InteGra (v 1.0)

Dokumentace k databázi

Nářečí českého jazyka interaktivně. Dokumentace a zpřístupnění mizejícího jazykového dědictví jako nedílné součásti regionálních identit.

DG20P02OVV029; poskytovatel podpory Ministerstvo kultury, Program NAKI II.

Obsah

Základní údaje

Název výsledku

InteGra

Druh výsledku

S - specializovaná veřejná databáze

Lokalizace výsledku

https://www.ceskanareci.cz/geoportal/integra/

Lokalizace dokumentace

https://www.ceskanareci.cz/geoportal/integra/dokumentace_integra.pdf

Vlastník

Univerzita Palackého v Olomouci a Ústav pro jazyk český AV ČR, v. v. i.

Stručný popis produktu

Specializovaná databáze gramatických dat, jejich interpolací a variant. Umožňuje vyhledávat, filtrovat a prohlížet uložená a interpolovaná data v tabelární formě, včetně vytváření pokročilých dotazů a filtrů.

Technické parametry výsledku

Databázový server MariaDB, jazyk SQL; webový server Apache; front-end uživatelské rozhraní využívá technologií php + HTML5 + JavaScript + CSS3.

Ekonomické parametry výsledku:

Během 18 měsíců vývoje a užívání vlastníkem byly do databáze uloženy záznamy o 5 830 363 vypočtených variantách, 15 092 oblastech (části obcí), 367 slovech (heslech) , ze 3301 zdroje (což je více než 3x více než počet uvedený v přihlášce projektu).

Druh možnosti využití výsledku jiným subjektem:

A - k využití výsledku jiným subjektem je vždy nutné nabytí licence

Požadavek na licenční poplatek:

N - poskytovatel licence na výsledek nepožaduje licenční poplatek

Typ licence:

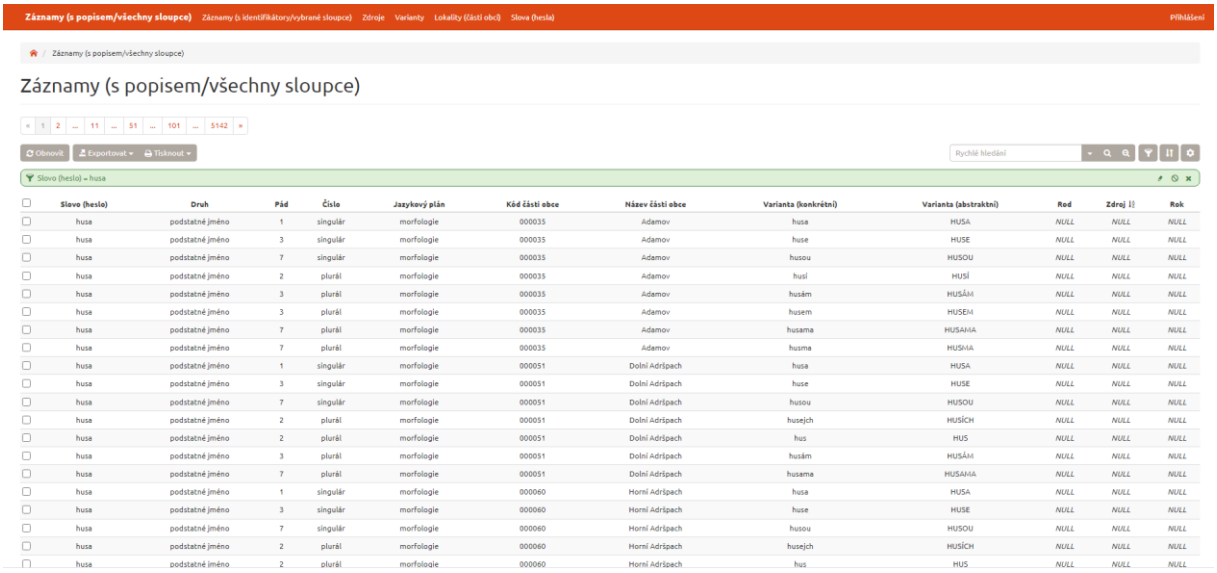
ODbL

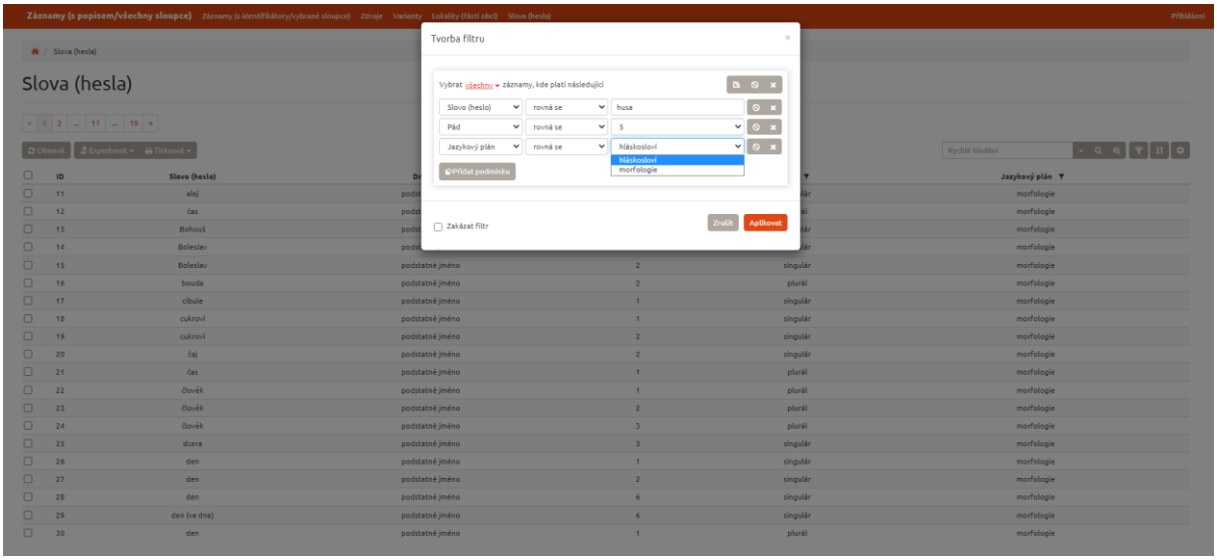
Databáze je zpřístupněna pod licencí ODbL 1.0 (Open Data Commons Open Database License), která umožňuje uživatelům databázi svobodně sdílet, upravovat a používat při dodržení zásady uvedení autora databáze, zachování licence a otevřenosti dat.

Pro uvedení zdroje doporučujeme formulaci: *Data: Databáze InteGra, Ústav pro jazyk český AV ČR, v. v. i., a Univerzita Palackého v Olomouci, 2022.*

Anglická verze: Source data: *InteGra database, Czech Language Institute, Czech Academy of Sciences, and Palacký University Olomouc, 2022.*

Náhled rozhraní





Charakteristiky výsledku

Databáze InteGra je třetí částí propojeného komplexu elektronických dialektologických nástrojů, vznikajících v rámci projektu Nářečí českého jazyka interaktivně. Dokumentace a zpřístupnění mizejícího jazykového dědictví jako nedílné součásti regionálních identit. (DG20P02OVV029; poskytovatel podpory Ministerstvo kultury, Program NAKI II.)

K čemu InteGra slouží?

InteGra je specializovaná databáze pro uložení nářečních gramatických dat, jejich interpolací a variant vygenerovaných softwarem ProMap. Webové rozhraní databáze umožňuje prohlížet data v uživatelsky přívětivé tabelární formě. Pokročilé vyhledávání a filtrování umožňuje generovat libovolné sestavy dat z výzkumů a rekonstruovat je i pro místa, kde výzkum neproběhl (na základě dat prostorově blízkých) a následně sestavy exportovat.

Databáze je určena laické veřejnosti (generování vlastních sestav o nářečí např. v zájmových oblastech nebo dle zájmových slov, případně reportování chyb) i odborné veřejnosti (otevřena k zavádění výsledků výzkumů).

InteGra obsahuje 6 částí:

- Záznamy (s popisem/všechny sloupce) – tabelární výpis všech (necelých 6 000 000) záznamů se všemi dostupnými atributy, vzniknuvší při interpolaci v softwaru ProMap a generování pro geoportál DiaMa
    - Atributy: Slovo (heslo); Druh; Pád; Číslo; Jazykový plán; Kód části obce; Název části obce; Varianta (konkrétní); Varianta (abstraktní); Rod; Zdroj; Rok
- Záznamy (s identifikátory/vybrané sloupce) - tabelární výpis všech (necelých 6 000 000) záznamů s vybranými atributy ve formě číselných identifikátorů
    - Atributy: ID; Slovo (heslo); Lokalita (část obce); Varianta
- Zdroje – číselník 3 300 použitých rešeršních zdrojů pro geoportál DiaMa
    - Atributy: ID, Zdroj (popis)
- Varianty – číselník 4 706 vygenerovaných variant slov (hesel) ze softwaru ProMap pro geoportál DiaMa
    - Atributy: ID; Konkrétní; Abstraktní
- Lokality (části obcí) – číselník 15 092 lokalit (části obcí) pro geoportál DiaMa
    - Atributy: ID; Kód části obce; Název části obce; Obec; Kraj; Okres
- Slova (hesla) – číselník 367 slov (hesel) pro geoportál DiaMa
    - Atributy: ID; Slovo (heslo); Druh; Pád; Číslo; Jazykový plán

Autentizace uživatelů (veřejná vs. neveřejná část)

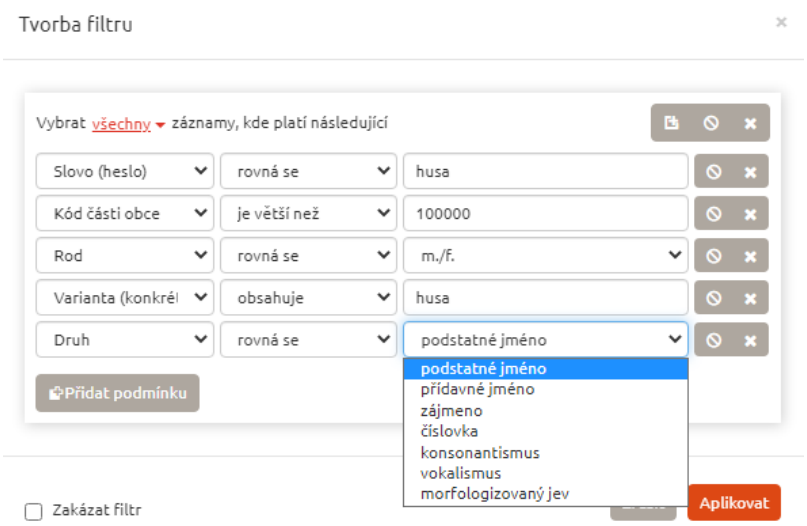
Veřejně dostupná část aplikace obsahuje všechnu funkcionalitu včetně vyhledávání a filtrováí pro generování sestav. Neveřejná část, dostupná po autentizaci jménem a heslem, obsahuje navíc možnost editace a mazání dat. Důvodem je eliminace neautorizovaných a nevhodných zásahů vedoucí ke znehodnocení obsahu „ostré“ databáze dat InteGRA a na ní navázaného vizualizačního portálu DiaMa (ať už záměrných – roboti, nebo nezáměrných plynoucích z neznalosti technického řešení nebo dialektologických pravidel a metodiky).

Autentizace uživatelů je řešena oproti tabulce „phpgen“ uložené v databázi, který zajišťuje autentizaci oproti jménu a heslu.

Implementovaná funkcionalita

Veřejná část

- Rychlé vyhledávání – fulltextové vyhledávání ve všech atributech/sloupcích
- Rychlé vyhledávání – vyhledávání v předem definovaných sloupcích
- Pokročilé vyhledávání - tvorba vlastního vyhledávacího filtru dle zadaných atributů, jejich kombinace a podmínek (rovná se, obsahuje, neobsahuje, větší, menší apod.) pro generování sestav – viz obrázek níže
- Seřazení dle vybraného sloupce
- Vícenásobné řazení sloupců
- Nastavení počtu zobrazených řádků a formy (tabulka, karty/detaily)
- Stránkování
- Export sestavy (PDF, EXCEL, CSV)
- Tisk sestavy (výběr, vše)



Neveřejná (po autentizaci) část

Obsahuje oproti veřejné části navíc:

- přidávání, editace a mazání záznamů

Čím je databáze výjimečná?

Databáze InteGra je inovativní díky kvalitě i kvantitě obsažených dat. V České republice neexistuje jiná dialektologická databáze, která by na jednom místě shromažďovala a umožňovala přístup k téměř 6 000 000 záznamů. Databáze umožňuje pracovat s rejstříkem relevantních tvarů jak interpolovaných oblastí, tak oblastí výzkumu ÚJČ pro danou konkrétní lokalitu, a tudíž bude dosahovat dosud nemyslitelné přesnosti.

Naplnění cílů programu NAKI II a jeho očekávaných přínosů

Databáze obohacuje dialektologické záznamy ověřené výzkumem o záznamy interpolocané pro celé zájmovém území a zároveň umožňuje ověřit (upravit) jak ověřené tak interpolované hodnoty v tabelární formě. Je přístupná externím výzkumným datům, umožňuje reporty nářečních mluvčích a již nyní shromažďuje necelých 6 000 000 záznamů.

Je bezesporu dalším potenciálním zdrojem sběru nářečních dat, v situaci postupného zanikání nářečních jevů. Pro uživatele umožňuje zobrazovat její data v systematickém uspořádání (např. vokalický systém, morfologické paradigma v jisté lokalitě) i v široce volitelných datových konfiguracích. Něčeho takového nelze při užití knižní podoby ČJA dosáhnout.

Odborníkům se tak nesmírně usnadňuje práce s výsledky dialektologických výzkumů a otevírají se jim nové možnosti srovnávání a zřejmě i nových zjištění a objevů. Zájemcům o nářečí se pak zpřístupňuje informace o českých dialektech obecně a prostřednictvím výsledků předkládaného projektu si mnohem snáze uvědomí a oživí regionální a lokální jazykovou identitu svou i svých předků.

Testování a využití softwaru

Databáze byl v beta verzi testován jednak výzkumným týmem Katedry geoinformatiky UP, jednak týmem dialektologického oddělení ÚJČ. Výzkumný tým testoval a vyvíjel databázi společně s softwarem ProMAP od ledna 2022 do září 2022, v září 2022 byla spolu se softwarem ProMap ustálena finální forma databáze tak, aby od září 2022 následně probíhalo plnění dat vygerovaných aplikací ProMap.

Databáze se ukázala jako dostatečně robustní a efektivní. Během 12 měsíců vývoje a užívání vlastníkem bylo prostřednictvím softwaru naimportováno a uloženo do databáze InteGRA 367 slov (hesel) a automatizovaně interpolováno 5 830 363 záznamů pro jednotlivá slova v jednotlivých oblastech (částech obcí).

Technické parametry

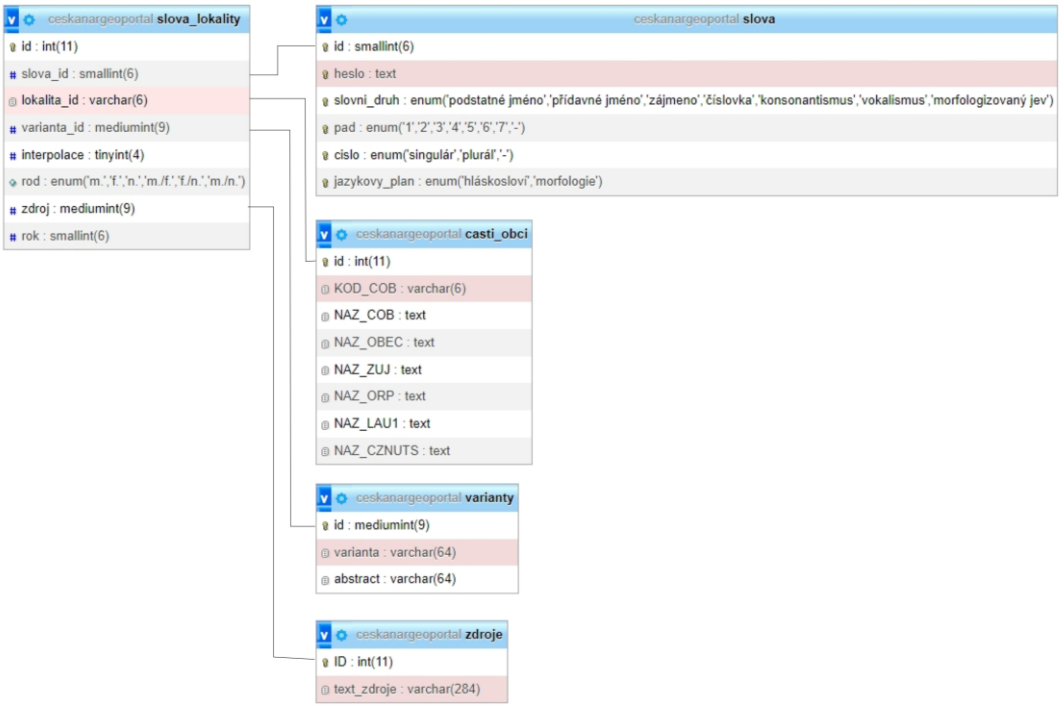
Databázové řešení InteGra seskládá ze 3 částí

- vlastní databáze MariaDB
- backendová část zajišťující funkcionalitu a dotazování do databáze
- frontendová část zajišťující zobrazení dat

Databáze

Vlastní databáze využívá úložiště MariaDB, verze 10.5.18-MariaDB-0+deb11u1-log - Debian 11, výchozí kódování UTF8. Veškeré příkazy jsou prováděny pomocí jazyka SQL. Databáze se skládá z 5 hlavních navzájem propojených tabulek: slova, casti_obci, varianty, zdroje, slova_lokality. Navíc obsahuje tabulku phpgen pro uložení autentizačních údajů.

Export databáze (tzv. dump) včetně struktury i dat ve formátu SQL je dostupný na https://www.ceskanareci.cz/geoportal/integra/dump.sql



Backend

Backendová část zajišťující funkcionalitu - generování SQL dotazů zasílaných do databáze na základě uživatelsky zvolených filtrů a následný přenos vygenerovaných odpovědí formou sestav na frontend. Jako technické řešení byl zvolen kompletně programovací jazyk PHP, pro generování výstupů do formátu PDF bylo využito rozšíření mpdf.

Frontend

Frontendová část zajišťující zobrazení dat formou přívětivého uživatelského rozhraní, v tabelární formě s výše popsanou funkcionalitou. Je postavená na osvědčené a zavedené kombinaci technologii HTML5 + JavaScript + CSS3. Aplikace je dostupná online prostřednictvím zabezpečeného protokolu HTTPS s certifikátem "Let's encrypt". Řešení je responsivní – přizpůsobené pro zobrazení na mobilních telefonech.