Masarykova univerzita
Fakulta informatiky

Projekt LINDAT/CLARIN

PV070 Digitální knihovny

Barbora Obluková
UČO: 437806
3. ročník B-FI PLIN

1.12. 2016

Obsah

Základní údaje

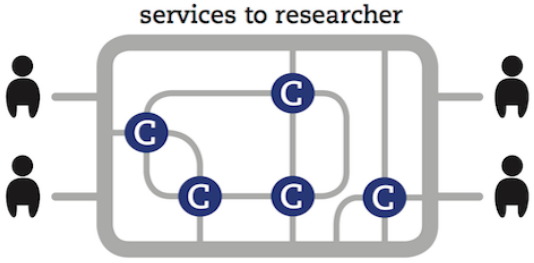
Název projektu: LINDAT/CLARIN
Nositel: prof. RNDr. Jan Hajič, Dr.
URL: https://lindat.mff.cuni.cz/cs/

CLARIN

CLARIN je zkratka pro Common Language Resources and Technology Infrastructure.
CLARIN je založen na distribuované síti center, která se starají o jazykové zdroje a další služby. Většina center je v Evropě a každé má své zaměření. V rámci jedné země se centra spojují do národních konsorcií (national consortium). Například v České republice je národním konsorciem LINDAT-CLARIN a v Polsku je národním konsorciem CLARIN PL.
Existuje několik typů center v CLARIN a každé si musí osvojit základní principy: nezávislost vnitřní organizace, jednoznačnost, konzistenci a zodpovědnost ve službách a dodržování protokolů interakce.
Co se týče podrobností o jednotlivých centrech a o jejich dělení, narazila jsem na rozporuplné informace. Jedním zdrojem mi byly samotné webové stránky www.clarin.eu a druhým zdrojem mi bylo pdf Short Guide CLARIN. Zmíněné pdf má datum Únor 2009, což je téměř sedm let zpátky. Předpokládám, že webové stránky budou obsahovat aktuálnější informaci, a proto se budu držet jejich dělení.
Existuje několik typů center. Velice důležitá jsou technická centra, konkrétně CLARIN B-Centres, někdy nazývaná Service providing centres. Tato centra se většinou nachází na univerzitě nebo na jiné akademické půdě a nabízí vědecké komunitě spolehlivý přístup ke zdrojů, službám a znalostem. Aby se instituce mohla stát CLARIN B-Centre, musí splnit přísná kritéria. V potaz se berou nejen technické možnosti instituce, ale také se posuzuje instituce jako taková. Tato kritéria posuzuje Center Assessment Committee.
Center Assessment Committee – rada, jejíž hlavní zodpovědností je zhodnotit úroveň a kvalitu CLARIN center typu A, B a E. Odehrávají se zde hlavně technické práce: psaní specifikací, plánování vývoje softwaru a organizování kontrol kvality kandidátských center.
V současnosti je přibližně 20 certifikovaných B-center a ještě je několik kandidátů.
Kromě B-center jsou:

- C-Centres (centra poskytující metadata (Metadata Providing Centres), jejich metadata jsou integrována s CLARIN)
- K- & L-Centres (obecná i lokální centra vědomostí (Knowledge Centre))
- T-Centres (centra důvěry (Trust Centres), poskytují přístup ke chráněným zdrojům pomocí Service Provider Federation)
- E-Centres (externí centra nabízející centrální služby, aniž by byly součástí nějakého národního konsorcia).

Instituce v zemích, které nejsou členy CLARIN ERIC (European Research Infrastructure Consortium), se mohou stát CLARIN centry typu C, K a L. Některé instituce také mohou mít více typů. Například CLARINO Bergen Center (což je norské centrum) je certifikováno jako centrum B, K.
LINDAT/ CLARIN je certifikován jako B-centrum a patří pod národní konsorcium LINDAT-CLARIN.



Nákres sítě CLARIN

Projekt LINDAT/ CLARIN

Projekt LINDAT/CLARIN se zasloužil o vytvoření stejnojmenného centra v České republice a v současné době se snaží o jeho udržení. Je zaměřen na tvorbu anotovaných dat pro NLP (Natural language processing) a jeho hlavním cílem je volné sdílení jazykových dat a pokročilých technologií mezi institucemi a mezi jednotlivci ve vědě a výzkumu. Podporuje vznik aplikací a nástrojů, které jsou z hlediska lingvistiky praktické a nadějné. Kromě toho podporuje vývoj dat pro počítačovou lingvistiku.
Projekt vznikl v roce 2010 a od té doby byl již jednou prodloužen. V současné době má platnost do roku 2019.
Na projektu spolupracují Univerzita Karlova v Praze, Masarykova univerzita, Akademie věd České republiky (konkrétně Ústav pro jazyk český AV ČR) a Západočeská univerzita v Plzni. Ústředním centrem LINDAT/CLARIN je Matematicko-fyzikální fakulta Univerzity Karlovy, která má v projektu největší rozsah moci a působí tam velmi významná osoba projektu – prof. RNDr. Jan Hajič, Dr. Ten řeší vazby k samotnému CLARINu.
Na Masarykově univerzitě je z pohledu projektu hlavní osobou doc. PhDr. Karel Pala, CSc. Dalšími lidmi, kteří jsou z Masarykovy univerzity a pracují na projektu, jsou doc. RNDr. Aleš Horák, Ph.D. a doc. Mgr. Pavel Rychlý, Ph.D.
Všechny zmíněné organizace pracují zvlášť, ale někdy mohou spojit své síly, což platí v případě Internetové jazykové příručky.

Jakým oblastem se projekt věnuje?

Oblast anotace dat:

- cílem projektu je pořídit jazyková data v dostatečném rozsahu pro praktickou aplikaci statistického modelování jazyka jako nutnou podmínku pro aplikaci těchto modelů v praxi

Oblast distribuce dat:

- cílem je poskytovat službu repozitáře pro úschovu, licencování a poskytování dat v rámci celoevropské sítě CLARIN a META-SHARE (což je součást projektu META-NET)

Oblast technologická a oblast lidských zdrojů:

- cílem je vybudování know-how v oblasti sběru, úschovy, tvorby a distribuce dat, které bude možno poskytovat i externím subjektům. Přitom je potřeba vyškolit jazykové odborníky i odborníky z oblasti technologií (informatika, statistika, matematické modelování). Takto vyškolení odborníci potom budou schopni efektivně pracovat v tomto mezioborovém projektu.
- dalším cílem je vychovat příští vědeckou generaci, která bude umět s jazykovými daty pracovat, správně je analyzovat a používat v národním i mezinárodním kontextu. Tato generace by také měla být schopna spolupracovat v rámci i mimo EU na budoucích projektech, které budou využívat moderní jazykové technologie.

Repozitář

Repozitář je knihovna pro ukládání a popisování digitálních objektů. Tyto objekty mohou být aplikace, publikace nebo datové modely.
Jedním z hlavních cílů CLARIN je zajistit, aby jazykové digitální zdroje byly poskytnuty široké komunitě na dlouhou dobu. To zajistí tak, že založí datové repozitáře v těch centrech, ve kterých budou uloženy digitální soubory a přiložená metadata.
Co se týče reference, tyto repozitáře přidají ke zdrojům trvalé identifikátory, takže určitý datový soubor může být jednoduše odcitován.
Jestliže chceme vědět, co vše v repozitáři můžeme najít, rozklikneme možnost rozšířené hledání, které nám nabídne mimo jiné možnost zúžit hledání podle určitého parametru. Lze hledat podle autora, klíčového slova, licence, jazyka, typu, komunity a podle toho, zda texty obsahují soubory.

Přidání příspěvku do repozitáře LINDAT/CLARIN:
Uživatel, který chce s repozitářem pracovat, se musí nejprve přihlásit. Stránka nám nabídne instituce, skrz které se můžeme přihlásit. Z univerzit máme na výběr Masarykovu univerzitu, Karlovu univerzitu, Ostravskou univerzitu, Vysoké učení technické... Našli bychom zde nabídku i zahraničních univerzit – německých i anglických; a různé instituce, jako fyziologický ústav AV ČR, biologické centrum AV ČR a Fakultní nemocnici Brno. Z knihoven bychom tu našli Moravskou zemskou knihovnu, městskou knihovnu Kutné Hory, městskou knihovnu České Třebové a vědeckou knihovnu v Olomouci.
Když vložíme příspěvek do repozitáře, zařadí se do kolekce vložených příspěvků a není zatím veřejný. Teprve až ho schválí redaktor, příspěvek se zveřejní a má otevřenou neboli veřejnou licenční kategorii. Jestliže nám při vkládání dat do repozitáře z nějakého důvodu nevyhovuje otevřená licence, licenční kategorii je možné změnit na akademickou nebo na omezenou. Tyto licence se nevztahují na metadata položek; ta jsou vždy veřejná a otevřeně dostupná.
Redaktor při schvalování ověřuje, zda příspěvek splňuje požadovanou kvalitu a úplnost metadat, konzistenci souborů s daty a zda neporušuje práva k duševnímu vlastnictví. Pokud podle redaktora něco chybí / je špatně, příspěvek může autorovi vrátit i s komentářem, co je potřeba změnit.
Publikovaný příspěvek jde upravit nebo dokonce vymazat a může o to požádat kdokoliv, nejen autor příspěvku. Každá žádost se posuzuje samostatně. V případě úpravy, kdyby by šlo o velký zásah do původního textu, je vkladatel příspěvku většinou vyzván, aby předložil novou verzi celého textu.
LINDAT/CLARIN má mimo jiné pravidla pro uchovávání dat. Zavazuje se k dlouhodobé péči o data a nástroje uložené v repozitáři a stará se o to, aby byly k dispozici i v budoucnu. To má zajistit tzv. Datová pečeť (Data Seal od Approval). Snaží se také používat aktuálně nejlepší osvědčené postupy v oblasti uchovávání digitálních záznamů.
Repozitář disponuje obnovovanou certifikací CLARIN ERIC. Ta potvrzuje, že LINDAT/CLARIN používá kompatibilní standardy, že je zajištěna ochrana práv duševního vlastnictví a osobních údajů, že má repozitář vysokou dostupnost a že poskytování služeb je na úrovni předepsané CLARIN ERIC.

Aplikace a nástroje vyvinuté díky projektu

- CzEngVallex – česko-anglický valenční slovník
- Česílko – systém pro strojový překlad úzce příbuzných jazyků
- Český morfologický analyzátor a tagger
- Dialogy.Org
- ElixirFM
- EngVallex – anglický valenční slovník
- EVALD 1.0
- EVALD 1.0 pro cizince
- Internetová jazyková příručka
- Keyword Extractor
- KonText
- Korektor
- MorphoDiTa: morfologický slovník a tagger
- Moses
- NameTag
- Parsito
- PDT-Vallex – český valenční slovník s odkazy do treebanků
- PML-Tree Query – vyhledávací nástroj pro všechny druhy lingvisticky anotovaných stromových korpusů
- Treex::Web – vysoce modulární NLP framework online
- UDPipe

Tyto aplikace a nástroje nejsou určeny jen úzké společnosti (např. vyučujícím), ale všem zájemcům. Nicméně jsou zde jistá omezení. Některé dokumenty a nástroje si lze prohlížet a stáhnout, aniž by bylo potřeba se přihlašovat na stránce projektu (www.lindat.mff.cuni.cz), přesto pro přístup k ostatním dokumentům a nástrojům (např. pokud chceme použít repozitář) musíme být přihlášeni. Typickou aplikací, ke které má přístup kdokoliv, je Internetová jazyková příručka.

Internetová jazyková příručka

Díky projektu LINDAT/ CLARIN a projektu Jazyková poradna na internetu vznikla tato příručka. Našli bychom ji na stránkách www.prirucka.ujc.cas.cz. Tato příručka se skládá z výkladové a ze slovníkové části. Slovníková část vychází hlavně z Pravidel českého pravopisu a ze Slovníku spisovné češtiny pro školu a veřejnost. Výkladovou část přesně a jasně popisuje samotná příručka: „*Do výkladové části byly zařazeny především ty jevy, na které se uživatelé češtiny v jazykové poradně ptají opakovaně. Nejde tedy o soustavný, komplexní popis současného systému češtiny."* [1] Na co se dá příručka použít? Tato příručka se vám může hodit kdykoliv, když píšete nějaký text a nejste si jisti správnou formou slova nebo jeho pravopisem. Při zadání určitého slova vám příručka ukáže jeho flexi, význam i příklad ve větě.

[1] *O internetové jazykové příručce* [online]. [cit. 2016-12-01]. Dostupné z: http://prirucka.ujc.cas.cz/?id=_about

Závěr

Cílem projektu je realizovat to, o co se pokouší samotný CLARIN. To je především péče o jazykové zdroje. Projekt k tomu přidává svůj cíl, což je volné sdílení jazykových dat a pokročilých technologií převážně v České republice.
Oproti jiným projektům se liší v tom, že jeho cílem není konečný bod, ke kterému se chce dostat. Projekt sice je časově omezený, ale jen z pohledu fondů a administrativy. Cílem projektu je rozvíjet centrum, sbírat další data, vytvářet další aplikace a nástroje, které pomohou v rozvoji. Jak se vyvíjí jazyk a informatika, bude se vyvíjet i tento projekt, který tyto dva směry kombinuje.

Zdroje

https://www.clarin.eu/content/repositories
https://www.clarin.eu/content/clarin-in-a-nutshell
https://www.clarin.eu/content/clarin-centres
https://www.clarin.eu/content/certified-centres
https://www.clarin.eu/content/clarin-technology-introduction
https://www.clarin.eu/sites/default/files/centres-CLARIN-ShortGuide.pdf
http://prirucka.ujc.cas.cz/?id=_about
https://lindat.mff.cuni.cz/repository/xmlui/page/about
https://lindat.mff.cuni.cz/cs/o-lindat-clarin
http://www.ujc.cas.cz/veda-vyzkum/vyzkum/grantove-projekty-ukoncene/LINDAT-CLARIN.html
https://www.muni.cz/vyzkum/projekty/13063
https://www.muni.cz/vyzkum/projekty/34244
https://ufal.mff.cuni.cz/grants/lindatclarin
https://www.korpus.cz/kontext_transition.php
https://wiki.korpus.cz/doku.php/cnk:uvod
https://is.muni.cz/auth/publication/935752/cs?fakulta=1421;obdobi=6644;studium=688599;lang=cs

Metadata

<dc:title>Projekt LINDAT/Clarin</dc:title>
<dc:description>V tomto dokumentu popisuji samotný CLARIN a jeho organizaci, poté se rozepisuji o projektu LINDAT/CLARIN a zaměřuji se na jeho repozitář a jak s ním pracovat. Ke konci zmíním aplikaci, na které se projekt podílel. </dc:description>
<dc:date>2016-12-01</dc:date>
<dc:creator>Barbora Obluková</dc:creator>
<dc:type>Text</dc:type>
<dc:language>cz</dc:language>