## WORKSHOP LEXICAL DATA MASTERCLASS 2018

Ve dnech 3.–7. 12. 2018 proběhl v Berlíně druhý ročník workshopu **Lexical Data Masterclass** (https://lexmc18.sciencesconf.org), který podpořilo francouzské Ministerstvo vysokého školství, výzkumu a inovací, infrastruktury DARIAH-EU, CLARIN, ELEXIS — European Lexicographic Infrastructure, BCDH (Belgrade Center for Digital Humanities), Berlínsko-braniborská akademie věd (Berlin-Brandenburgische Akademie der Wissenschaften) a společnost SyncRO Soft. Díky posledně jmenované firmě získali účastníci bezplatnou půlroční licenci softwaru oXygen XML Editor (https://www.oxygenxml.com/xml_editor.html), který umožňuje pohodlnou práci s dokumenty ve formátu XML a se souvisejícími technologiemi. Dalším benefitem byla úhrada nákladů na dopravu a ubytování na základě žádosti o proplacení účetních dokladů u DARIAH-EU.

Hlavními organizátory workshopu byli zejména Laurent Romary a Toma Tasovac, kteří stojí za přípravou nového standardu pro značkování digitálních slovníků, jež nese označení TEI-Lex0 (https://github.com/LingSIG/Dictionaries). Významnou měrou se na akci podílel také Mohamed Khemaken, který vyvíjí aplikaci GROBID--Dictionaries (viz níže).

Pětidenní akce je rámovaná prezentací účastníků. První den představují své projekty, na nichž budou v rámci workshopu pracovat. Tato část má pevná pravidla: do tříminutové prezentace (12 snímků po 15 sekundách) je potřeba vtěsnat gros lexikografického problému. Závěrečná vystoupení, která už nebyla tak striktně limitovaná časem, slouží ke shrnutí pokroku v práci a k prezentaci dosažených výsledků. V roce 2018 bylo pro účast na workshopu vybráno 16 projektů.

Od úterý se program rozdělil do tří skupin. Jedna se zaměřila na práci s aplikací **GROBID-Dictionaries** (https://github.com/MedKhem/grobid-dictionaries), která pomocí strojového učení dokáže v naskenovaném materiálu rozpoznat a označit jednotlivé části slovníkové heslové stati (ve formátu XML TEI P5); na tuto sekci bylo potřeba se přihlásit pomocí samostatného formuláře. Ve zbývajících dvou sekcích se na začátku dne probírala dílčí témata a poté následovala asistovaná práce nad konkrétním materiálem jednotlivých účastníků. Asistovali nejen organizátoři, ale i samotní účastníci, pokud se v probíraných technologiích dostatečně vyznali.

Mezi prezentovaná témata patřil úvod do kódování slovníků pomocí doporučení TEI (**Introduction to encoding dictionaries with the TEI guidelines**), pokročilé techniky s programem oXygen XML Editor (**Advanced Oxygen techniques**), využití XPath pro hledání ve slovnících (**Advanced Path for searching dictionary content**), transformace pomocí XSLT do jiných formátů (**XSLT**), představení projektu Standardization Survival Kit (**Data management**; http://ssk.parthenos-project.eu), který nabízí doporučené postupy v různých oblastech digitálních humanitních věd a výzkumu kulturního dědictví. Obecně lze říci, že jsou tyto prezentace určeny zejména pro začátečníky, ale pokročilejší a specializovaná témata mohou účastníci probrat se školiteli během zbytku dne, kdy se věnují svým vlastním projektům.

V plenární přednášce **The VICAV Dictionaries: Working on a Virtual Research Environment** představil Karlheinz Moerth z Rakouského centra pro digitální hu-

 

manitní vědy (ACDH, https://www.oeaw.ac.at/de/acdh/) infrastrukturu pro vídeňský korpus variet arabštiny.

Lexikografická témata, s nimiž se workshopu účastnili jednotliví badatelé, byla značně různorodá. Část se věnovala retrodigitalizaci slovníků s pomocí aplikace GROBID-Dictionaries, a to jak historických tištěných (německo-srbský hornický slovník) a strojopisných (terminologický slovník makedonštiny), tak i moderních (výkladové slovníky turečtiny). Ke specializovaným slovníkům bychom mohli zařadit slovníky italštiny z 19. století, právní slovníky španělštiny nebo slovník slovanské korpusové terminologie. Některé projekty se týkaly nejen zachycení slovní zásoby, ale i dalších jazykových a metajazykových rovin, např. dialektů Àbèsàbèsì (v Nigérii) nebo dokumentace lingvistické terminologie v citátech z Koránu. Část účastníků řešila vylepšení dosavadního způsobu značkování (Elektronický slovník staré češtiny, Historický tezaurus angličtiny, akademický slovník portugalštiny), analýzu dostupných slovníkových dat (Deutsches Wörterbuch bratří Grimmů) nebo transformaci údajů do formátu TEI (WordNet bulharštiny). Podrobnější přehled všech témat je k vidění na adrese https://digilex.hypotheses.org/551.

Jak prezentace, tak studijní materiály z dosavadních ročníků jsou dostupné na úložišti GitHub (https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Events/). V ideálním případě je zde k dispozici i pracovní materiál jednotlivých účastníků, pokud k jeho publikaci mají svolení, takže může posloužit jako inspirace pro zájemce o specifickou problematiku. Další možností, jak informovat vědeckou komunitu o dílčích problémech a jejich řešeních, jsou blogové příspěvky na stránkách https://digilex.hypotheses.org; přístupové údaje k tomuto médiu zasílají pořadatelé během workshopu jeho účastníkům.

Podobná setkání jsou důležitá pro zájemce o prezentované standardy a technologie, kteří se o nich můžou dozvědět nejen základní informace, ale v případě potřeby i detaily, které pomohou vyřešit jejich problémy. Zároveň je tento druh akcí přínosný i pro organizátory, kteří získají zpětnou vazbu a mohou ovlivňovat další směřování standardů a vývoj softwarových aplikací.

**Boris Lehečka** | Ústav pro jazyk český AV ČR, v. v. i. | Valentinská 1, 116 46 Praha 1
ORCID ID: 0000-0003-4893-5537
boris@daliboris.cz