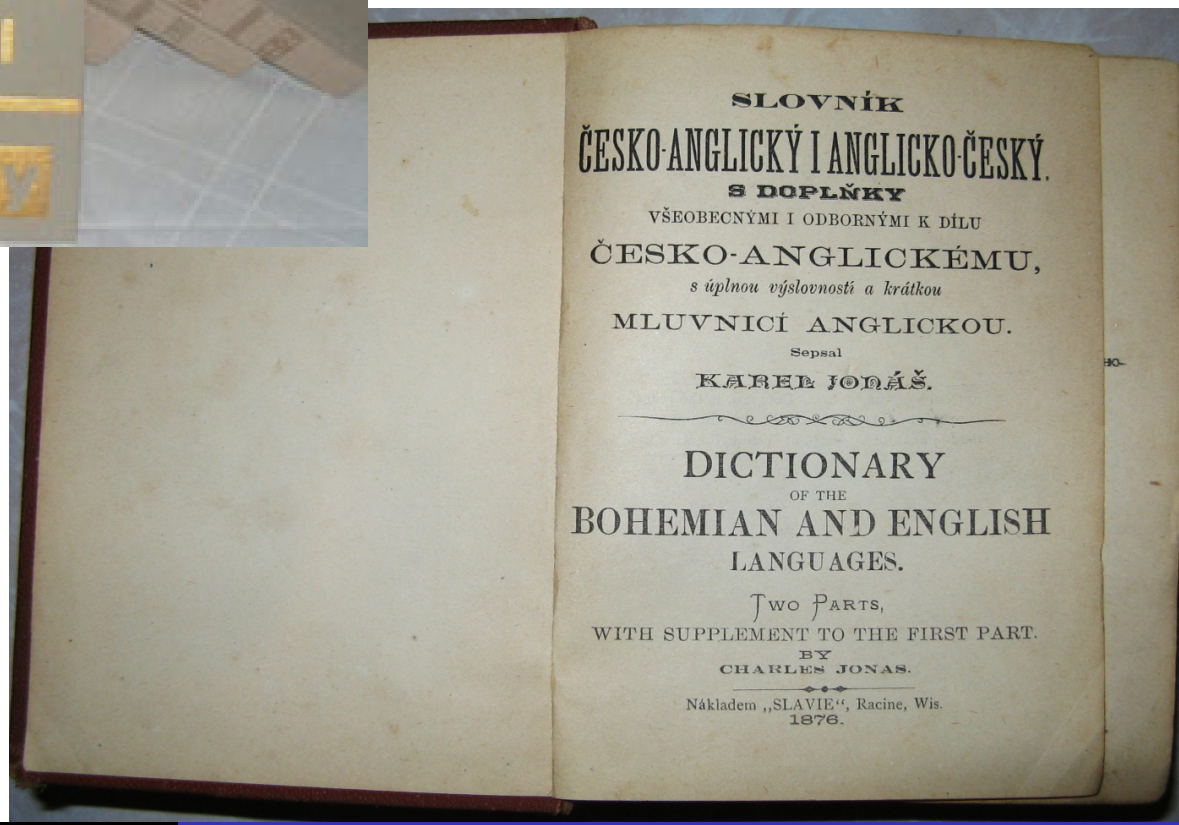
DEB II – Lexikografická platforma pro vývoj slovníkových aplikací

Adam Rambousek

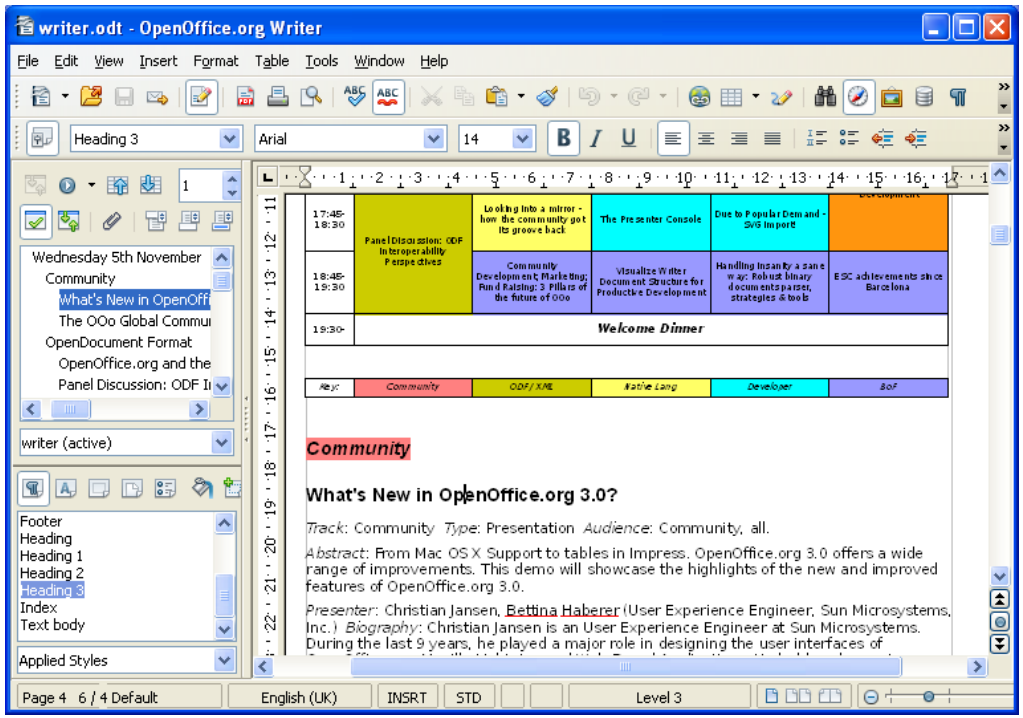
Centrum zpracování přirozeného jazyka
Fakulta informatiky, Masarykova univerzita
xrambous@fi.muni.cz
deb@aurora.fi.muni.cz
http://deb.fi.muni.cz

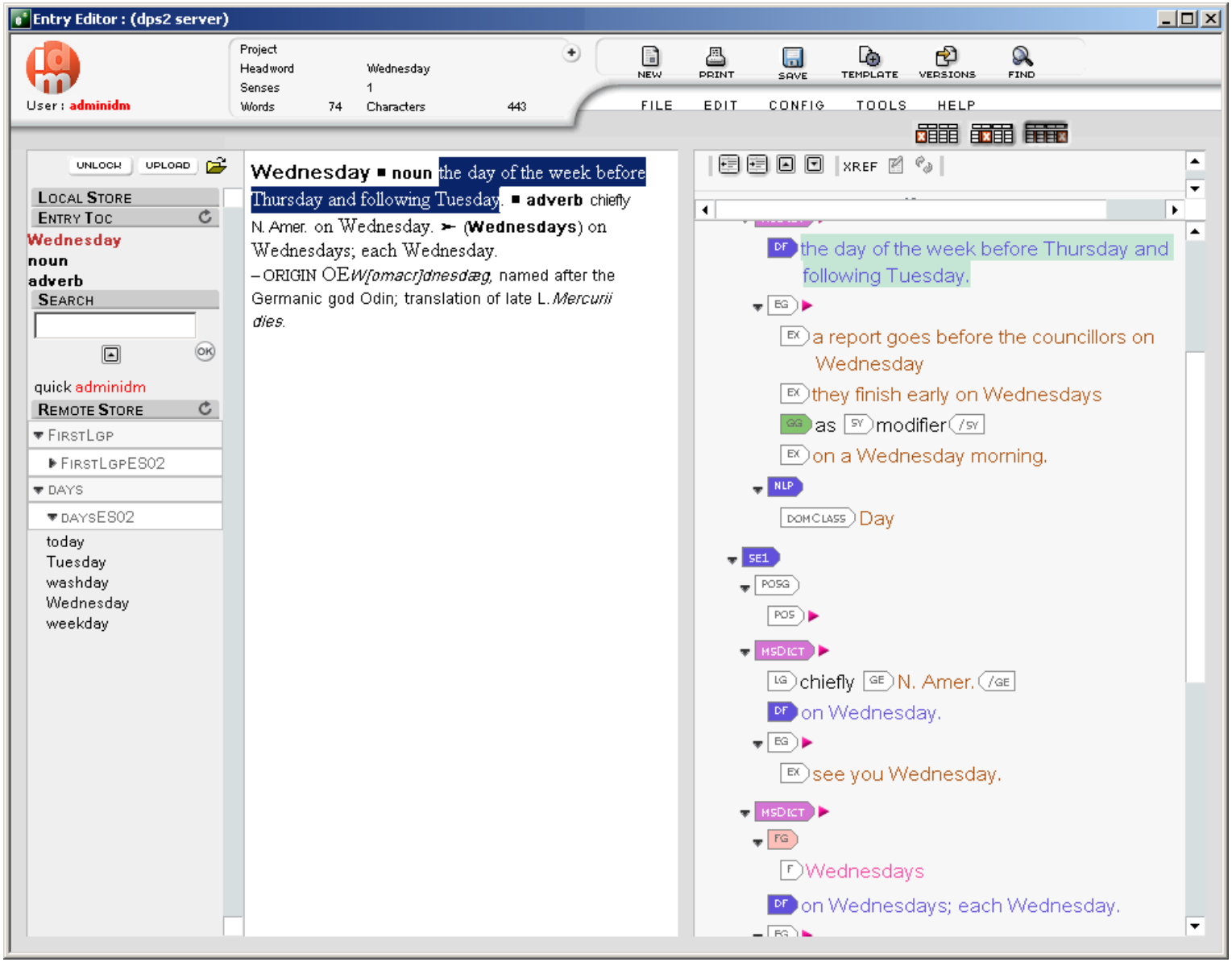

SLOVNÍK
ČESKO-ANGLICKÝ I ANGLICKO-ČESKÝ.
S DOPLŇKY
VŠEOBECNÝMI I ODBORNÝMI K DÍLU
ČESKO-ANGLICKÉMU,
s úplnou výslovností a krátkou
MLUVNICÍ ANGLICKOU.
Sepsal
KAREL JONÁŠ.

DICTIONARY
OF THE
BOHEMIAN AND ENGLISH
LANGUAGES.
TWO PARTS,
WITH SUPPLEMENT TO THE FIRST PART.
BY
CHARLES JONAS.
Nákladem „SLAVIE“, Racine, Wis.
1876.


Underwood


Underwood


writer.odt - OpenOffice.org Writer
File Edit View Insert Format Table Tools Window Help
Heading 3
Arial
14
Wednesday 5th November
Community
What's New in OpenOff
The OOo Global Commu
OpenDocument Format
OpenOffice.org and the
Panel Discussion: ODF I
writer (active)
Footer
Heading
Heading 1
Heading 2
Heading 3
Index
Text body
Applied Styles
Panel Discussion: ODF Interoperability Perspectives
Looking into a mirror - how the community got its groove back
The Presenter Console
Due to Popular Demand - SVG Import
Community Development, Marketing, Fund Raising: 3 Pillars of the future of OOo
Visualize Writer Document Structure for Productive Development
Handling Insanity a sane way: Robust binary documents parser, strategies & tools
ESC achievements since Barcelona
Welcome Dinner
Community
What's New in OpenOffice.org 3.0?
Track: Community Type: Presentation Audience: Community, all.
Abstract: From Mac OS X Support to tables in Impress. OpenOffice.org 3.0 offers a wide range of improvements. This demo will showcase the highlights of the new and improved features of OpenOffice.org 3.0.
Presenter: Christian Jansen, Bettina Haberer (User Experience Engineer, Sun Microsystems, Inc.) Biography: Christian Jansen is an User Experience Engineer at Sun Microsystems. During the last 9 years, he played a major role in designing the user interfaces of
Page 4 6 / 4 Default
English (UK)
INSRT
STD
Level 3


Entry Editor : (dps2 server)
User : adminidm
Project
Headword Wednesday
Senses 1
Words 74 Characters 443
NEW PRINT SAVE TEMPLATE VERSIONS FIND
FILE EDIT CONFIG TOOLS HELP
UNLOCK UPLOAD
LOCAL STORE
ENTRY TOC
Wednesday
noun
adverb
SEARCH
quick adminidm
REMOTE STORE
FIRSTLGP
FIRSTLGPES02
DAYS
DAYSES02
today
Tuesday
washday
Wednesday
weekday
Wednesday ■ noun the day of the week before Thursday and following Tuesday. ■ adverb chiefly N. Amer. on Wednesday. ➢ (Wednesdays) on Wednesdays; each Wednesday.
– ORIGIN OE Wōdnesdæg, named after the Germanic god Odin; translation of late L. Mercurii dies.
XREF
DF the day of the week before Thursday and following Tuesday.
EG
EX a report goes before the councillors on Wednesday
EX they finish early on Wednesdays
GG as SY modifier /SY
EX on a Wednesday morning.
NLP
DOMCLASS Day
SE1
POSG
POS
MSDICT
LG chiefly GE N. Amer. /GE
DF on Wednesday.
EG
EX see you Wednesday.
MSDICT
FG
F Wednesdays
DF on Wednesdays; each Wednesday.
EG

Slovníky a počítače

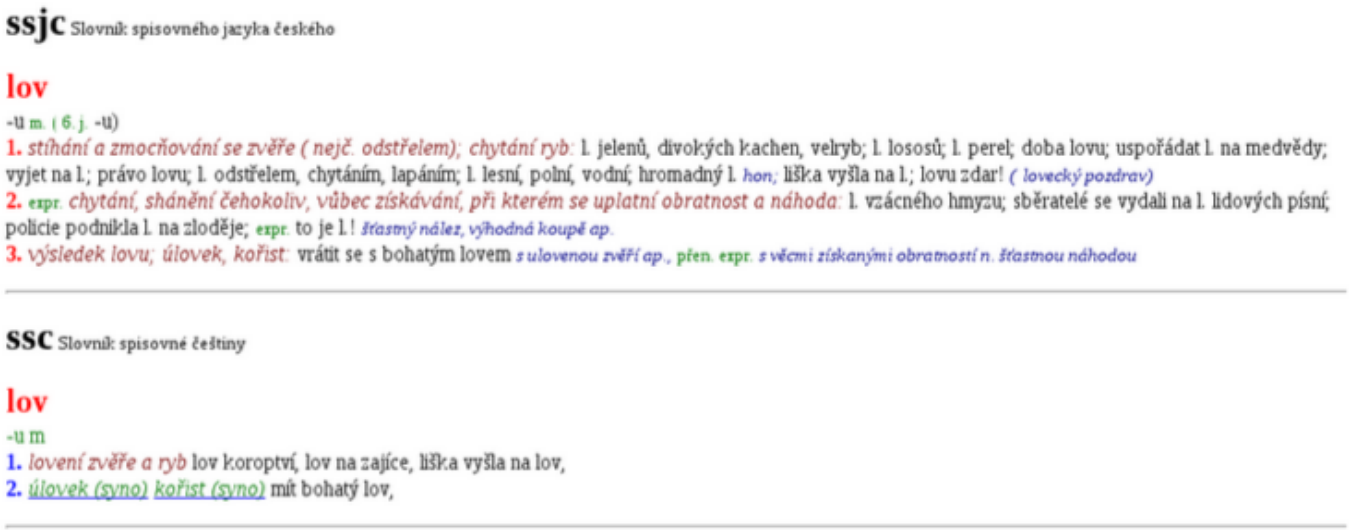
- 60. léta - používají se počítače, lexikografové píší na papír, specialisté přepisují do databáze, Brown Corpus
- 1978, *Longman Dictionary of Contemporary English*
  - první s omezeným slovníkem definicí, kontrolováno strojově
  - kódování pro NLP výzkum
- 1980, *COBUILD*, University of Birmingham + Collins
  - korpus současných textů (Bank of English)
  - 1987, *Collins COBUILD English Language Dictionary*
  - první slovník založený na korpusových datech
  - nový styl definice - celé věty
  - *If a person, animal, or other living thing* **is killed**, *something or someone causes them to die.*
- 90. léta - vývoj specializovaných systémů pro tvorbu slovníků
- 1987, Text Encoding Initiative

XML

- PB138 Moderní značkovací jazyky
- eXtensible Markup Language - značkovací (meta)jazyk
- pravidla, jak má vypadat správně vytvořený dokument - snadné strojové zpracování a výměna informací
- konkrétní názvy značek určuje uživatel (standardy, vlastní)

Zobrazení

- **XSLT** – eXtensible Stylesheet Language (Transformations)
- převod XML na jiné formáty
  - jiné XML značkování, text, HTML, LaTeX, PDF
- šablony pro části XML dokumentu, postupné procházení dokumentu
- funkcionální programovací jazyk



SSJC Slovník spisovného jazyka českého

lov

-u m. (6. j. -u)

1. *stíhání a zmocňování se zvěře (nejč. odstřelem); chytání ryb:* l. jelenů, divokých kachen, velryb; l. lososů; l. perel; doba lovu; uspořádat l. na medvědy; vyjet na l.; právo lovu; l. odstřelem, chytáním, lapáním; l. lesní, polní, vodní; hromadný l. *hon;* liška vyšla na l.; lovu zdar! *(lovecký pozdrav)*

2. expr. *chytání, shánění čehokoliv, vůbec získávání, při kterém se uplatní obratnost a náhoda:* l. vzácného hmyzu; sběratelé se vydali na l. lidových písní; policie podnikla l. na zloděje; expr. to je l.! *šťastný nález, výhodná koupě ap.*

3. *výsledek lovu; úlovek, kořist:* vrátit se s bohatým lovem *s ulovenou zvěří ap.*, přen. expr. *s věcmi získanými obratností n. šťastnou náhodou*

SSC Slovník spisovné češtiny

lov

-u m

1. *lovení zvěře a ryb* lov kaprů, lov na zajíce, liška vyšla na lov,

2. *úlovek (syno) kořist (syno)* mít bohatý lov,

Ukládání

- XML databáze
- ukládají se přímo XML dokumenty
- vyhledávání - XPath, XQuery
- např. eXist, BaseX, Sedna

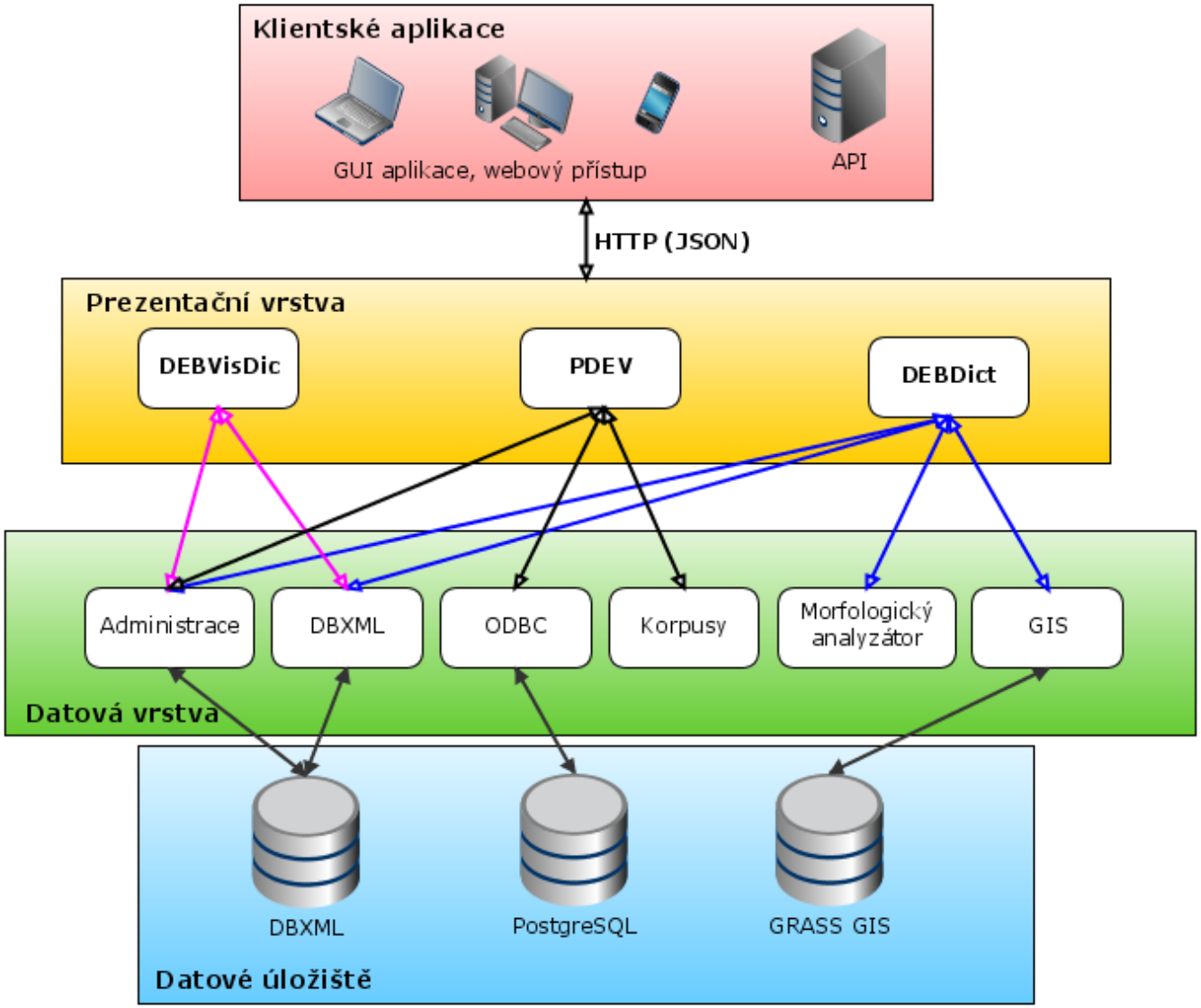
TEI

- *Text Encoding Initiative*, `http://www.tei-c.org/`
- *TEI Guidelines* (aktuálně verze 5 z roku 2007)
- XML formát pro sémantický popis textových dokumentů
- velký rozsah značek
- *TEI Lite* – osekaná verze, *"90 % potřeb 90 % uživatelů"*
- romány, poezie, divadelní hry, dokumentace, slovníky, korpusy, grafy, rukopisy, zarovnání, odkazy, změny textu, notové zápisy...
- nástroje - sada XSLT pro převod na LaTeX, docx, EPUB, HTML

Úvod

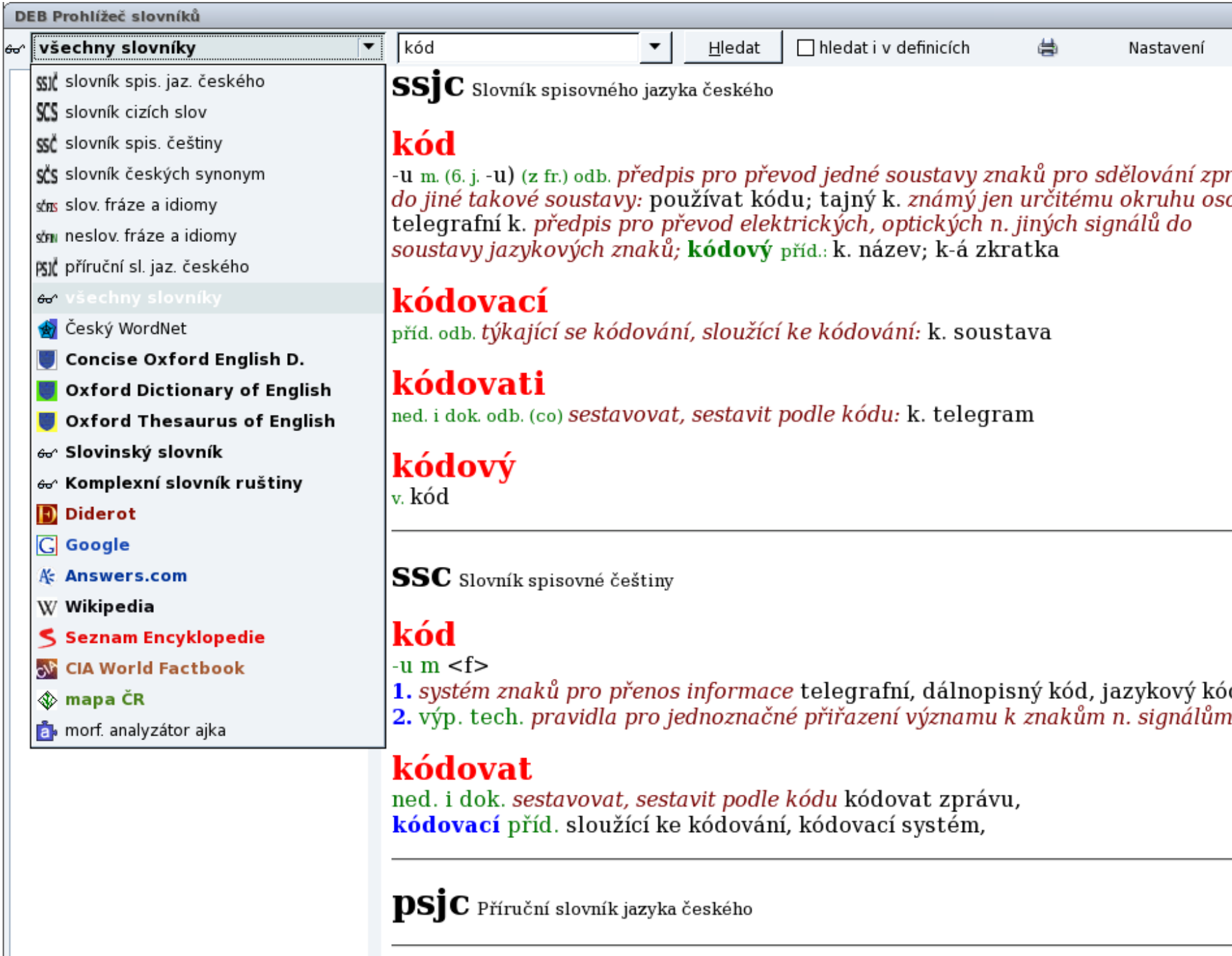
DEB – Dictionary Editor and Browser

- platforma pro vývoj slovníkových aplikací
  - všechna data ve formátu *XML*, *Unicode*
  - libovolná struktura, jakýkoliv jazyk
- architektura klient-server
- *server*
  - práce s daty, velká část funkcí
  - databázové úložiště
  - správa uživatelů, slovníků, spolupráce
  - rozděleno do modulů, spojování podle potřeb
- *klient*
  - omezená funkcionalita
  - grafické nebo webové rozhraní


Klientské aplikace
GUI aplikace, webový přístup
API
HTTP (JSON)
Prezentační vrstva
DEBVisDic
PDEV
DEBDict
Administrace
DBXML
ODBC
Korpusy
Morfologický analyzátor
GIS
Datová vrstva
DBXML
PostgreSQL
GRASS GIS
Datové úložiště

DEBDict

- prohlížeč slovníků
- 7 slovníků češtiny pro veřejnost, další přístupné jen pro část uživatelů
- napojení na morfologický analyzátor
- data z externích zdrojů
- přes 700 uživatelů (podepsané prohlášení)


DEB Prohlížeč slovníků
všechny slovníky
kód
Hledat
hledat i v definicích
Nastavení
slovník spis. jaz. českého
slovník cizích slov
slovník spis. češtiny
slovník českých synonym
slov. fráze a idiomy
neslov. fráze a idiomy
příruční sl. jaz. českého
všechny slovníky
Český WordNet
Concise Oxford English D.
Oxford Dictionary of English
Oxford Thesaurus of English
Slovinský slovník
Komplexní slovník ruštiny
Diderot
Google
Answers.com
Wikipedia
Seznam Encyklopedie
CIA World Factbook
mapa ČR
morf. analyzátor ajka
ssjc Slovník spisovného jazyka českého
kód
-u m. (6. j. -u) (z fr.) odb. předpis pro převod jedné soustavy znaků pro sdělování zpr
do jiné takové soustavy: používat kódu; tajný k. známý jen určitému okruhu osc
telegrafní k. předpis pro převod elektrických, optických n. jiných signálů do
soustavy jazykových znaků; kódový příd.: k. název; k-á zkratka
kódovací
příd. odb. týkající se kódování, sloužící ke kódování: k. soustava
kódovati
ned. i dok. odb. (co) sestavovat, sestavit podle kódu: k. telegram
kódový
v. kód
ssc Slovník spisovné češtiny
kód
-u m <f>
1. systém znaků pro přenos informace telegrafní, dálnopisný kód, jazykový kó
2. výp. tech. pravidla pro jednoznačné přiřazení významu k znakům n. signálům
kódovat
ned. i dok. sestavovat, sestavit podle kódu kódovat zprávu,
kódovací příd. sloužící ke kódování, kódovací systém,
psjc Příruční slovník jazyka českého

Praled

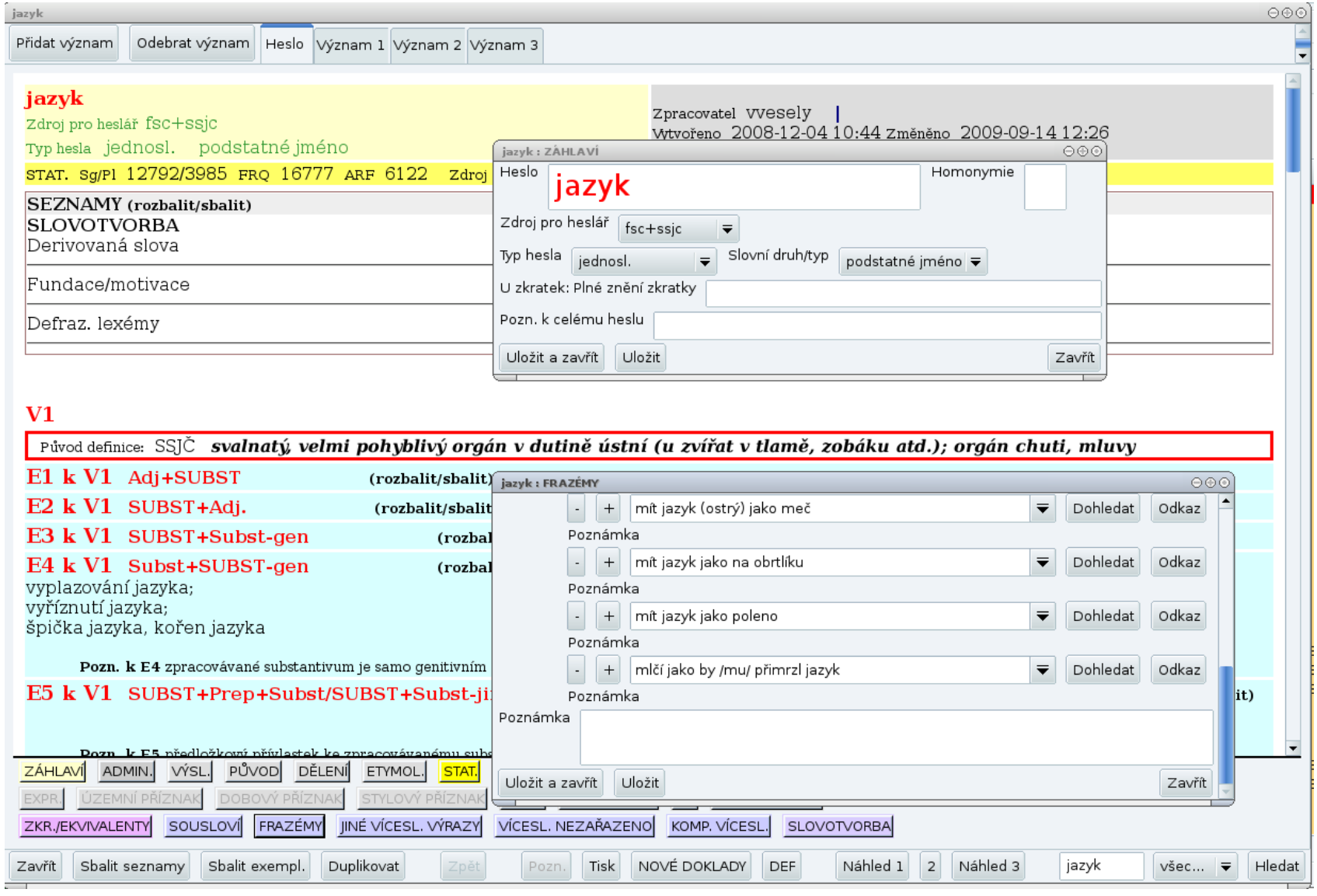
- příprava lexikální databáze češtiny v Ústavu pro jazyk český

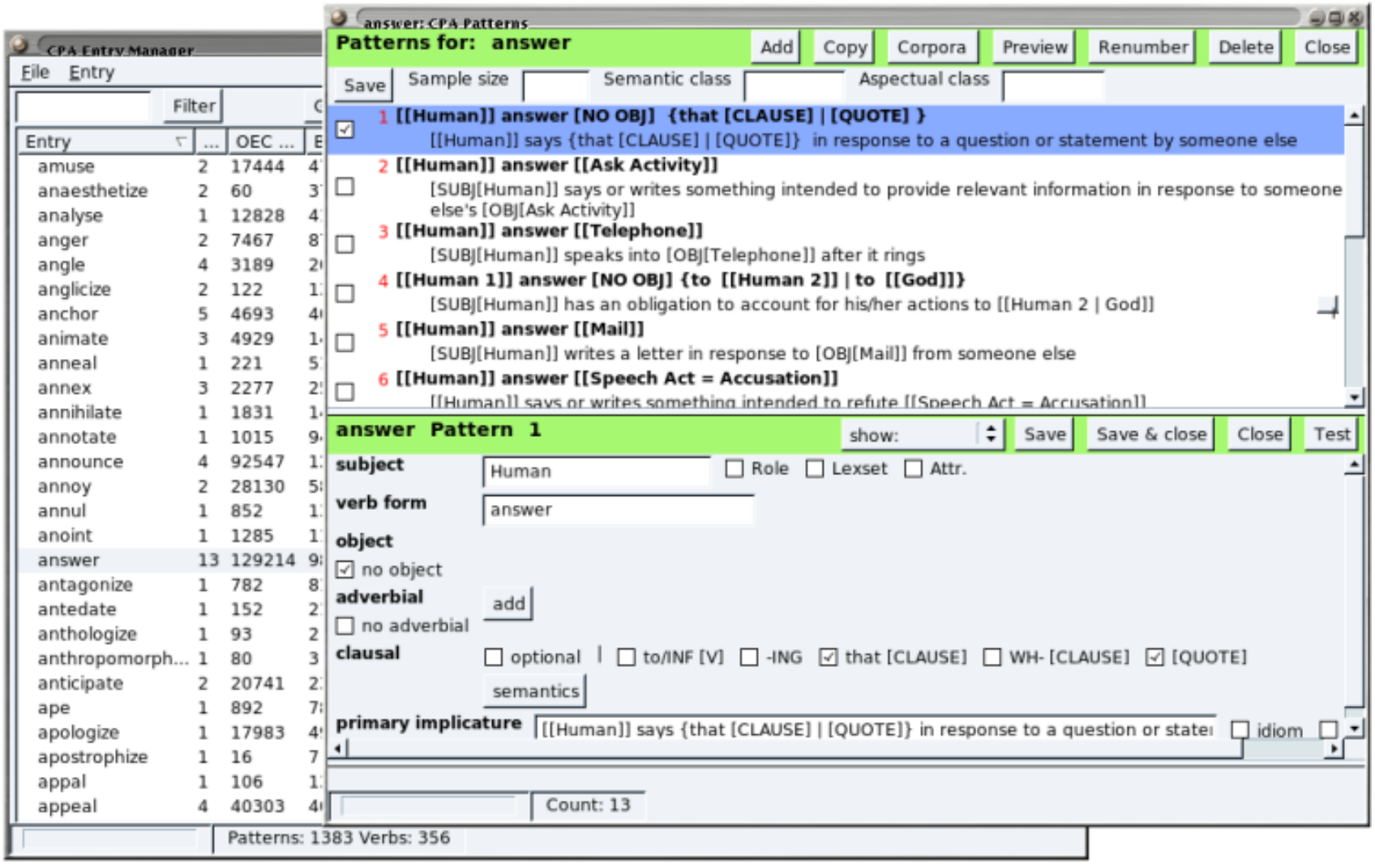
PDEV (CPA)

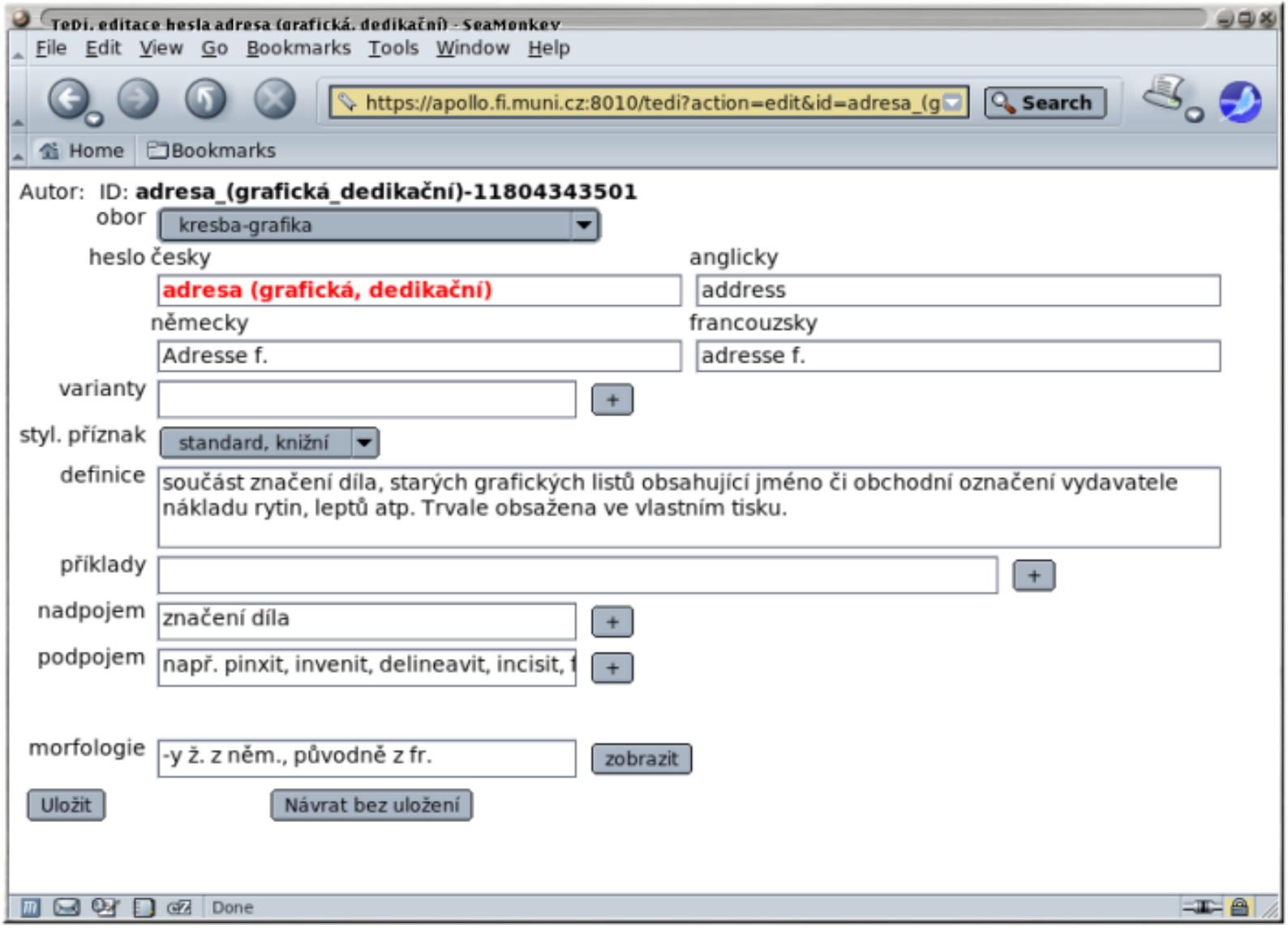
- editace slovníku vzorů anglických sloves
- varianty také pro češtinu, italštinu a španělštinu

TeDi – Terminologický slovník

- společný projekt s Fakultou výtvarných umění VUT
- glosář výtvarných pojmů
- multimediální prvky
- nově také Divadelní fakulta JAMU, Agronomická fakulta MU

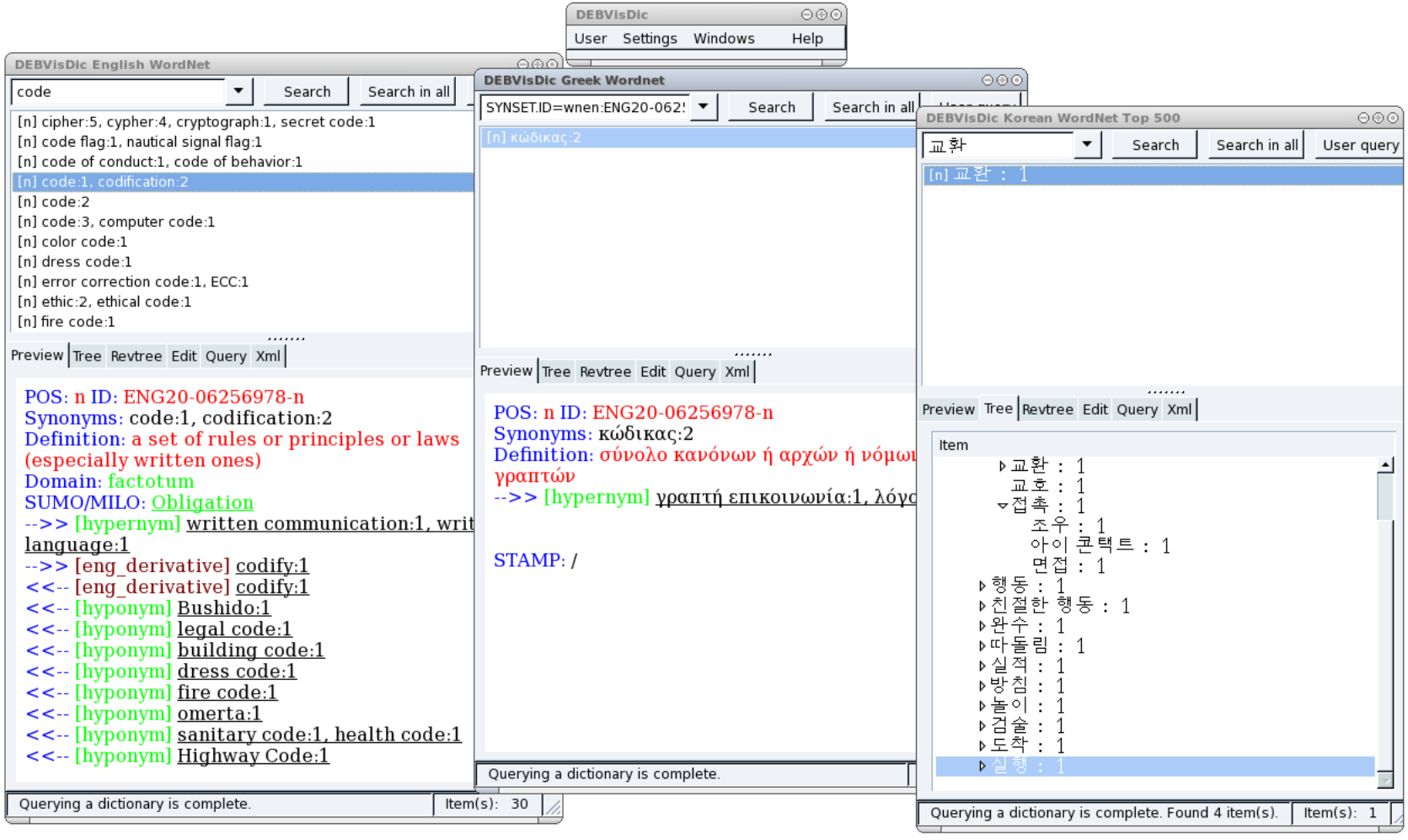

jazyk
Přidat význam
Odebrat význam
Heslo
Význam 1
Význam 2
Význam 3
jazyk
Zdroj pro heslář fsc+ssjc
Typ hesla jednosl. podstatné jméno
Zpracovatel vvesely
Vytvořeno 2008-12-04 10:44 Změněno 2009-09-14 12:26
STAT. Sg/Pl 12792/3985 FRQ 16777 ARF 6122 Zdroj
SEZNAMY (rozbalit/sbalit)
SLOVOTVORBA
Derivovaná slova
Fundace/motivace
Defraz. lexémy
jazyk : ZÁHLAVÍ
Heslo
jazyk
Homonymie
Zdroj pro heslář
fsc+ssjc
Typ hesla
jednosl.
Slovní druh/typ
podstatné jméno
U zkratek: Plné znění zkratky
Pozn. k celému heslu
Uložit a zavřít
Uložit
Zavřít
V1
Původ definice: SSJČ svalnatý, velmi pohyblivý orgán v dutině ústní (u zvířat v tlamě, zobáku atd.); orgán chuti, mluvy
E1 k V1 Adj+SUBST (rozbalit/sbalit)
E2 k V1 SUBST+Adj. (rozbalit/sbalit
E3 k V1 SUBST+Subst-gen (rozbal
E4 k V1 Subst+SUBST-gen (rozbal
vyplazování jazyka;
vyříznutí jazyka;
špička jazyka, kořen jazyka
Pozn. k E4 zpracovávané substantivum je samo genitivním
E5 k V1 SUBST+Prep+Subst/SUBST+Subst-ji
jazyk : FRÁZÉMY
mít jazyk (ostrý) jako meč
Dohledat
Odkaz
Poznámka
mít jazyk jako na obrtlíku
mít jazyk jako poleno
mlčí jako by /mu/ přimrzl jazyk
Poznámka
Uložit a zavřít
Uložit
Zavřít
ZÁHLAVÍ
ADMIN.
VÝSL.
PŮVOD
DĚLENÍ
ETYMOL.
STAT.
EXPR.
ÚZEMNÍ PŘÍZNAK
DOBOVÝ PŘÍZNAK
STYLOVÝ PŘÍZNAK
ZKR./EKVIVALENTY
SOUSLOVÍ
FRÁZÉMY
JINÉ VÍCESL. VÝRAZY
VÍCESL. NEZAŘAZENO
KOMP. VÍCESL.
SLOVOTVORBA
Zavřít
Sbalit seznamy
Sbalit exempl.
Duplikovat
Zpět
Pozn.
Tisk
NOVÉ DOKLADY
DEF
Náhled 1
2
Náhled 3
jazyk
všec...
Hledat


CPA Entry Manager
File Entry
Filter
Entry ... OEC ...
amuse 2 17444
anaesthetize 2 60
analyse 1 12828
anger 2 7467
angle 4 3189
anglicize 2 122
anchor 5 4693
animate 3 4929
anneal 1 221
annex 3 2277
annihilate 1 1831
annotate 1 1015
announce 4 92547
annoy 2 28130
annul 1 852
anoint 1 1285
answer 13 129214
antagonize 1 782
antedate 1 152
anthologize 1 93
anthropomorph... 1 80
anticipate 2 20741
ape 1 892
apologize 1 17983
apostrophize 1 16
appal 1 106
appeal 4 40303
Patterns: 1383 Verbs: 356
answer: CPA Patterns
Patterns for: answer
Add Copy Corpora Preview Renumber Delete Close
Save Sample size Semantic class Aspectual class
1 [[Human]] answer [NO OBJ] {that [CLAUSE] | [QUOTE] }
[[Human]] says {that [CLAUSE] | [QUOTE]} in response to a question or statement by someone else
2 [[Human]] answer [[Ask Activity]]
[SUBJ[Human]] says or writes something intended to provide relevant information in response to someone else's [OBJ[Ask Activity]]
3 [[Human]] answer [[Telephone]]
[SUBJ[Human]] speaks into [OBJ[Telephone]] after it rings
4 [[Human 1]] answer [NO OBJ] {to [[Human 2]] | to [[God]]}
[SUBJ[Human]] has an obligation to account for his/her actions to [[Human 2 | God]]
5 [[Human]] answer [[Mail]]
[SUBJ[Human]] writes a letter in response to [OBJ[Mail]] from someone else
6 [[Human]] answer [[Speech Act = Accusation]]
[[Human]] says or writes something intended to refute [[Speech Act = Accusation]]
answer Pattern 1
show: Save Save & close Close Test
subject Human Role Lexset Attr.
verb form answer
object
no object
adverbial add
no adverbial
clausal optional to/INF [V] -ING that [CLAUSE] WH- [CLAUSE] [QUOTE]
semantics
primary implicature [[Human]] says {that [CLAUSE] | [QUOTE]} in response to a question or state idiom
Count: 13


TeDi, editace hesla adresa (grafická, dedikační) - SeaMonkey
File Edit View Go Bookmarks Tools Window Help
https://apollo.fi.muni.cz:8010/tedi?action=edit&id=adresa_(g
Search
Home
Bookmarks
Autor: ID: adresa_(grafická_dedikační)-11804343501
obor
kresba-grafika
heslo česky
anglicky
adresa (grafická, dedikační)
address
německy
francouzsky
Adresse f.
adresse f.
varianty
+
styl. příznak
standard, knižní
definice
součást značení díla, starých grafických listů obsahující jméno či obchodní označení vydavatele nákladu rytin, leptů atp. Trvale obsažena ve vlastním tisku.
příklady
+
nadpojem
značení díla
+
podpojem
např. pinxit, invenit, delineavit, incisit, f
+
morfologie
-y ž. z něm., původně z fr.
zobrazit
Uložit
Návrat bez uložení
Done

DEBVisDic

- editor slovníků typu WordNet
- samostatný modul pro každý jazyk (modifikace)
- použit pro tvorbu několika wordnetů
- poskytuje API – napojení externích aplikací
- možnost rozšíření a modifikací


DEBVisDic
User Settings Windows Help
DEBVisDic English WordNet
code
Search
Search in all
[n] cipher:5, cypher:4, cryptograph:1, secret code:1
[n] code flag:1, nautical signal flag:1
[n] code of conduct:1, code of behavior:1
[n] code:1, codification:2
[n] code:2
[n] code:3, computer code:1
[n] color code:1
[n] dress code:1
[n] error correction code:1, ECC:1
[n] ethic:2, ethical code:1
[n] fire code:1
Preview Tree Revtree Edit Query Xml
POS: n ID: ENG20-06256978-n
Synonyms: code:1, codification:2
Definition: a set of rules or principles or laws (especially written ones)
Domain: factotum
SUMO/MILO: Obligation
-->> [hypernym] written communication:1, writ
language:1
-->> [eng_derivative] codify:1
<<-- [eng_derivative] codify:1
<<-- [hyponym] Bushido:1
<<-- [hyponym] legal code:1
<<-- [hyponym] building code:1
<<-- [hyponym] dress code:1
<<-- [hyponym] fire code:1
<<-- [hyponym] omerta:1
<<-- [hyponym] sanitary code:1, health code:1
<<-- [hyponym] Highway Code:1
Querying a dictionary is complete.
Item(s): 30
DEBVisDic Greek Wordnet
SYNSET.ID=wnen:ENG20-062
Search
Search in all
[n] κώδικας:2
Preview Tree Revtree Edit Query Xml
POS: n ID: ENG20-06256978-n
Synonyms: κώδικας:2
Definition: σύνολο κανόνων ή αρχών ή νόμω
γραπτών
-->> [hypernym] γραπτή επικοινωνία:1, λόγ
STAMP: /
Querying a dictionary is complete.
DEBVisDic Korean WordNet Top 500
교환
Search
Search in all
User query
[n] 교환 : 1
Preview Tree Revtree Edit Query Xml
Item
교환 : 1
교호 : 1
접촉 : 1
조우 : 1
아이 콘택트 : 1
면접 : 1
행동 : 1
친절한 행동 : 1
완수 : 1
따돌림 : 1
실적 : 1
방침 : 1
놀이 : 1
검술 : 1
도착 : 1
실행 : 1
Querying a dictionary is complete. Found 4 item(s).
Item(s): 1