Příloha č. 1 – Návrh Projektu velké infrastruktury schválený vládou

Projekt LINDAT/CLARIN
Vybudování a provoz českého uzlu pan-evropské infrastruktury pro výzkum

Uchazeč

Uchazečem o podporu je Univerzita Karlova v Praze; s prostředky bude hospodařit Ústav formální a aplikované lingvistiky MFF UK. Univerzita Karlova v Praze zajistí dvoustrannými smlouvami převod prostředků nezbytných pro provoz centrálního uzlu Clarin a pro zajištění národních úkolů na pracoviště, která se na projektu budou podílet podle schváleného rozpočtu. Univerzita Karlova v Praze rovněž zajistí (nebude-li do té doby rozhodnuto jinak) převod plánovaných prostředků do budoucí asociace Clarin-ERIC.

Osoba odpovědná za organizační, technické a personální zajištění projektu (řešitel-koordinátor): Prof. RNDr. Jan Hajič, Dr. (hajic@ufal.mff.cuni.cz), ředitel Ústavu formální a aplikované lingvistiky MFF UK, Malostranské nám. 25, 11800 Praha 1 (tel. +420 607 209 212).

Ústav formální a aplikované lingvistiky MFF UK (http://ufal.mff.cuni.cz) je předním národním pracovištěm v oblasti počítačového zpracování přirozeného jazyka, které bylo rozvíjeno v rámci programu výzkumných center i výzkumných záměrů a které má výrazné mezinárodní postavení v rámci již existujících celoevropských projektů. Plánovaný projekt výzkumné infrastruktury přímo navazuje na projekt CLARIN (FP7-RI-2122230), jehož je pracoviště uchazeče národním koordinátorem.

- ***Popis výzkumné infrastruktury***

1.1. Koncepce

Projekt LINDAT-Clarin je koncipován jako český „uzel“ mezinárodní sítě Clarin (Common Language Resources and Technology Infrastructure, FP7-RI-2122230, zatím 2008-2010) a je rozšířen i na projekt “T4ME Net” (Technologies for the Multilingual European Information Society, NoE, 2011-2014, FP7-ICT-4-249119) pro volné sdílení jazykových dat a základních technologií mezi institucemi a jednotlivci ve vědě a výzkumu. Projekt Clarin je přitom cílen na humanitní vědy, projekt T4ME je širší s dosahem především na jazykové technologie a jejich aplikace.

Centrum LINDAT-Clarin bude tedy v českém jazykovém prostředí propojovat tyto dvě oblasti a bude zaměřeno na sběr jazykových dat a především jejich anotaci (tj. formální manuální, poloautomatickou a automatickou jazykovou analýzu). Sběr a anotace budou probíhat v takovém rozsahu, kvalitě a technologické přípravě (specifikace, schémata, formáty), která bude přímo využitelná jak v humanitní oblasti (jazykovědný a mezioborový výzkum, kde přirozený jazyk hraje podstatnou roli), tak i pro výzkum a vývoj v oblasti jazykových technologií za použití moderních statistických a hybridních metod.

Náplň práce centra a jeho výsledky se dotýkají řady oborů – v humanitních vědách to jsou jazykověda obecná a pro jazykověda zabývající se konkrétními jazyky, zejména češtinou, dále překladatelství, lexikografie, sociolingvistika, částečně i obory příbuzné (psychologie, sociologie, knihovnictví, neurovědy, cognitive science) s významným přesahem do informatiky (computer science, computational linguistics), matematiky (statistika a pravděpodobnost), a elektrotechniky (zpracování akustického signálu).

Z hlediska národních priorit aplikovaného výzkumu (dle dokumentu Národní politika výzkumu, vývoje a inovací České republiky na léta 2009 – 2015, Hlava VI) navrhované centrum spadá do priority 6 (Informační společnost) a 8 (Priority rozvoje české společnosti).

1.2. Současný stav

Jazykové zdroje a technologie pro jejich zpracování v jednotlivých evropských zemích (stejně jako v USA a Asii) již existují, ale současné centralizované distribuční agentury (Linguistic Data Consortium v USA a European Language Resources Association v Evropě) nevyhovují současným požadavkům na jednoduchý, nebyrokratický a zejména volný přístup k jazykovým datům pro většinu vědeckých, výzkumných a vývojových komunit.

Dochází tak k fragmentované, nekoordinované distribuci dat se všemi negativními důsledky (nekompatibilita formátů a tím obtížná použitelnost software pro jejich zpracování, velké množství různých licenčních podmínek, často i nemožnost přistupovat přímo k datům samotným a nutnost používat mnoho různých vyhledávacích systémů apod.). V ČR se data sbírají a anotují především na čtyřech pracovištích, která by se měla stát podílet na činnosti centra LINDAT-Clarin: UK v Praze, MU v Brně a ZČU v Plzni spolu s ÚJČ AV ČR v Praze. Tato pracoviště v současné době tvoří Centrum komputační lingvistiky (CKL – projekt MŠMT LC536), které je ovšem vědeckým pracovištěm, jazykové zdroje tvoří spíše okrajově a po roce 2010 už nebude existovat.

Projekt 7. RP (ESFRI) Clarin je zaměřen na odstranění těchto nedostatků v evropském rámci a ve prospěch humanitních a sociálních věd, zejména výzkumu ve všech oblastech lingvistiky. Projekt T4ME Net (zahajuje v březnu 2010) je zaměřen obdobně, ale jeho primárním cílem je sloužit vědecké komunitě mezioborově (lingvistika, informatika, statistika), a to zejména v jazykových technologiích, tzv. počítačové lingvistice (computational linguistics), která už má následné dopady i v aplikační oblasti.

1.3. Věcná struktura projektu

V souladu s koncepcí projektu a provázaností s projekty Clarin a T4ME bude projekt strukturován takto:

1. Hlavní část: uzel distribuované evropské sítě Clarin
2. Národní část: české a multilingvální jazykové zdroje: sběr a tvorba jazykových korpusů a databází
3. Koordinační část: koordinace v oblasti právní, technologické, vzdělávací a vnějších vztahů

Na základě tohoto věcného členění bude vytvořena i organizační struktura projektu (viz kap. □) a členěn rozpočet projektu (kap. □).

1.3.1. Uzel distribuované evropské sítě Clarin

Jak už vyplývá z povahy projektu a jeho napojení především na ESFRI a projekt Clarin, tato část projektu je zásadní pro úschovu, sdílení a poskytování dat a služeb, což je jádrem projektu Clarin. Centrum LINDAT-Clarin bude ve své funkci distribuovaného uzlu zajišťovat technologické zázemí pro uzel typu „A“ (podle definice projektu Clarin v přípravné fázi), tj. nejvyšší model takového distribučního uzlu. Uzel typu „A“ zajišťuje úschovu jazykových dat (včetně přijímání nových dat a jejich začlenění do systému včetně přidělení unikátních persistentních identifikátorů), autorizovaný přístup s využitím celoevropské federace identit, přidělování identit pro uživatele systému, a webové služby lokální (plně) a distribuované (zprostředkovaně) pro zpracování jazykových dat, přístup k nim a jejich

poskytnutí dalším uzlům Clarin. Tato část projektu bude zajištěna navrhovatelem prostřednictvím MFF UK v Praze.

Rozhodování o směrování této části (na rozdíl od části národní, viz 1.3.2) bude probíhat ve struktuře projektu Clarin. Na tomto rozhodování se bude česká strana podílet podle pravidel projektu Clarin, a to rovněž prostřednictvím MFF UK, která byla v přípravné fázi pověřena koordinováním národních aktivit Clarin v ČR.

1.3.2. České jazykové zdroje: sběr a tvorba jazykových korpusů a databází

Zatímco jazyková data, která má distribuovaný uzel sítě Clarin uschovávat a sdílet, existují (a stále se tvoří) na řadě pracovišť v zahraničí, v českých podmínkách j nutno pro sběr a tvorbu jazykových korpusů stále aktivně podporovat. Z pochopitelných důvodů v zahraničí vznikají jazyková data v dané oblasti lokální (v Německu jde o němčinu, a Holandsku o holandštinu, v Belgii francouzštinu a vlámštinu atd.), a česká jazyková data všech druhů je tedy nutno zajistit v ČR.

Zatímco sběrem dat se zabývají i jiná pracoviště v ČR a činnost LINDAT-Clarin zde bude pouze doplňovat chybějící oblasti, především mluvená data, paralelní data (s jinými jazyky) ve velkém rozsahu, a data kombinovaná (definice viz níže), při tvorbě anotovaných jazykových korpusů, které jsou základem pro další výzkum a vývoj jak v humanitních, tak v technologických a aplikačních oblastech klíčové, bude centrum LINDAT-Clarin v ČR jedinečné.

Sběrem jazykových dat se rozumí:

- získávání psaných (případně mluvených) jazykových dat z veřejně dostupných zdrojů (internet, otevřené zdroje); tato data mohou být i v kombinaci s dalšími modalitami, např. videem, znakovou řečí zaznamenanou symbolicky apod.
- získávání psaných nebo mluvených jazykových dat (s případnou vizuální složkou) od nakladatelů a majitelů těchto dat na základě smluv
- nahrávky řečových dat v předem určeném prostředí (laboratorní, aplikačně přizpůsobené, v terénu)
- elektronická data ze senzorů při expozici subjektů přirozenému jazyku v libovolné podobě (video, haptika, elektrické nebo magnetické snímání reakcí apod.)

Tvorbou jazykových korpusů a databází se rozumí takové zpracování sebraných dat, které je obohatí pro použití v humanitních i technických vědách v těchto směrech:

- tzv. „čištění“, unifikace a normalizace dat do standardních datových formátů, které lze využít pro další zpracování nebo distribuci
- párování (alignment) pro paralelní jazyková data (psaná, mluvená, vizuální nebo jiná), s cílem zajistit explicitní vztahy mezi jazyky nebo různými modalitami dat navzájem, a to pomocí manuálních nebo automatických metod, jejichž výsledkem jsou synchronizační značky obsahového nebo časového rázu
- extrakce jazykových databází z jazykových dat; databázemi se rozumí především slovníky (fonetické, morfologické, syntaktické a sémantické, příp. s vazbou na ontologie)
- anotace jazykových korpusů; anotací se rozumí manuální nebo (polo)automatická analýza jazykových dat (s výstupem v elektronické formě), která je nutná pro další využití dat v humanitní i technologické oblasti, a to

podle světových standardů a nových postupů s ohledem na vlastnosti českého jazyka.

Cílem této části projektu je připravit jazyková data (jazykové korpusy a databáze, podle situace i s anotací) k publikaci v nejkratším možném termínu prostřednictvím uzlu Clarin (viz 1.3.1) a dalších distribučních kanálů tak, aby byla k dispozici co nejširší veřejnosti pomocí jednoduché licenční a autentizační politiky bez technologických nebo právních překážek, a to od institucionální úrovně až na úroveň jednotlivého výzkumníka nebo studenta.

Na této části projektu se budou podílet všechna čtyři česká partnerská pracoviště podle svého zaměření a specializace. V jednotlivých případech se očekává společný postup při tvorbě zvláště pracovně nebo technologicky náročných jazykových korpusů a databází.

1.3.3. Koordinace v oblasti právní, technologické, vzdělávací a vnějších vztahů

Sběr, tvorba, anotace a distribuce jazykových dat vyžaduje podle dosavadních zkušeností řadu navazujících činností, které je vhodné soustředit mimo vlastní jazykovědnou nebo technologickou práci.

Pro sběr dat na jedné a pro distribuci dat na druhé straně je třeba důsledně dbát na vhodnou aktivní i pasivní licenční politiku, s cílem co nejvíce zjednodušit přístup k vytvořeným jazykovým datům koncovým uživatelům a poskytnout jim rovněž nejvyšší reálně možnou právní jistotu, že data, která si pomocí distribuční sítě Clarin pořídí, nemají z hlediska vědeckého užití žádná omezení. Situace je nyní v EU jednodušší než před několika lety díky změnám v zákonech týkajících se autorských práv (v ČR zákon 121/200 Sb. v aktuálním znění), nicméně stále není pro vědecké využití jazykových dat ideální a je třeba ji dostatečně zajistit.

Technologickým cílem je sjednotit softwarové nástroje pro sběr, čištění, párování i (manuální) anotaci dat tak, aby byla zajištěna interoperabilita minimálně na úrovni datových formátů (XML a jeho varianty a restrikce pro jazykovou oblast). Stejného cíle je nutno dosáhnout pro uchovávání a distribuci dat a webové služby v rámci sítě Clarin (na základě rozhodnutí koordinačního centra Clarin).

Nedílnou součástí centra LINDAT-Clarin je i výchova odborníků v oblasti jazykové, jazykově-technologické a organizační ve všech oblastech práce centra, a to na úrovni navazujících magisterských programů, na úrovni doktorandské a při dalším vzdělávání odborníků z vysokých škol, Akademie věd a aplikované praxe. Cílem je vzdělaná skupina pracovníků, kteří budou dále efektivně pracovat v centru LINDAT-Clarin, ve výzkumu humanitním i technologickém a praxi s jazykovými daty a technologiemi.

Centrum LINDAT-Clarin bude rovněž široce propagovat využití jazykových technologií a ukazovat jejich možnosti pro rozvoj znalostní společnosti pořádáním seminářů určených pro jednotlivé cílové skupiny (od vědeckých až po manažerské), podílem na organizaci celoevropských akcí ve výzkumně-jazykové oblasti (zejména ve spolupráci z projektem T4ME Net) a účastí pracovníků centra na obdobných akcích.

Tyto aktivity budou koordinovány jak s projektem Clarin (resp. jeho centrálou), a projektem T4ME Net na „evropské“ straně a rovněž v rámci národní účasti (ZČU, MU, ÚJČ AV ČR) prostřednictvím zástupců MFF UK v těchto projektech a vnitřní organizační struktury projektu LINDAT-Clarin (viz kap. □).

- ***Přínos pro výzkum a vývoj v Evropě a v České republice***

Centrum bude pokračovat v mezinárodní spolupráci při tvorbě a distribuci dat. V rámci projektu Clarin již nyní spolupracuje 171 institucí z 27 členských a 5 asociovaných zemí EU (na úrovni univerzit, akademických ústavů atd.). V projektu T4ME bude při jeho zahájení 17 partnerů, kteří jsou financování EU, a přepokládá se vytvoření konsorcia uživatelů s cca 200-300 institucionálními členy z EU a asociovaných zemí. V obou projektech je partnerem z ČR Univerzita Karlova (prostřednictvím ÚFAL MFF UK).

Předpokládaní partneři projektu byli a jsou rovněž zapojeni do dalších projektů v ČR i EU, ve kterých (dosud nekoordinovaně) vznikají jazyková data a různými způsoby se distribuují, například EuromatrixPlus a Faust (ÚFAL MFF UK - jazyková data pro automatický překlad), Companions (ÚFAL MFF UK, KKY FAV ZČU – řečová anotovaná data pro vývoj dialogových systémů), KYOTO (MU Brno, lexikální data). ÚFAL MFF UK je dále zapojen do existujícího menšího projektu NoE FlareNet, kde vznikají standardy a specifikace společného základu pro jazykové zdroje.

ÚFAL MFF UK spolupracuje dále s USA a některými asijskými pracovišti na projektech anotace jazykových dat: projekt NSF PIRE (anotace pro pokročilé porozumění mluvenému jazyku, strojový překlad: Johns Hopkins University, MD, USA, Brown University, RI, USA), projekty anotace syntaxe a valence (University of Colorado, CO, Brandeis University, MA, IIIT Hyderabad, Indie), projekt anotace jazykových dat pro výzkum diskursu (University of Pennsylvania, PA, USA). ÚFAL MFF UK rovněž spravuje část dat a přístupový bod k 116 tisícům hodin nahrávek vzpomínek přeživších obětí holocaustu („Centrum Malach“, smlouva s University of Southern California, CA, USA), které jsou významným a velikostí dosud nepřekonaným zdrojem mluvených dat v mnoha jazycích.

ÚFAL MFF UK má spolu s Katedrou kybernetiky FAV ZČU v Plzni rovněž dlouholeté zkušenosti s vydáváním a „klasickou“ distribucí jazykových dat přes střediska LDC (USA) a ELRA/ELDA (Francie/EU).

Význam navrhované infrastruktury lze shrnout do těchto bodů:

1. vytvoří národní referenční zdroj jazykových dat pro veřejně dostupné, snadné a právně bezproblémové vědecké, pedagogické, výzkumné i aplikačně-vývojové použití
2. umožní široký přístup odborné komunitě i uživatelům k expertíze, která je ověřená
3. umožní široký přístup odborné komunitě i uživatelům k již vyvinutým počítačovým nástrojům a technologiím i službám
4. umožní široký přístup odborné komunitě i uživatelům nejen k jednojazyčným (českým) zdrojům, ale ke zdrojům multilingválním a k odpovídajícím technologiím
5. bude tvořit významnou „přidanou mezinárodní hodnotu“ k národním, v našem případě českým, iniciativám, umožní propojení center v celoevropském měřítku
6. poskytne významný potenciál k inovacím
7. posílí zájem o národní jazyk jako součást národní kultury a národního dědictví
8. v neposlední řadě významnou měrou přispěje k modernizaci pedagogického procesu (výuka jazyků, jazykové technologie, velká data a jejich zpracování)

Klíčová je přitom otázka, kdo a jak bude vytvořené centrum LINDAT-Clarin využívat; na základě dosavadních zkušeností s využitím jazykových dat předpokládáme, že uživatelé se budou rekrutovat z těchto kategorií:

1. v oblasti vědecké: všechna akademická pracoviště (na univerzitách i v AV ČR) zabývající se vědeckým zpracováním, výukou a počítačovým zpracováním);

předpokládá se využití i ze zahraničí (pracoviště stejného nebo podobného typu) v rámci projektů Clarin a T4ME Net i mimo ně, studenti zapojení do vědecké práce (magisterské a doktorské studium, zahraniční studenti v ČR);
2. v oblasti aplikačního vývoje a inovací: všechna pracoviště a průmyslové organizace zabývající se informačními systémy (knihovnictví, dokumentační střediska, atd.), překladatelskými službami, vyhledáváním dokumentů a informací, terminologickými databázemi, podpůrnými nástroji pro tvorbu textů (korektory), dokumentací v oblasti historie, průzkumy sociologickými, psychologickými a aplikacemi podobného charakteru, v nichž se pracuje s texty a textovými záznamy; i zde je vysoká pravděpodobnost využití ze zahraničí (lokalizační projekty zahraničních firem, firem vyvíjejících nástroje pro podporu automatického překladu apod. – viz seznam výše);
3. v oblasti výuky: využití ve školství všech stupňů v rámci jazykové výuky, výuky IVT a případně dalších předmětů.

- ***Cíle projektu***

3.1. Úvod

Projekt EU/ESFRI Clarin, stejně jako budoucí T4ME Net a ve stejném duchu i navrhované centrum LINDAT-Clarin má překážky volného přístupu k jazykovým datům postupně odstranit a umožnit distribuované, nicméně jednotné poskytování jazykových dat a souvisejících technologií. Ambice Clarinu a do značné míry také T4ME (s výjimkou evaluačních dat pro konkurenční testování nástrojů pro zpracování přirozeného jazyka) jsou zejména ve sjednocení technologií a distribuce – tvorba je ponechána na kompetenci (a finančním zajištění) v jednotlivých státech, neboť se jedná často o národní jazyky. Sběr a anotace (tj. tvorba dat) je tedy nedílnou součástí navrhovaného centra LINDAT-Clarin.

Anotace dat v dostatečném rozsahu je nezbytná k tomu, aby výsledky tohoto výzkumu mohly být aplikovány v praxi (korektory textu, automatický překlad, extrakce informací z textu, porozumění textu, dialogové systémy apod.), neboť vývoj softwarových nástrojů je dnes založen na statistických metodách a ty vyžadují velké množství zejména jazykově interpretovaných (anotovaných) dat. Tato data jsou potřebná i pro to, aby čeština byla zařazena do programu lokalizace textových produktů velkých firem, které tyto metody vesměs využívají.

Existence anotovaných dat v daném jazyce rozhoduje často o zařazení do velkých výzkumných projektů EU (v minulosti UK získala několik takových projektů mj. díky vlastnictví anotovaných korpusů – např. projekty Euromatrix, EuromatrixPlus, Companions - se ZČU, Faust a samozřejmě i projekty Clarin a T4ME Net).

Anotace dat je časově i kapacitně velmi náročná - jedná se především o manuální práci vysoce odborně školených jazykových odborníků, a to studentů doktorského (výjimečně i navazujícího magisterského) studia jazykovědných oborů, zejména mezioborových v kombinaci s informatikou, s vysokým podílem technického zajištění odborníky z oblasti informatiky. Sběr dat je rovněž náročná práce, která vyžaduje odborníky několika profesí, od informatiky přes organizační zajištění, právní zabezpečení, až po výuku a školení uživatelů.

Jednotná distribuce pak zajistí širší publicitu a využití dat a tedy i vložených prostředků.

Nezanedbatelným cílem projektu je vychovat další vědeckou generaci, která bude umět s jazykovými daty pracovat, správně je tvořit a používat v národním i mezinárodním kontextu, a spolupracovat v rámci EU i mimo ni na budoucích projektech využívajících jazykové technologie.

3.2. Dílčí cíle projektu

Dílčí cíle projektu vycházejí z hlavních dvou hlavních cílů projektu popsaných v předchozím odstavci – pořídit jazyková data pro češtinu (a data paralelní) a zpřístupnit je. Mezi kontinuálně pojaté dílčí cíle projektu patří a zajištění výchovy mladé vědecké generace v oblasti jazykových dat a jazykových technologií v humanitní i technické oblasti.

3.2.1. Výstavba distribučního uzlu Clarin

Výstavba distribučního uzlu typu „A“ v metodologii Clarin je základním předpokladem fungování centra LINDAT-Clarin v rámci evropského výzkumného prostoru a v rámci projektu Clarin (a T4ME Net). Tento uzel bude budován postupně v letech 2010 (příprava, dokončení specifikací) a 2011-2013 (konstrukční fáze, soustředění technologií, postupné získávaní dat a zahájení zkušebního provozu). V plném provozu se očekává v roce 2014.

Výstavba bude mít investiční část (ve smyslu pořízení výpočetní a síťové techniky na UK v Praze a na ostatních pracovištích) a část softwarovou při zavedení celosvětově jednotné identifikace dat, jednotné a alespoň celoevropsky platné autentikace uživatelů pomocí federace identit (v ČR: EduID), zavedení dohodnutého (v rámci Clarin) repozitáře pro úschovu vlastních i „cizích“ dat, zavedení jednotného popisu metadat, zavedení jednotného API pro webové služby v oblasti zpracování jazykových dat a zajištění právního rámce (vzorové smlouvy pro otevřené licence pro poskytovatele dat, licence pro uživatele).

3.2.2. Národní infrastruktura pro sběr a tvorbu jazykových dat

Jednotlivá pracoviště navrhovaného centra LINDAT-Clarin spolu již v oblasti sběru a anotace dat spolupracují na základě jednotlivých případů (projekt 6. RP „Companions“ a další). V rámci navrhovaného centra LINDAT-Clarin budou vzájemně předány nástroje na budování a anotaci jazykových dat, aby se dosáhlo co nejvyšší efektivity a co nejširší dostupnosti i mimo partnerské instituce.

Nástroje na sběr dat z otevřených zdrojů, na základní zpracování („čištění“ dat, tokenizace, segmentace) a na základní jazykové automatické předzpracování budou vzájemně volně dostupné a v případech, kdy to bude výhodné, budou sjednoceny (pokud se nyní používají podobné, nikoli však identické a všem vyhovující typy nástrojů).

Pro distribuci dat a pro ty služby v oblasti zpracování jazykových dat, kde i v rámci ČR to bude efektivnější, bude využíván uzel Clarin (viz 3.2.1).

3.2.3. Sběr dat

Sběr dat bude prováděn v závislosti na typu dat. Získávání textových českých a textových paralelních dat z otevřených zdrojů (tj. především internetu) je otázkou vhodné technologie, která bude mezi jednotlivými českými pracovišti sdílena.

Řečová data je nutno získat od subjektů-lidí, kteří budou předčítat vhodně vybrané pasáže pro daný konkrétní úkol. Podobně bude probíhat i záznam dialogů a senzorických reakcí lidí na jazykové podněty. Ve všech případech bude vyžádán souhlas subjektu na základě zákona o ochraně osobních dat; povaha získávání dat však nevyžaduje speciální opatření ohledně etických norem.

Součástí sběru dat je i jejich prvotní zpracování a ukládání s jednotnými metadaty a ve formátu vhodném pro další zpracování, zejména manuální transkripci a anotaci (viz 3.2.4).

3.2.4. Anotace dat

Lingvistická anotace jazykových dat, která je klíčová pro zhodnocení dat zejména pro jejich budoucí aplikační využití, je časově i finančně nejnákladnější a nejnáročnější operací při jejich přípravě. Základní metoda anotace se skládá ze čtyř fází:

- výběr dat (z dat sebraných a připravených)
- příprava anotačního nástroje pro přímou anotaci v elektronické formě
- vlastní anotace (včetně organizačního a administrativního zajištění)
- kontrola kvality anotace, příprava dat k uložení a veřejnému zpřístupnění

Rozdělení těchto čtyř oblastí z hlediska objemu práce je takové, že vlastní anotace je zastoupena cca polovinou. Kontrola kvality rovněž záleží na komplexnosti anotace – jedná se většinou o porovnání dvojité nebo trojité anotace a opravu neshody mezi anotátory v datech; lze aplikovat i různé automatické nebo poloautomatické metody kontroly formátu i obsahu anotace.

3.2.5. Příprava dat k publikaci, vlastní zpřístupnění a distribuce dat

Cílem projektu LINDAT-Clarin i samotných evropských projektů Clarin a T4ME Net je co nejvíce zpřístupnit (a to co nejjednodušeji) jazyková data a metody jejich zpracování výzkumné komunitě v České republice i Evropě. Data je tedy třeba publikovat – v novém konceptu Clarin to znamená, že data budou po jejich vytvoření a anotaci co nejdříve zpřístupněna v síti uzlů Clarin.

Pro tento účel je třeba vytvořená data popsat z hlediska budoucích uživatelů (anotační příručky, jejich překlad do relevantních jazyků), Data je pak nutno odpovídajícím způsobem uložit v repozitářích sítě Clarin.

3.2.6. Začlenění softwarových nástrojů do poskytovaných webových služeb

Pro přístup k jazykovým datům je obvykle třeba vytvořit software včetně vyhledávacích a statistických modulů (pokud se data nezpřístupňují prostým poskytnutím v kopii) a realizovat jej jako webovou službu v síti Clarin. Vzhledem k různorodosti jednotlivých datových souborů a databází je často nutno vytvářet specializovaný software, který umožní uživatelům např. i z humanitních oblastí plně a efektivně daná data využívat v jejich práci.

3.2.7. Výchova vědeckých pracovníků v oblasti jazykových dat

Nedílnou součástí centra LINDAT-Clarin je výchova vědeckých pracovníků pro práci s jazykovými daty, a to jak z hlediska lingvistického (teoretického i aplikovaného), tak z hlediska informatického. Všechna čtyři pracoviště disponují odborníky v těchto oblastech; budou využívány (v případě zájmu studentů, zapojených do prací v centru) i akreditované magisterské a doktorandské programy, neboť jazyková data i způsob jejich sběru a anotace jsou vhodným materiálem pro diplomové a doktorské práce.

V rámci možnost bude centrum LINDAT-Clarin poskytovat toto vzdělání pro studenty EU i mimo EU podle platných předpisů a regulí, s využitím vzdělávacích programů EU i přímo.

3.2.8. Organizace vědeckých seminářů a konferencí v oblasti jazykových dat

Pro odborníky (ze všech relevantních oblastí) bude centrum LINDAT-Clarin pořádat semináře a případně se podílet i na organizaci konferencí v oboru řeči a zpracování přirozeného jazyka, v čemž mají spolupracující organizace již dlouholetou tradici (konference Text, Speech and Dialogue v ČR, konference a workshopy Association for Computational Linguistics).

3.2.9. Organizace seminářů a školení pro uživatele služeb

Uživatelé služeb uzlu sítě Clarin budou školeni v metodách používání jazykových dat zpřístupněných pomocí této sítě, a to jak v ČR, tak ve spolupráci s ústředím Clarin i v zahraničí.

3.3. Časový harmonogram pro roky 2010-2015

Obecně lze harmonogram rozdělit na tři části:

- 2010 příprava centra
- 2011-2013 konstrukční fáze, ve spolupráci s Clarin (Clarin-ERIC, bude-li založeno)
- 2014-2015(-2020 a dále) provoz (operační fáze) / uzel Clarin-ERIC.

V současné době jsou jazykové zdroje a příslušné programové nástroje uchovávány na pracovištích, kde vznikají, a také jimi zpřístupněny nebo distribuovány (web, fyzická média). Výjimečně jsou distribuovány dvěma existujícími mezinárodními středisky (LDC - USA, ELRA - Francie). Tvorba zdrojů se provádí zejména na pracovištích UK a ZČU, ÚFAL MFF UK má vybudované menší úložiště (několik TB diskového prostoru) a výpočetní kapacitu cca 200 CPU jednotek na základní zpracování jazykových dat.

1. Vybudování úložného a distribučního centra (Centrum „A“ v intencích projektu Clarin a podobné centrum v rámci projektu T4ME) se předpokládá ve dvou etapách: (i) vybudování technické infrastruktury (úložiště, sítě, připojení) do r. 2013 (konstrukční fáze), a poté (ii) její provoz a případné kapacitní rozšiřování (operační fáze, kontinuálně). Zajištění toho distribučního centra je plánováno na ÚFAL MFF UK.
2. Sběr dat, archivace, zpřístupnění: tato aktivita v ČR probíhá na několika pracovištích průběžně (UK, MU, ZČU). Jedná se tedy o převedení těchto aktivit na společnou platformu, začlenění relevantních datových kolekcí pod distribuční schéma centra ad (1) do konce r. 2013. Operační fáze kontinuálně – částečně již paralelně s konstrukční, plně od r. 2014.
3. Specifikace a tvorba anotovaných dat, softwarových nástrojů pro zpracování dat: převedení těchto činností ze současného Centra komputační lingvistiky a částečně i několika „Výzkumných záměrů“ po jejich skončení, organizační zajištění, propojení s distribucí ad (1) do konce r. 2013. Operační fáze (včetně poskytnutí nástrojů a kapacit pro přípravu vlastních dat) od r. 2014 kontinuálně.