IB047

Lexikografie

Pavel Rychlý

pary@fi.muni.cz

3. května 2024

Lexikografie

- věda, která se zabývá slovníky
- zejména budováním slovníků

Lexikografie

- věda, která se zabývá slovníky
- zejména budováním slovníků
- Jak budovat slovník?

Lexikografie

- věda, která se zabývá slovníky
- zejména budováním slovníků
- Jak budovat slovník?
  - většinou rozsáhlé projekty
  - nutnost dělat řadu rozhodnutí

Návrh slovníku

Vydavatel

- produkt
  - trh, uživatelé, konkurence
  - formát, obsah, cena
- rozpočet
  - časový plán
  - personální obsazení
    editoři, lexikografové
  - zdroje
    korpus, počítače, nástroje

Návrh slovníku

Editor

- výběr obsahu
  - seznam slov (slovních spojení)
  - typy informací ve slovníku
- metody prezentace
  - mapování jednotlivých typů informací do grafické podoby

Pro koho je slovník určen?

- dospělí / děti
- rodilý mluvčí / cizinec
- akademické použití / veřejnost
- překladový / jednojazyčný

Hlavní typy hesel

- lexikální
  a = spojka, částice
- encyklopedické
  a = ar, *A* = ampér
- rozdělení není vždy jasné
  a = česká samohláska a písmeno (SSJČ)
  a = první písmeno většiny abeced (Wikipedie)
- Oxford Advanced Learner's Encyclopedic Dictionary
- encyklopedie = naučný slovník (SSJČ)

Hlavní typy hesel

lexikální

- běžný jazyk
- substantiva, adjektiva, slovesa, příslovce (otevřené slovní druhy)
- zájmena, číslovky, spojky
- částice, citoslovce, pomocná slova

Hlavní typy hesel

encyklopedické

- oborově zaměřené
- vlastní jména
- obecná
  - názvy měsíců, svátků, jména osob
- jednoznačná
  - osoby, místa, firmy, obchodní značky

Jak vybrat seznam hesel

- nečetnější slova
- dokumentová četnost
- vyfiltrovat jména

Obsah heslového odstavce 1

- heslo (heslové slovo)
- číslo hesla (u homografie)
- výslovnost
- varianty (dubletní tvary)
- slovní druh
- gramatické informace
  - rod, skloňování, časování
- oddělovač významů

kurs

-u [-zu] m. (6. j. -u, -e) (z lat.)

1. řidč. *doba platnosti (peněz, cenin ap.);*

2. peněž. *hodnota peněz a cenných papírů*

3. *směr, ráz (zejm. politiky); režim:* se zn

4. dopr. *směr, jímž loď pluje;* let. *úhel, kt. s*

dopravního prostředku

5. *soubor přednášek, učebných lekcí urč.*

lidové kursy ruštiny; --- **kursový** [-zo-] p

mezinárodní dopravě) přímý

Obsah heslového odstavce 2

- lingvistické informace
- syntaktické doplňky
- definice
- glosa (vysvětlení)
- překlad
- příklad
- víceslovná spojení
- podheslo

Obsah heslového odstavce 2

- odvozená slova
- křížové odkazy
- poznámky k použití
- četnost slova

Lingvistické informace

- region
- styl
- obor
- čas
- postoj
- obraznost

Návrh zásad slovníku

Style Guide

- sada pravidel pro každou část heslového odstavce, kterými se budou řídit editoři a lexikografové při tvorbě slovníku

Návrh zásad slovníku

Proč?

- lexikograf
  - jistota, konzistence
- uživatel
  - snadnost použití
  - důvěryhodnost

Návrh zásad slovníku

- používání zkratek a značek
- pravopis
- použitá slovní zásoba
- v průběhu budování slovníku se může doplňovat/měnit

Heslové šablony

Template Entry

- nástroj pro systematickou tvorbu heslových odstavců pro slova ze stejné lexikální množiny

Heslové šablony

Template Entry

- nástroj pro systematickou tvorbu heslových odstavců pro slova ze stejné lexikální množiny
- lexikální množina
  - skupina slov majících společný prvek významu

Heslové šablony

příklady lexikálních množin

- ptáci, květiny, kovy, ...
- téměř absolutní upřesnění zásad slovníku pro danou lexikální množinu

Formats for storing dictionary entries

- text files vs. database
  - database contains entries
  - each entry in text format
- XML, JSON, NVH
- tree structure
- schema

XML

- TEI Guidelines
- TEI Lex-0
  - A baseline encoding for lexicographic data
  - Europian projects: ENeL, ELEXIS, DARIAH
  - https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

```
<entry type="mainEntry" xml:lang="cz" xml:id="en000008">
  <form type="lemma" xml:id="en000008.hw1">
     <orth>abeceda</orth>
  </form>
  <pc>,</pc>
  <form type="inflected">
     <gramGrp>
        <gram type="case" value="genitiv"/>
        <gram type="number" value="singular"/>
        <gram type="gender" value="feminine"/>
     </gramGrp>
     <orth extent="suffix" expand="abecedy">-y</orth>
  </form>
  <!--...-->
</entry>
```

JSON

- JavaScript Object Notation
- data interchange format
- attribute–value pairs and arrays

JSON – example

```
{ "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

NVH: Name-Value Hierarchy

- less verbose alternative to XML
- name, value, list of child elements

```
headword: bank
  partOfSpeech: noun
  definition: an institution where you store or borrow money
    translation: банка
    example: I got a large loan from the bank.
      translation: Я получил крупный кредит в банке.
  definition: a stretch of land along a river
    translation: берег
    example: The house is on the north bank of the river.
      translation: Дом находится на северном берегу реки.
```

NVH: Compare to XML

```
<entry>
  <headword>bank</headword>
  <partOfSpeech>noun</partOfSpeech>
  <sense>
    <definition>an institution where you store or borrow money</definition>
    <translation>банка</translation>
    <exampleContainer>
      <example>I got a large loan from the bank.</example>
      <translation>Я получил крупный кредит в банке.</translation>
    </exampleContainer>
  </sense>
  <sense>
    <definition>a stretch of land along a river</definition>
    <translation>берег</translation>
    <exampleContainer>
      <example>The house is on the north bank of the river.</example>
      <translation>Дом находится на северном берегу реки.</translation>
    </exampleContainer>
  </sense>
</entry>
```

LEXIDMA

- OASIS Lexicographic Infrastructure Data Model and API
- OASIS Open Groups
- Technical Committee (Chair: Michal Měchura - FI MU)
- simple, modular, and easy to adopt data model
- standard serialization independent interchange objects
- open standard - anyone can use