Nástin koncepce rozvoje Ústavu formální a aplikované lingvistiky MFF UK v období 2025-2028

Předkládá: Barbora Vidová Hladká, hladka@ufal.mff.cuni.cz
Datum: 14. září 2024

V tomto dokumentu nastiňuji rozvoj Ústavu formální a aplikované lingvistiky MFF UK pro období 2025-2028. Návrh vychází z aktuálního fungování ústavu, které je základem jeho dosavadních úspěchů a uznání jak na národní, tak mezinárodní úrovni. Je sestaven jako popis ústavu a budoucích plánů.

Pracoviště se tradičně soustředí na dva cíle: poskytovat studentům kvalitní vzdělání a rozšiřovat vědecké poznatky. K dosažení těchto cílů nestačí jen otevřená mysl a nadšení pro výzkum a výuku. Potřebná je také efektivní administrativní a technická podpora, která je tím důležitější, čím více lidí se na výuce a výzkumu podílí. Proto se popis pracoviště zaměřuje především na číselné ukazatele jeho aktivit, které ovlivňují rozsah potřebné podpory. Ukazatele jsou extrahovány převážně z fakultních a univerzitních systémů, stejně jako z interních zdrojů ústavu. Tato data, jejichž časový rozsah je dán nastavením jednotlivých systémů, považuji za klíčová pro sledování trendů, stanovení kapacit a identifikaci limitů pracoviště.

Současný chod pracoviště není třeba zásadně měnit, ale je důležité pokračovat na rozvoji několika oblastí. Ty uvádím v budoucích plánech.

Velice děkuji Ing. Stanislavě Gráf za spolupráci při sběru dat. Vzhledem k absenci centrálního datového bodu na úrovni univerzity či fakulty je nutné data dohledávat převážně manuálně, typicky exportem excelových souborů s následnou netriviální konverzí a filtrováním dat.

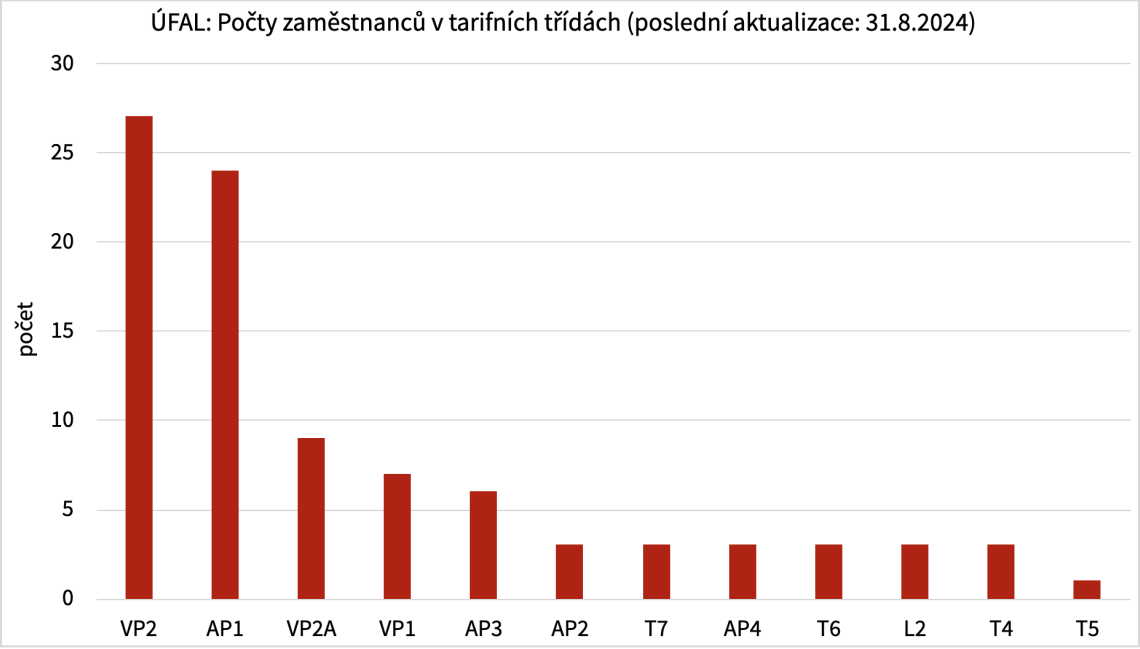
Dále velice děkuji Mgr. Milanu Fučíkovi za detailní rekapitulaci rozvoje výpočetních kapacit ústavu. Jeho přehled poskytuje cenný náhled na to, jakým způsobem se tato infrastruktura na pracovišti postupně vyvíjela a jakým způsobem dnes podporuje výzkumné projekty a výuku. Zvláště v kontextu práce s daty a experimentování s modely strojového učení jsou výpočetní kapacity klíčové.

1. Úvod

Ústav formální a aplikované lingvistiky je součástí informatické sekce Matematicko-fyzikální fakulty (ÚFAL MFF UK). Byl založen v roce 1990 v návaznosti na neformální skupinu počítačové lingvistiky v rámci bývalé Katedry aplikované matematiky. Jeho hlavním zaměřením je výzkum na pomezí informatiky a lingvistiky v oblasti počítačové lingvistiky a jazykových technologií. Spolu s tím zajišťuje výuku předmětů souvisejících s těmito výzkumnými oblastmi.

Organizační struktura ÚFAL je plochá – vedení tvoří ředitel a jeho zástupce, týmy se soustřeďují kolem vedoucích konkrétních výzkumných projektů, tajemník se stará o agendu týkající se rozvrhování výuky a administrativní správu pracoviště a projektů a správu IT zajišťují techničtí pracovníci. K 31. srpnu 2024 má ÚFAL 92 zaměstnanců, kteří pracují v budovách na Malé Straně a v Troji.

Operativní činnost ÚFAL zahrnuje správu pracoviště a administraci projektů. Vysoký počet zaměstnanců (aktuálně 92) soustředěných ve dvou vzdálených lokalitách a vysoký počet řešených projektů (aktuálně 46) na sebe vážou i vysoké počty dílčích administrativních agend, jako je např. zpracování cestovních příkazů (viz Příloha B), přijetí zahraničních hostů (viz Příloha C), nákupy (viz Příloha D). Takové podmínky vyžadují systematický přístup k vedení ústavu ve třech klíčových oblastech: formalizace agendy, optimalizace procesů a sdílení informací. Tato trojkombinace je zásadní pro to, aby zaměstnanci mohli svou energii a čas maximálně věnovat svým vědeckým, administrativním či technickým úkolům, a nemuseli se zbytečně zdržovat neefektivními agendami. Od září 2021 se na pracovišti intenzivně rozvíjí snaha tyto oblasti zlepšovat, přičemž součástí tohoto úsilí je i digitalizace procesů.



2. Personální složení

K 31. srpnu 2024 má ÚFAL 92 zaměstnanců. Téměř třetina zaměstnanců spadá do tarifní třídy VP2, což reflektuje výzkumné zaměření pracoviště. Další významnou skupinu zaměstnanců tvoří doktorandi zařazení do tarifní třídy AP1, kteří jsou zaměstnáni v rámci Programu zaměstnávání doktorandů informatické sekce MFF UK.[1]

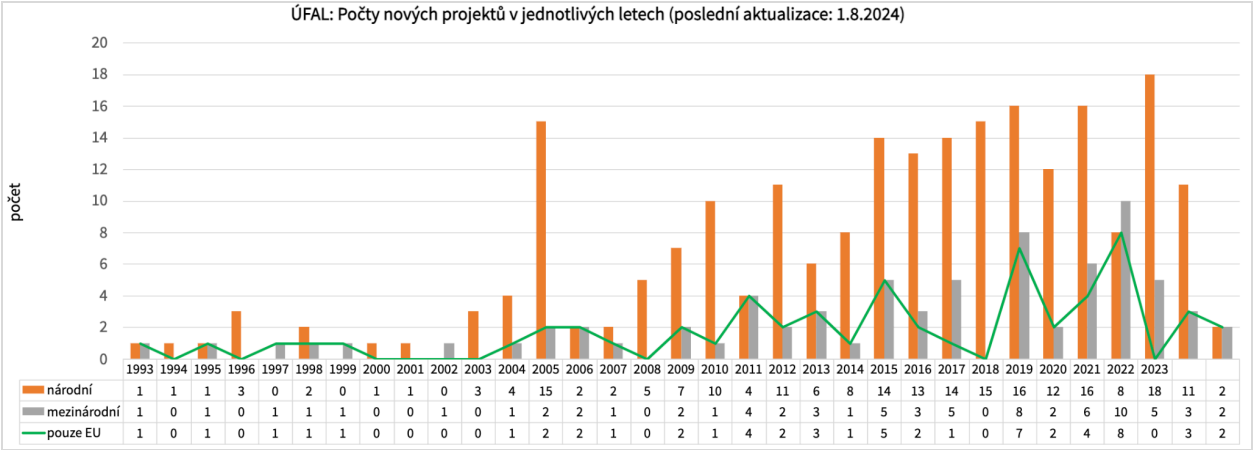
Za posledních osm let (2017-2024) vzrostl počet zaměstnanců o třetinu, přičemž průměrná výše úvazku klesla ze 72 na 67 %. Změny v alokacích/úvazcích jsou vázány na změny v návrhových listech (NL). Těch bylo předloženo ke schválení od roku 2019 průměrně 198 ročně.

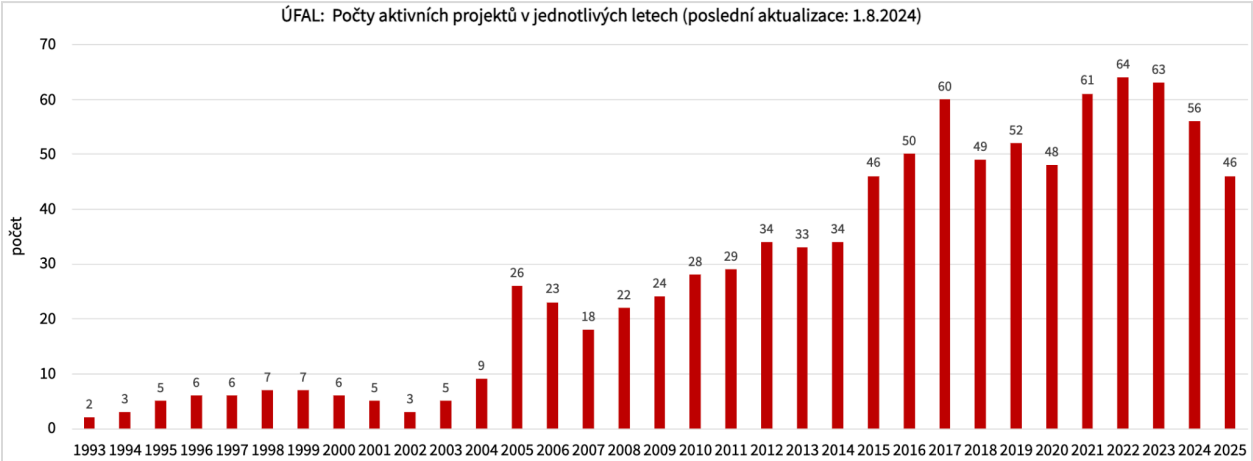
Řešitelé historicky administrovali své projekty sami, avšak tak, jak pro ně rostla administrativní zátěž projektové správy na úkor jejich vědecké a výzkumné role, tým administrátorek se rozšiřoval. Celková alokace projektových administrátorek (aktuálně celkem 5,2 FTE) odpovídá počtu projektů, velikosti jejich řešitelských týmů, počtu partnerů a variabilitě poskytovatelů grantových prostředků. Plánování využití kapacit administrátorek zohledňuje jak nové projekty, tak uzavírání těch, které končí.

Výuka a výzkumné činnosti s experimentálním a aplikačním zaměřením vyžadují robustní IT architekturu. IT podpora je stejně klíčová jako administrativní. Interní IT oddělení ÚFAL je personálně zajištěno aktuálně ve výši 2 FTE.

rok	počet zaměstnanců (k 30.3.)	∑ FTE (k 30.3.)	počet změn v NL
2017	69	49.61	
2018	76	50.27	
2019	93	59.60	206
2020	102	63.87	173
2021	102	66.04	193
2022	105	63.98	226
2023	106	64.45	259
2024	92	66.36	128

[1] https://cs.mff.cuni.cz/cs/pro-studenty/program-zamestnavani-doktorandu





3. Vzdělávací činnost

ÚFAL garantuje na MFF UK (1) zaměření Zpracování přirozeného jazyka v bakalářském programu Umělá inteligence, (2) magisterské programy Informatika - Jazykové technologie a počítačová lingvistika a Computer Science - Language technology and computational linguistics a (3) doktorský program Matematická lingvistika. Dále je ústav zapojen do evropského magisterského programu European Masters Program Language and Communication Technologies.[2]

Členové ÚFAL od ak. roku 2018/19 nabízí průměrně 46 kurzů ročně. Kurzy jsou zaměřené převážně na počítačovou lingvistiku a jsou určeny zejména magisterským studentům a doktorandům v programech uvedených výše. Kurzy strojového učení (Hluboké učení, Úvod do strojového učení v Pythonu) jsou určeny studentům všech informatických oborů. Podíl ÚFAL na výuce informatické sekce MFF UK se pohybuje mezi 8 a 10 %, například podíl Katedry aplikované matematiky nebo Katedry softwaru a výuky informatiky je průměrně 20 %. Podíl je určen na základě počtu hodin připadajících na učitele z daného pracoviště a počtu studentů na jednoho

[2] https://ufal.mff.cuni.cz/grants/lct

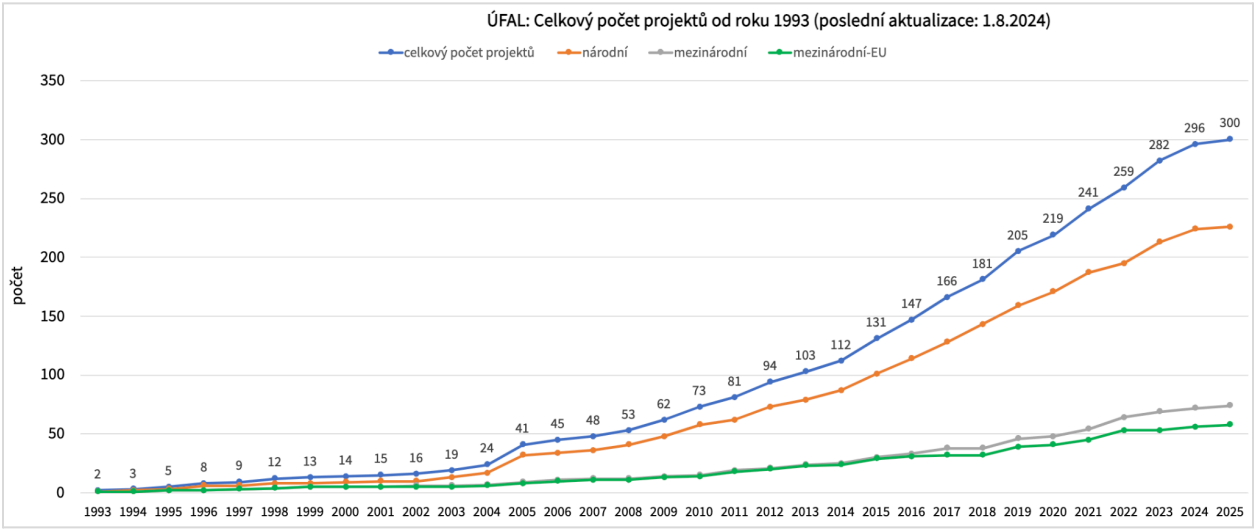
učitele z tohoto pracoviště. Nižší podíl ÚFAL je dán nízkým počtem kurzů, které ÚFAL učí pro bakalářské studenty, a primárně výzkumným zaměřením pracoviště.

Členové ÚFAL rovněž zajišťují kurzy, které zahrnují přednášky hostujících odborníků. Na ÚFAL jsou pravidelně zváni specialisté z akademické i byznysové sféry, kteří se zabývají stejnými nebo příbuznými tématy. Nejdelší tradici má Seminář z formální lingvistiky, v rámci něhož vznikl cyklus přednášek zahraničních hostů *Fred Jelinek Seminar Series*, které se pořádají na počest čechoameričana Freda Jelinka, průkopníka v oblastech automatického rozpoznávání řeči a statistického strojového překladu.[3] Volitelné kurzy umožňují flexibilně a efektivně reagovat na aktuální výzvy a příležitosti v oboru. Příkladem je kurz AI v kontextu, který ÚFAL pořádá ve spolupráci s Katedrou softwaru a výuky informatiky MFF UK. Kurz zahrnuje přednášky od odborníků z různých oblastí, které v souvislostech poskytují rozmanité pohledy na umělou inteligenci.[4]

Školitelé z ÚFAL vedli od ak. roku 2005/06 úspěšně dokončených 190 bakalářských, 205 diplomových a 58 doktorských prací (údaje jsou ze systému SIS k 11. září 2024).

4. Výzkumná činnost

Výzkumná činnost pracoviště se zaměřuje na počítačovou lingvistiku a jazykové technologie a pokrývá jak základní, tak aplikovaný výzkum empirické povahy, protože se opírá o reálná jazyková data. Jak v lingvistických studiích, tak v experimentálních výzkumech jsou významným zdrojem dat jazykové korpusy. Ty poskytují důležité podklady pro analýzu, testování teorií a trénování modelů, čímž umožňují hlubší porozumění přirozeným jazykům a rozvoj jazykových technologií. Výzkumná témata jsou často interdisciplinární, např. se zaměřením na digitální humanitní vědy nebo lékařské vědy. Relativně novým tématem je počítačové zpracování hudby.



[3] https://tinyurl.com/2p8xw2ka, https://tinyurl.com/49c587pp
[4] https://tinyurl.com/yc657tfz

Dle interní evidence bylo na ÚFAL od jeho vzniku v roce 1990 řešeno celkem 300 projektů ať už v roli hlavního řešitele, nebo spoluřešitele.[5] Křivka vizualizující roční přírůstky projektů vykazuje rostoucí průběh jak pro národní, tak pro mezinárodní projekty, což v rámci posouzení významnosti celkového počtu projektů signalizuje rozvoj pracoviště. V Příloze A jsou uvedeny počty projektů po jednotlivých letech v kategoriích národní a mezinárodní, včetně projektů financovaných EU.

V letošním roce je na pracovišti řešeno 46 projektů, přičemž k 1. srpnu bylo řešeno 38 z nich (viz Příloha E).

Významnost počtu projektů lze posuzovat dle úspěšnosti podání projektových návrhů. V interním přehledu pracoviště jsou od roku 2013 evidované i podané návrhy, které nebyly přijaty k financování. Celkem jich v letech 2013-2024 bylo 157, a proto je horní odhad úspěšnosti pracoviště od roku 1993 65.65 %.

K objektivnímu posouzení vysoké úspěšnosti podání je nezbytné srovnání s ostatními pracovišti na fakultě nebo univerzitě. Momentálně to však není možné, protože podle našich informaci na UK neexistuje centrální evidence, z níž by bylo možné čerpat. Pokud taková evidence přece jenom existuje, není dostupná pro vedoucí pracovišť. I přesto je možné hodnotit úspěšnost v širším kontextu za hranicemi ÚFAL na základě statistik od samotných poskytovatelů. V období 2013-2024 byla úspěšnost 354 návrhů podaných řešiteli z ÚFAL 55.6 % a dílčí úspěšnosti ve srovnávacím období byly následující: GA ČR 48.3 %, TA ČR 62.5. %, MŠMT 75.5 %, výzvy EU 72.5 %. Obecně platí, že čím nižší je úspěšnost soutěže a vyšší úspěšnost ÚFAL, tím výraznější je úspěch ÚFAL.

GA ČR poskytuje úspěšnosti ve všech soutěžích a oborech v časovém období 2010-2020; nejvyšší úspěšnost byla v roce 2017, a sice 33.1 % (viz Příloha F).[6] V Příloze F rovněž uvádíme úspěšnosti těch soutěží TA ČR, kterých se ÚFAL zúčastnil - většinově se zapojil do společenskovědních soutěží s vysokou úspěšností. Hodnocení úspěšnosti u MŠMT ČR by bylo kvůli velkému množství soutěží velmi časově náročné. Proto alespoň uvádíme, že v soutěži OP JAK Mezisektorová spolupráce byla v roce 2023 celková úspěšnost 21.7 % a návrh projektu "Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím" hlavního řešitele prof. Jana Hajiče z ÚFAL byl přijat k financování.[7] U evropských projektů jsou rovněž dostupné údaje o úspěšnosti výzev a navíc jsou k dispozici informace o celkových finančních příspěvcích z jednotlivých rámcových programů, které lze v interaktivní mapě filtrovat na úroveň organizace. V rámci UK činil příspěvek ÚFAL v rámci programů FP7, H2020 a HE 10.06 % (příspěvek MFF 35.17 %), viz Příloha F.

[5] Jejich kompletní přehled je publikován na https://wiki.ufal.ms.mff.cuni.cz/cisla-grantu.
[6] https://gacr.cz/wp-content/uploads/2020/04/GRAFY-NA-WEB-2020.pdf
[7] Č. projektu: CZ.02.01.01/00/23_020/0008518

V rámci smluvního výzkumu bylo v letech 2014-2024 uzavřeno 18 smluv s komerčními partnery.

Webové jazykové nástroje a služby vyvinuté na pracovišti ÚFAL jsou publikovány výzkumnou infrastrukturou LINDAT/CLARIAH-CZ s podmínkami užívání jak pro osobní a nekomerční, tak komerční použití.[8] Licenční ujednání pro komerční užití se sjednává se společností CUIP, a.s.[9]

Výsledky výzkumné činnosti ústavu jsou publikace, datové sady a software. Jejich evidenci podle let, zdrojů, typů aj. k dispozici nemáme. V univerzitním systému OBD je evidováno 2 513 záznamů. Mezi výsledky se řadí i dva US patenty evidované pod čísly US 12,056,457 B2 a US 11,037,028. První z nich, původců Ondřeje Bojara a Dominika Macháčka, je z letošního roku a pokrývá simultánní strojový překlad řeči z více jazykových zdrojů. Druhý, původců Ondřeje Bojara a Romana Sudarikova, je z roku 2021 a pokrývá trénování překladového systému pro malé jazyky a specifické domény, např. textové zprávy SMS.

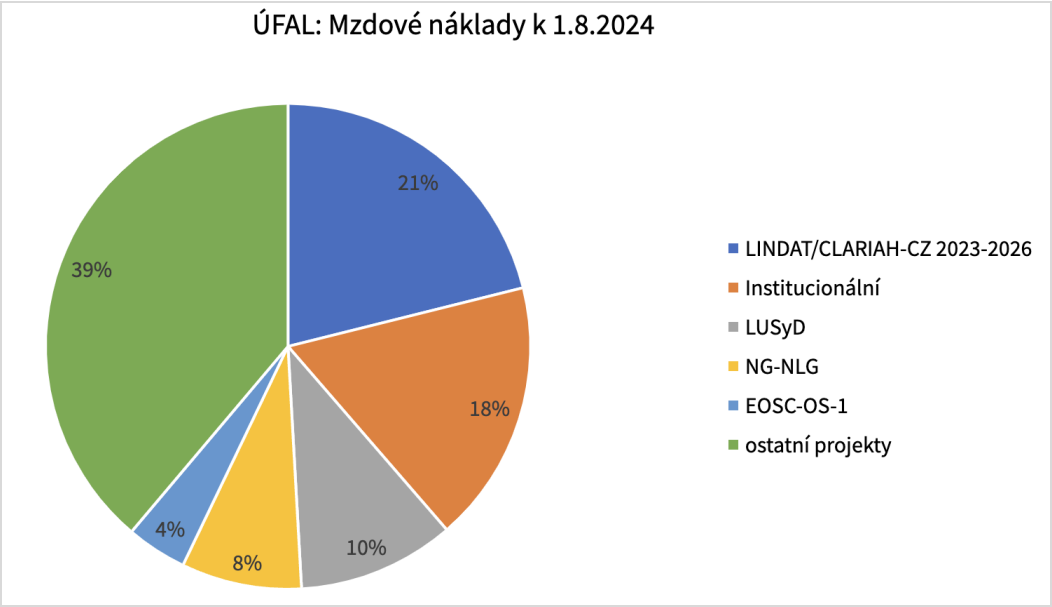
5. Finance

Roční rozpočet pracoviště je složen z institucionálních a projektových prostředků. V letech 2020–2021 činily průměrné výdaje pracoviště 95 milionů Kč.

Mzdové náklady vyplacené z institucionálních zdrojů tvoří v průměru cca 20 % celkových nákladů. 80 % je vypláceno z projektových zdrojů, nejvíce (v průměru 20 %) z velké výzkumné infrastruktury LINDAT/CLARIAH-CZ. Díky projektovým zdrojům bylo možné v letech 2023 a 2024 navýšit všem zaměstnancům ÚFAL mzdy v návaznosti na vysokou inflaci.

rok	výdaje celkem (mil. Kč)	odvedená režie (mil. Kč)	mzdové náklady celkem (mil. Kč, bez dohod)
2020	95.6	12.3	64.0
2021	95.0	9.7	63.81
2022	89.5	8.7	59.7
2023	101.3	14.6	67.81
2024(k 07)	58.5	11.8	37.45

[8] https://lindat.cz/en/services
[9] https://lindat.cz/form/smlouva-licence



6. IT Infrastruktura

Výzkum zaměřený na experimentální a aplikovaná témata vyžaduje robustní IT infrastrukturu, která musí být neustále rozvíjena. Rozvoj přináší zlepšení výkonnosti, zvýšení kapacity pro ukládání a zpracování dat a nezbytnou podporu výzkumným projektům. Vývoj infrastruktury pracoviště ÚFAL od roku 1998 je popsán v Příloze G.

7. Propagace

ÚFAL je pracoviště, které se aktivně, v rámci svých dostupných kapacit, snaží propagovat svou činnost a výsledky. Pravidelně se účastní Dnů otevřených dveří MFF UK a akce Jeden den s informatikou, kde populární formou představuje projekty a přibližuje je zájemcům o studium. ÚFAL se prezentuje i na mimofakultních/univerzitních akcí, včetně přednášek pro středoškoláky.[10]

Aktivity ústavu jsou pravidelně zveřejňovány na webových stránkách https://ufal.mff.cuni.cz.

Do roku 2022 byly aktivně evidovány popularizační ohlasy v médiích, avšak za poslední dva roky aktuální záznamy z kapacitních důvodů chybí.[11]

Aktivita ÚFAL na síti X (@ufal_cuni) je cílena převážně na vědeckou komunitu.

[10] Např. "Hoď to do stroje. Uplatní se strojový překlad v umělecké literatuře?" (V. Kloudová, O. Bojar) "Jak Transformer ukradl AI" (O. Bojar), https://www.mff.cuni.cz/cs/verejnost/fakultni-skoly/nabidka-akci-pro-fakultni-skoly

[11] https://ufal.mff.cuni.cz/popularization

8. Budoucí plány

Činnost ÚFAL zahrnuje jak tvůrčí, tak praktické aspekty, které jsou vzájemně propojené. K rozvoji tvůrčích oblastí je klíčové mít stabilní a efektivní praktickou podporu. Naší ambicí je udržet silnou pozici v obou těchto oblastech, a tím zvyšovat atraktivitu ústavu pro stávající i nové zaměstnance, studenty, akademické kolegy a komerční partnery. Níže představuji plány v oblastech, které jsou výrazně ovlivněny technologickými tendencemi současnosti. Je důležité, abychom v nich uspěli nejen odborně, ale i lidsky.

Personální činnost

Klíčové je nadále udržovat přátelské a kolegiální prostředí, které podporuje spolupráci uvnitř ÚFAL, a to v rámci projektových týmů, operativy pracoviště, mezi studenty a jejich školiteli či mezi samotnými studenty. Stejně tak je to důležité i směrem k informatické sekci a k jednotlivým složkám děkanátu/rektorátu.

Je důležité pokračovat v pravidelných schůzkách pracoviště a tematicky zaměřených poradách, například školitelů doktorandů, řešitelů projektů a projektových administrátorek.

Při počtu 100 zaměstnanců je intenzivní individuální přístup vedení pracoviště k jednotlivcům obtížně realizovatelný. V tomto ohledu považujeme plán kariérního rozvoje, který je administrativně zakotven v Opatření děkana č. 12/2024,[12] za velmi užitečný. Vnímáme ho jako příležitost pro hodnocené pracovníky a vedení pracoviště, aby se na chvíli zastavili, zhodnotili dosavadní profesní rozvoj a promysleli budoucí cíle a plány.

Vzdělávací činnost

Magisterský program Informatika - Počítačová lingvistika a jazykové technologie má akreditaci do roku 2029. S ohledem na rychlý vývoj umělé inteligence a jejích různých aspektů je rok 2029 poměrně vzdálený a předpovědět budoucí trendy není snadné. Proto je nezbytné již nyní začít diskutovat o tom, jakým způsobem reflektovat aktuální trendy ve výuce, jak přizpůsobit studijní programy a jejich interdisciplinaritu budoucím změnám a jak do nich začlenit flexibilitu a modularitu.

Výzkumná činnost

ÚFAL se věnuje všem aktuálním tématům z oblasti počítačové lingvistiky a počítačového zpracování přirozeného jazyka, multidisciplinárním tématům (např. multimodální modely, úloze generování textu) i interdisciplinárním tématům (např. z oblasti zpracování hudby, lékařských věd, historie), a tak je připraven (1) vést diskusi o tom, jakým směrem se vývoj v těchto oblastech ubírá a jaká témata zkoumat a podporovat, a (2) nadále vyhledávat vhodné projektové výzvy.

[12] https://www.mff.cuni.cz/cs/vnitrni-zalezitosti/predpisy/opatreni-dekana/opatreni-dekana-c-12-2024

Digitalizace

- Pro ÚFAL se ukázalo jako zásadní využití interně vyvíjeného administrativního systému PAKT, který hraje klíčovou roli při digitalizaci operativy pracoviště a projektové správy. Tento systém poskytuje komplexní přehled o datech, eliminuje duplicitní zadávání dat, snižuje chybovost při správě dat a celkově zefektivňuje procesy. Zajištění finančních prostředků pro jeho další rozvoj je nutné.
- Momentálně není možné PAKT propojit se správními systémy fakulty, čímž nelze dostatečně efektivně využít již existující digitalizovaná data. Tématu propojení se chceme nadále aktivně věnovat v diskusích v rámci digitalizační skupiny děkanátu MFF UK.
- Existující správní systémy fakulty poskytují specializované pohledy, které nejsou pro vedoucí pracovišť užitečné. Například data o rozpočtu pracoviště jsou dostupná pouze v ekonomickém systému CIS ve formě účetních kategorií, které nejsou pro vedoucí relevantní a kupříkladu rozlišení mezi institucionálními a projektovými zdroji vyžaduje složitou transformaci dat. I toto je téma pro diskusi v digitalizační skupině MFF UK.

Příloha A – Počty nových projektů pracoviště ÚFAL v jednotlivých letech[13]

	počet nových projektů v daném roce				kumulativní počet projektů od roku 1993			
rok	počet	národní	mezinárodní	pouze EU	počet	národní	mezinárodní	pouze EU
1993	2	1	1	1	2	1	1	1
1994	1	1	0	0	3	2	1	1
1995	2	1	1	1	5	3	2	2
1996	3	3	0	0	8	6	2	2
1997	1	0	1	1	9	6	3	3
1998	3	2	1	1	12	8	4	4
1999	1	0	1	1	13	8	5	5
2000	1	1	0	0	14	9	5	5
2001	1	1	0	0	15	10	5	5
2002	1	0	1	0	16	10	6	5
2003	3	3	0	0	19	13	6	5
2004	5	4	1	1	24	17	7	6
2005	17	15	2	2	41	32	9	8
2006	4	2	2	2	45	34	11	10
2007	3	2	1	1	48	36	12	11
2008	5	5	0	0	53	41	12	11
2009	9	7	2	2	62	48	14	13
2010	11	10	1	1	73	58	15	14
2011	8	4	4	4	81	62	19	18
2012	13	11	2	2	94	73	21	20
2013	9	6	3	3	103	79	24	23
2014	9	8	1	1	112	87	25	24
2015	19	14	5	5	131	101	30	29
2016	16	13	3	2	147	114	33	31
2017	19	14	5	1	166	128	38	32
2018	15	15	0	0	181	143	38	32
2019	24	16	8	7	205	159	46	39
2020	14	12	2	2	219	171	48	41
2021	22	16	6	4	241	187	54	45
2022	18	8	10	8	259	195	64	53
2023	23	18	5	0	282	213	69	53
2024	14	11	3	3	296	224	72	56
2025	4	2	2	2	300	226	74	58

[13] Zdroj: Interní evidence ÚFAL

Příloha B – Počty zpracovaných cestovních příkazů na pracovišti ÚFAL[14]

rok	počet cestovních příkazů
2007	138
2008	157
2009	202
2010	213
2011	190
2012	229
2013	190
2014	229
2015	214
2016	206
2017	224
2018	231
2019	226
2020	28
2021	108
2022	172
2023	252
2024	182

[14] Zdroj: systém SIS

Příloha C – Počty přijatých zahraničních hostů na pracoviště ÚFAL[15]

rok	počet přijatých zahraničních hostů
2007	13
2008	21
2009	27
2010	17
2011	24
2012	16
2013	24
2014	16
2015	41
2016	37
2017	40
2018	32
2019	30
2020	8
2021	0
2022	11
2023	10
2024	14

[15] Zdroj: systém SIS

Příloha D – Počty žádanek vystavených pracovištěm ÚFAL[16]

rok	počet žádanek
2018	8 (XI-XII)
2019	95
2020	120
2021	124
2022	119
2023	148
2024	179

[16] Zdroj: systém CIS

Příloha E – Výzkumné projekty řešené na ÚFAL k 1.8.2024[17]

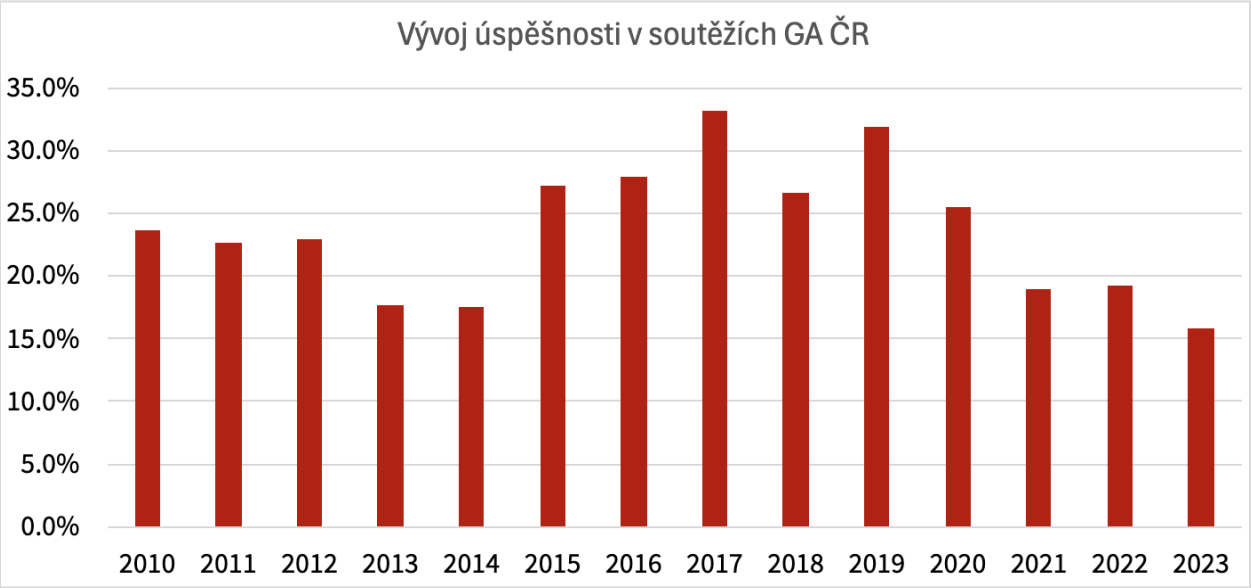
1. NG-NLG Dušek (ERC)
   Next-Generation Natural Language Generation
2. HPLT / Hippolyta Hajič (EC HE)
   High Performance Language Technologies
3. MEMORISE Pecina (EC HE)
   Virtualisation and Multimodal Exploration of Heritage on Nazi Persecution
4. RES-Q Plus Pecina (EC HE)
   Comprehensive solutions of healthcare improvement based on the global Registry of Stroke Care Quality
5. EVERSE Straňák (EC HE)
   European Virtual Institute for Research Software Excellence
6. ATRIUM Straňák (EC HE)
   Advancing FronTier Research In the Arts and hUManities
7. FSTP InCroMin Bojar (EC HE - FSTP)
   Interactive Crosslingual Minutes
8. CLS INFRA Cinková (EC H2020)
   Computational Literary Studies Infrastructure
9. HumanE-AI-Net Hajič (EC H2020)
   HumanE AI Network
10. UMR Hajič (MŠMT Inter-Action)
    Univerzální reprezentace významu UMR
11. IE-II InterAction Pecina LUABA24 (MŠMT Inter-Action)
    Nové metody pro vyšetření žaludku pomocí umělé inteligence: Využití hlubokého učení pro asistovanou gastroskopii
12. EOSC-OS-1 Hajič (MŠMT OP JAK)
    Národní repozitářová platforma pro výzkumná data
13. LINDAT/CLARIAH-CZ 2023-2026 Hajič (MŠMT VVI)
    Digitální výzkumná infrastruktura pro jazykové technologie, umění a humanitní vědy
14. VI-I LINDAT Hajič (MŠMT OP JAK)
    LINDAT/CLARIAH-CZ Přístrojové vybavení
15. Cooperatio (UK)
    ÚFAL je zapojen do dvou vědních oblastí, a sice Computer Science (COMP, koordinuje MFF UK) a Linguistics (LING, koordinuje FF UK)
16. CVHM Hajič (UK)
    Centrum vizuální historie Malach
17. UNCE Multilingual Lens Žabokrtský (UK)
    Výzkum velkých textových korpusů prizmatem vícejazyčnosti a komplementárních metodologických přístupů
18. PRIMUS 2023 Libovický (UK)
    Jazykově neutrální a kulturně kontrolovatelné vícejazyčné neuronové reprezentace vět
19. GAUK 2023 Jon (UK)
    Metody pro zlepšení neuronového strojového překladu různorodých textů
20. GAUK 2023 Mayer (UK)
    Generování syntetických trénovacích dat a jiné metody pro rozpoznávání psaných notopisů
21. GAUK 2023 Hledíková (UK)
    Morfematická komplexita slovesné slovní zásoby ve čtyřech jazycích: Kvantitativní výzkum založený na korpusových datech

[17] Červeně označené projekty mají hlavního řešitele projektu z ÚFALu.

22. GAUK 2023 Javorský (UK)
    Používanie pomocných podúloh na učenie obmedzení v NLP
23. GAUK 2024 Abishek (UK)
    Modelování toku morfémů mezi jazyky
24. GAUK 2024 Gamba (UK)
    Přizpůsobení formalismu UMR
25. NAKI 2023 Hajič jun. (MK)
    OmniOMR – rozpoznávání hudebního záznamu pomocí strojového učení pro digitální knihovny
26. NAKI 2023 Rysová (MK)
    Automatické hodnocení mluveného projevu v češtině
27. TAČR SIGMA Rosa EduPo
    Generování české poezie v edukačním a multimediálním prostředí
28. TAČR SIGMA Poláková edUKate
    Podpora digitálního vzdělávání cizojazyčných dětí prostřednictvím počítačového překladu
29. TAČR SIGMA Hladká PONK
    PONK - Asistent přístupné úřední komunikace
30. TAČR SIGMA Hladká Newsroom AI
    Newsroom AI: veřejná služba v éře automatizované žurnalistiky
31. TAČR SIGMA Zeman HiČKoK
    HiČKoK: Historie češtiny v korpusovém kontinuu
32. LUSyD Hajič (GAČR EXPRO)
    Porozumění jazyku: od syntaxe k diskurzu
33. GAČR 2022 Kolářová
    Odraz slovotvorných vztahů ve valenci substantiv
34. GAČR 2022 Mírovský
    Metody pro rychlou diskurzní anotaci ve vybraných korpusech
35. GAČR 2023 Štěpánková
    Prostředky vyjadřování epistémické modality a evidenciality v češtině
36. GAČR 2023 Mareček
    Identifikace a prevence nechtěné genderové zaujatosti v neuronových jazykových modelech
37. GAČR 2023 Mikulová
    Funkce a formy okolnostních určení
38. GAČR 2024 Zikánová
    Neshoda v korpusové anotaci ve vztahu k víceznačnosti textu

Příloha F – Úspěšnost podání projektových návrhů pro vybrané poskytovatele

GA ČR poskytuje úspěšnosti ve všech soutěžích a oborech v časovém období 2010-2020.



V TA ČR se zaměřujeme pouze na soutěže, kterých se ÚFAL účastnil do roku 2023 (včetně).

program a soutěž	rok	úspěšnost soutěže (%)	úspěšnost ÚFAL (%)
Alfa 2. veřejná soutěž	2020	náročné dohledat	100.0
Éta 3. veřejná soutěž	2020	19.0	100.0
Éta 4. veřejná soutěž	2020	10.7	0.0
Éta 5. veřejná soutěž	2021	11.7	25.0
Trend 3. veřejná soutěž	2021	37.6	100.0
Trend 3. veřejná soutěž	2022	22.8	0.0
Sigma 1. veřejná soutěž DC5	2023	18.2	85.7

V evropských projektech se zaměřujeme na rámcové programy FP7, H2020, HE a na ty soutěže, kterých se ÚFAL zúčastnil.[18]

soutěž	rok	úspěšnost soutěže (%)	úspěšnost ÚFAL (%)
ERC-2020-STG	2020	13.36	100.00
HORIZON-CL2-2021-HERITAGE-01	2021	3.39	100.00
HORIZON-HLTH-2021-TOOL-06	2021	5.95	100.00
HORIZON-CL4-2021-HUMAN-01	2021	1.46	0.00
HORIZON-CL4-2021-DIGITAL-EMERGING-01	2021	2.75	0.00
HORIZON-INFRA-2021-EOSC-01	2021	6.67	0.00
HORIZON-CL4-2021-DATA-01	2021	4.59	100.00
PPPA-LANGEQ-2021	2021	není k dispozici	100.00
HORIZON-CL2-2022-DEMOCRACY-01	2022	1.24	0.00
HORIZON-CL4-2022-HUMAN-01&02	2022	3.91	0.00
HORIZON-INFRA-2022-TECH-01	2022	35.90	0.00
HORIZON-MSCA-2022-DN-01-01	2022	16.70	0.00
HORIZON-CL4-2023-DIGITAL-EMERGING-01	2023	2.60	0.00
HORIZON-INFRA-2023-SERV-01	2023	12.50	100.00
HORIZON-INFRA-2023-EOSC-01	2023	5.88	100.00
HORIZON-CL4-2023-HUMAN-01	2023	1.48	0.00
HORIZON-WIDERA-2023-ACCESS-02	2023	8.96	0.00
ERC-2023-SyG	2023	9.14	0.00
HORIZON-CL2-2024-DEMOCRACY-01	2024	není k dispozici; 287 návrhů	0.00
průměrná úspěšnost		10.70	26.92

[18] https://tinyurl.com/yc63tc92

Pro evropské projekty jsou rovněž k dispozici výše příspěvků EU. Níže uvádíme příspěvky pro ÚFAL a MFF v rámci UK. Data jsou k 9. září 2024. V databázi nejsou sjednoceny názvy institucí, resp. jména oddělení nejsou uvedena pod stejným jménem/ identifikátorem, proto přikládáme názvy, které jsme v datech filtrovali.

část UK	příspěvek EU (EUR)	podíl na celkovém příspěvku UK (%)
MFF UK bez ÚFAL[1)]	22,347,193.57	25.11
ÚFAL pouze[2)]	8,948,949.25	10.06
MFF UK celkem	31,296,142.82	35.17

1) unikátní názvy organizačních jednotek, které byly uvedeny v návrzích a které EC uvažuje v uvedených statistikách: Computer Science Institute of Charles University, CUNI MFF| Department of Mathematics and Physics, Department of Algebra, Faculty of Mathematics and Physics, Department of Applied Mathematics, Department of Atmospheric Physics, Department of Chemical Physics and Optics, Fac. of Math. and Phys, Department of Condensed Matter Physics, Department of Distributed and Dependable Systems, Department Of Distributed And Dependable Systems, Faculty Of Mathematics And Physics, Charles University, Department Of Geophysics, Faculty Of Mathematics And Physics, Department of Mathematics and Mathematics Education, Department Of Surface And Plasma Science, Faculty Of Mathematics And Physics, Department Of Surface And Plasma Science/Faculty Of Mathematics And Physics, Department of Theoretical Computer Science and Mathematical Logic, Dept. Chemical Physics And Optics, Faculty Of Mathematics And Physics, Dept. of Atmospheric Physics, Fac. of Mathematics and Physics, Dept. of Atmospheric Physics, Faculty of Mathematics and Physics, Dept. Of Theoretical Computer Science And Math Logic, Faculty Of Mathematics And Physics, Faculty of Mathematics and Physics, Faculty of Mathematics and Physics, Computer Science Institute of Charles University, Faculty Of Mathematics And Physics, Dept.Of Physics Education, Institute of Nuclear and Particle Physics, Institute of Particle and Nuclear Physics, Institute of Particle and Nuclear Physics (IPNP), Institute Of Physics Of Charles University, Faculty Of Mathematics And Physics, Institute of Physics of the Charles University, Institute of Theoretical Physics, Faculty of Mathematics&Physics, Mathematics And Physics Faculty, School Of Computer Science, Faculty Of Mathematics And Physics, Theoretical Computer Science and Mathematical Logic
2) sečteno přes všechny uvedené varianty názvu ÚFAL: Institute of Formal and Applied Linguistics, Institute Of Formal And Applied Linguistics, Mff, Institute of Formal and Applied Linguistics (IFAL)

Příloha G – Rozvoj IT infrastruktury pracoviště ÚFAL od roku 1998

V roce 1998 se pracovalo na počítačích s procesory Intel Pentium první generace a staršími (486, 386, 286) a s CRT monitory, vážícími i desítky kilogramů. Experimentovalo se na poměrně výkonných a drahých strojích Sun Microsystems. V této době v budově na Malé Straně ještě nebyla servrovna a veškerý výzkum probíhal na běžných kancelářských počítačích. Pracovalo se na počítačích jednak s Windows, jednak s Linuxem a Solarisem. Velikost operační paměti byla z dnešního pohledu velmi nízká, 16MB vs. 16GB, tj, 1 000x menší. Běžný pevný disk pro data měl kapacitu ve stovkách MB (dnes typicky jednotky TB, takže opět 1 000x výše). Objem dat, se kterými se mohlo pracovat, byl řádově nižší. Stejně tak tomu bylo i s komunikací – datová síť používala koaxiální kabely, které dnes známe jako kabel pro připojení k televizi. Rychlost připojení obvykle nepřekračovala 10Mbps (zhruba 1MB za vteřinu). Dnes jsou pracovní stanice připojeny typicky rychlostí 1 000Mbps. Nejvýkonnější stroje, které v té době ÚFAL měl, byla čtveřice počítačů od Sun Microsystems. Měly vlastní operační systém Solaris, velké objemné monitory, vlastní procesory a vlastní diskové pole. Toto úložiště společně se stroji bylo velmi pravděpodobně to nejlepší, co se bylo v té době na MFF UK k dispozici – optické připojení diskového pole s kapacitou asi 40GB, 512MB operační paměti, rychlý procesor. Pořizovací náklady nebyly zanedbatelné, ale umožnily práci s rozsáhlými (na tu dobu) daty a zobrazování složitých objektů na obrazovce. Mírná odlišnost systému byla považována za přijatelnou cenu. Veškerou správu počítačového vybavení zajišťovali zaměstnanci sami, s podporou studenta na částečný úvazek. Přibližně v této době začalo formování IT oddělení ústavu.

Období let 2000-2004 bylo v mnoha ohledech klíčové. Bylo založeno Centrum komputační lingvistiky, významný projekt vedený prof. Evou Hajičovou, který zajistil dostatek finančních prostředků k výraznému rozvoji výpočetní infrastruktury ústavu. IT oddělení tehdy tvořil jeden zaměstnanec na plný úvazek, později podpořený studentem pracujícím na částečný úvazek. Namísto výpočtů na pracovních stanicích byly zakoupeny tři rackové skříně, které obsahovaly stroje sloužící jako zárodek budoucího výpočetního clusteru a centrální úložiště s kapacitou přibližně 200 GB. ÚFAL získal samostatné připojení 1Gbps optickou linkou, čímž bylo možné pracovat s výrazně většími daty. Srovnatelný výpočetní výkon tehdy nabízelo pouze ÚVT UK. Praktické zkušenosti z provozu byly využívány také v informatické sekci MFF, přičemž cílem ÚFAL vždy bylo přispět k rozvoji centrálních infrastruktur, které sloužily všem a umožňovaly i další rozvoj ústavu. O několik let později byla v budově na Malé Straně zřízena centrální serverovna, kam se výpočetní stroje a datová úložiště přesunuly do klimatizovaných prostor se záložním zdrojem. Tento krok výrazně zvýšil bezpečnost provozu, a to díky spolupráci Střediska informatické sítě a laboratoří a informatické sekce MFF UK.

V roce 2009 zahájil ÚFAL spolupráci s University of Southern California (USC) a vzniklo Centrum vizuální historie Malach (CVHM) s cílem zpřístupnit audiovizuální archiv svědectví holocaustu a později i dalších genocid. K realizaci tohoto centra byla podepsána meziuniverzitní dohoda o spolupráci. V knihovně MFF UK na Malé Straně byla vyhrazena místnost vybavená počítači a projekcí, ÚFAL zřídil zabezpečené datové úložiště a ve spolupráci se společností CESNET zajistil

přenos dat z USA. V té době se jednalo o mimořádně objemná data a jejich přenos na tak velkou vzdálenost představoval značnou výzvu, přičemž optimalizace trvala několik týdnů. K archivu má dnes přístup také Židovské muzeum, v CVHM se pořádají výukové programy a archiv využívají i historici. Kromě této funkce se data využívají ve spolupráci ÚFAL, USC a ZČU na projektech zaměřených na zpracování zvukových záznamů, jejich přepisů, indexaci obsahu a překlad titulků do více jazyků, což umožňuje efektivní prohledávání rozsáhlého archivu. V dalších letech byl vybudován robotizovaný páskový archiv v Areálu Troja, jehož současná maximální kapacita je cca 45PB (45 tisíc TB). Nedávno bylo zahájeno doplňování dat, která jsou nyní přístupná přímo v Evropě. V současnosti jsou tři kopie archivu uloženy v USA, přičemž další je plánována v Paříži. Z technického hlediska je tento archiv v rámci Univerzity Karlovy jedinečný. Přírodovědecká fakulta UK zvažuje podobné vybavení, což otevírá možnost sdílet zkušenosti s jeho architekturou a provozem. Modernizaci archivu financuje USC, zatímco ÚFAL zajišťuje správu a rozvoj systému na místě.

V roce 2009 začaly přípravy na vybudování infrastrukturního centra LINDAT/CLARIN, což představuje další milník v technologickém rozvoji ústavu. V roce 2010 byl v budově na Malé Straně zprovozněna druhá serverovna z finančních prostředků LINDAT/CLARIN. Díky této investici do infrastruktury se podařilo vytvořit bezpečné prostředí umožňující výrazné zvýšení výkonu výpočetního clusteru Linguistic Research Cluster (LRC) a infrastruktury pro provoz repozitáře jazykových dat a nástrojů. Tento repozitář, společně s rozvojem softwaru DSpace, dodnes významně přispívá k podpoře oboru a má klíčový význam i pro Univerzitu Karlovu.

Rok 2011 přinesl další významný milník, kdy byly do výpočetního clusteru LRC zařazeny první dva stroje s GPU kartami nVidia GeForce GTX 570 a nainstalovány Cuda balíčky. Tyto kroky položily základy pro současné využívání velkých jazykových modelů a generativní umělé inteligence na pracovišti. Díky tomu bylo možné zahájit experimenty se strojovým překladem na neuronových sítích využívajících GPU karty. V budově na Malé Straně se postupně začal budovat GPU cluster Deep Learning Laboratory (DLL), který se stal specializovanou součástí LRC. Byly nasazeny servery s desítkami GPU karet. V této době došlo také k navýšení alokace IT zaměstnanců na dva plné úvazky, což umožnilo efektivní správu rostoucí infrastruktury.

Na základě zkušeností s výstavbou LRC/DLL byl vytvořen výukový cluster Artificial Intelligence Cloud (AIC). Cílem bylo umožnit bakalářským a magisterským studentům přístup k technologiím, se kterými pracují doktorandi na ÚFAL. Přístup k AIC je na žádost poskytován také studentům z dalších pracovišť informatické sekce MFF UK. V posledních letech je na obdobných základech budován centrální výpočetní cluster Chiméra, který slouží všem studentům MFF UK, přičemž IT oddělení ÚFAL působí jako konzultant a předává své zkušenosti realizačnímu týmu. MFF UK tak postupně směřuje k jednotnému modelu práce s daty a výpočetními prostředky.

Na rozvoj DLL bylo navázáno díky rozvojovému projektu v roce 2016, kdy byla v režii fyzikální sekce MFF UK vybudována další serverovna v Areálu troja a ÚFAL tak mohl přejít od používání nejvýkonnějších GPU karet určených i pro herní sestavy na profesionální karty, určené pro datová

centra. Úměrně růstu výkonu clusteru rostla i velikost dat a bylo tedy nutné rozšířit kapacity clusterového paralelního úložiště Lustre na asi 400TB.

V následujících letech pokračovala modernizace a rozšiřování stávající infrastruktury, která dosáhla současné podoby. Došlo k posílení IT oddělení a zrychlení počítačových sítí na 100-200 Gbps na hlavních trasách. Úložiště clusteru nyní mají kapacitu přes 1 PB a zahrnují různé typy, od kapacitních a pomalejších po vysoce rychlá, založená na NVMe discích. LRC spravuje přístup k několika tisícům výpočetních jader a více než 200 GPU kartám různých kategorií. Kromě toho ÚFAL provozuje GPU karty také v rámci výukového clusteru a aplikačního cloudu, kde jsou prezentována dema i finální produkty projektů. ÚFAL disponuje celkem sedmnácti rackovými skříněmi ve čtyřech serverovnách, přičemž každá skříň měří přibližně 2 metry, což by vytvořilo sloupec vysoký 34 metrů odpovídající výšce jedenáctipatrového domu. Spotřeba energie je značná a maximální zatížení by znamenalo příkon serverů kolem 100 kW, k čemuž je třeba připočítat chlazení. S průměrným koeficientem 1,6 se maximální spotřeba dostává až na 160 kW. Tato situace se však nevyskytuje často, protože nejvytíženější je nejmodernější vybavení, které je energeticky nejefektivnější.