# Datace primárních jazykových pramenů

Boris Lehečka, 3. 5. 2017

Při dataci primárních jazykových pramenů pro potřeby diachronního bádání se můžeme setkat s následujícími otázkami a specifiky.

1. Není vždy jasné, zda se časovým údajem myslí období, během něhož pramen vznikl, nebo čas, kdy byl pramen dokončen a začal působit jako samostatné dílo.
   1. Deskriptivní pohled (jak údaje jednoznačně zachycovat):
      1. Předpokládám, že většinou budeme znát spíš jenom druhou možnost – (přibližný) časový interval, kdy byl pramen dokončen (třeba 1510–1512). Navrhuju ale první variantu (období, kdy pramen vznikal) odlišovat zavedením dodatečného údaje pro dobu vzniku.
      2. Ve výsledku pak bude možné uvádět kromě (anebo namísto) @notBefore a @notAfter pro období dokončení ještě @notBefore a @notAfter pro celou dobu, kdy dílo vznikalo (třeba 1503–1512).
      3. Nejsložitější možnost, zavést místo předchozího dokonce interval pro čas, kdy dílo začalo vznikat (1502/1504), a čas, kdy bylo dokončeno (1510/1512), bych zvolil jen tehdy, kdyby pro větší množství pramenů bylo žádané tyto údaje zaznamenávat.
   2. Problematika řazení zdrojů při hledání v korpusu (i při filtrování):
      1. Období přípravy díla může mít pro účely řazení a filtrování přednost před údajem o období dokončení. Navenek může nový údaj vystupovat namísto základního.
2. V některých případech, např. u souboru listin nebo archivních zápisů, je potřeba počítat s tím, že jednotlivé zápisy tvoří samostatné části a vznikaly v průběhu několika (desítek) let, jazykověda je však (pro snazší uchopení velkého množství pramenů) bere jako pramen jediný (např. AktaBratr 1 a AktaBratr 2). Obdobný případ představují např. konvoluty, které mohou zahrnovat několik samostatných jazykových pramenů, jež se však nerozlišují a za „samostatný“ pramen se považuje konvolut jako celek (např. PasMuzA: „str. 275–436 a 451–490 obsahují zbytky starého Pasionálu z pol. 14. stol., ostatek z 2. pol. 14. stol.“). Podobně se to má s jazykovými prameny sekundárními, které obsahují samostatná díla z delšího časového období, např. Vilikovského edice Staročeská lyrika (LyrVil) se skladbami ze 14. a 15. století.
   1. Deskriptivní pohled:
      1. Opět podle mě dává smysl, kdybychom kromě hlavního časového údaje, který je v hlavičce šablony, měli možnost v dokumentu vymezit části s vlastním datováním.
      2. V textu by se taková část (rozsah stran) vymezila třeba pomocí dalšího znakového stylu. Příklad: *text na straně 274* [označení začátku strany 275 pomocí k tomu určeného znakového stylu] **[označení začátku odlišně datovaného textu pomocí nového znakového stylu s obsahem „1. pol. 14. stol.“]** *text na stranách 275–436* **[označení konce odlišně datovaného textu]** [označení začátku strany 437]
   2. Problematika řazení zdrojů při hledání v korpusu (i při filtrování):
      1. Tím, že se vymezí zvláštní, odlišně datované části díla, je možné dokument pro účely řazení výsledků v korpusu považovat za několik samostatných, odlišně datovaných dokumentů. (Zobrazení v edičním modulu se nezmění, jde stále o jeden celek.)
3. Přesná datace, určení konkrétního data vzniku/působení díla či alespoň roku není vždy možná. Čím starší je období, z něhož pramen pochází, tím složitější bývá určit přesnou dataci. V takových případech se určuje období vzniku pramene pomocí časového rozsahu, který může mít různé časové rozpětí: od jednotek po desítky či stovky let.
4. Určení rozmezí mívá podobu přesných milníků (1408/1410, 1401–1404), ustálených období (století, polovina, čtvrtletí, třetina, léta) i „neměřitelných“ údajů (začátek, konec, okolo).
5. K čemu datace primárních jazykových pramenů slouží? Dataci lze využít jako pomocné kritérium pro utřídění velkého množství materiálu. Primární jazykové prameny ze stejného období budou s velkou pravděpodobností obsahovat stejné jazykové jevy (zejména v oblastí hláskosloví a morfologie). Časové určení vzniku pramene (před)určuje jeho vztah k dalším pramenům a autorům (tj. např. předlohy a opisu, možného autorství ap.). Datace pramene může mít dále vliv na interpretaci zápisu jazyka (např. zda se u zápisu „ye“ jedná o jotaci (ě), nebo o zvratnou písařskou analogii, kterou je potřeba interpretovat jako depalatalizované e).
6. Jakým způsobem se využívá datace v korpusu/textové bance/edičním modulu?
   1. uživatel si omezí prohledávaná díla na určité časové období;
   2. uživatel seřadí nalezené doklady chronologicky, čím se k sobě dostanou doklady ze stejného časového období, resp. ukáže vývoj výrazu od nejstarších forem po nejnovější;
   3. obdobně je v případě chronologického seřazení děl;
   4. přehled o zastoupení děl/žánrů/autorů ap. v jednotlivých časových obdobích (současné pojetí datace ve Vokabuláři webovém počítá s časovým obdobím po 50 letech)
7. S čím je potřeba při výše uvedených specificích datace počítat?
   1. Pro časové rozmezí se určuje hranice (obvykle rok), kdy pramen ještě nemohl vzniknout (v XML TEI atribut @notBefore) a kdy už vzniknout nemohl (v XML TEI atribut @notAfter).
   2. Pokud pramen obsahuje texty, které vznikaly v rámci více období (např. v rámci jednoho století), měl by být pramen zahrnut v přehledu/filtru pro každé jedno období? (Tj. měl by se např. GuallCtnostK, který vznikl v 15. století, objevit v přehledu pro období 1501–1550 i 1551–1600, nebo pouze v jednom z nich? Toto řešení může badatele mást: pokud se pramen objeví v obou obdobích, součet pramenů nebo např. tokenů ve všech obdobích bude vyšší než celkový součet za jednotlivé prameny.)
      1. Souhlasím s tím, aby se pramen zobrazoval v obou obdobích. Problémy z toho plynoucí se pokouším řešit níže.
      2. Pokud budou části z různých období vyznačené ve zdrojovém textu (víz 2.1 a 2.2) a tím pádem pro účely vyhledávání tvořící samostatné dokumenty, zobrazoval bych pro období 1501–1550 pouze části GuallCtnostK, které do něj spadají, a analogicky pro období 1551–1600.
      3. Naopak, pokud nebude u GuallCtnostK možné datovat jednotlivé části a bude proto nutné uvést období vzniku pouze pro celek, dal bych to uživateli najevo: zobrazoval bych třeba u pramene dobře viditelnou poznámku, že se objevuje ještě v dalších obdobích (případně při zobrazení přehledu více období bych zobrazoval vedle GuallCtnostK indexy 1 a 2, aby bylo zřejmé, že má dílo dva výskyty).
      4. Problém s tím, že se pramen nalezne (pro účely počítání tokenů) při práci se širším obdobím dvakrát bych v případě, že opravdu půjde o celek (který nebude rozparcelovaný na různě datované části), nejraději řešil tak, že se ve výsledcích zobrazí jen jednou.
      5. Pokud by nebylo možné se duplicitám ve výsledku vyhnout, alespoň bych na to uživatele upozornil.
      6. Pokud nebude ani to možné, bude nutné zvolit, ve kterém období pramen uvádět. K tomu bych zavedl (pokud to ještě neexistuje) nový údaj, který se odvodí z období vzniku: jedno přesné kanonické/**reprezentativní datum**, které bude ležet buď na začátku, uprostřed, nebo na konci období vzniku. Sloužit by mohlo i pro řazení.
   3. Při seřazení pramenů/dokladů podle datace v chronologické řadě, v níž se mísí přesně a volněji datované prameny, je potřeba někam umístit prameny s časovým údajem obsahujícím rozmezí. Mají se tyto případy zařadit před přesně datované prameny, až za ně, nebo doprostřed? Tj. např:
      1. 1415 > 20. léta > 1423 > 1426 > 1430;
      2. 1415 > 1423 > 1426 > 20. léta > 1430;
      3. 1415 > 1423 > 20. léta > 1426 > 1430.

Jako nejvhodnější se jeví varianta 7.3.2. I u ní je však potřeba počítat s tím, že se např. při chronologickém seřazení nalezených dokladů mohou na konci časové řady objevit doklady, které z formálního hlediska (hláskosloví) patří někam jinam (např. forma „cuzie“ v prameni, který obsahuje jazykovou produkci 2 století).

* + 1. Zvolenou variantu (řadit 20. léta na konec jejich období) podporuje i rada na <http://stackoverflow.com/questions/2378873/how-do-i-store-and-sort-vague-date-ranges> (20. léta řadí jako 1410 + 10 let bez jednoho dne a dříve zmíněné reprezentativní datum by bylo na Silvestra 1419).
    2. S důsledky řazení (na konec, ale obecně s jakýmkoli) pak prostě musí uživatel počítat. Na pomoc bych mu (třeba na straně, kdyby se zobrazovaly konkordance) začátek i konec období vzniku pramene ukazoval.
    3. Pokud by to uživatel potřeboval, mohl by si třeba výsledky řadit přetažením anebo měnit strategii řazení (20. léta na začátek), ale to je už překomplikované.
  1. Aby řazení pramenů bylo stabilní (tj. sa stejným pořadím pramenů při každém seřazení), je potřeba počítat s případy, kdy bude mít stejnou dataci více pramenů (např. KatoU, PodkU, RadaOtcK, které pocházejí ze stejného rukopisu datovaného do roku 1409, nebo HusAbecV, HusBludM, HusDcerkaH, HusPostH, HusPuncta, SlovLucP, které všechny vznikly v roce 1414). Podle jakého (pomocného) kritéria se mají řadit tyto prameny? Podle názvu, zkratky pramene, žánru, počtu tokenů v prameni?
     1. Zkratka pramene vypadá dostatečně stabilně a možná i očekávaně a dá se říct, že je svým způsobem standardizovaná; a rozhodně je unikátní.
     2. Žánry nutně na rozlišení více pramenů ze stejného období nemusí stačit. Název je oproti zkratce svou formou podstatně volnější, není tak snadno rozpoznatelný a zapamatovatelný (a předvídatelný) a hlavně někdy chybí (a musí se vytvářet náhradní, který se zapisuje do hranatých závorek).
     3. Počet tokenů je v danou chvíli stabilní, ale není tak vypovídající a dobře uchopitelný („proč je který zdroj před jiným, přestože má delší slova?“).
  2. Ve StčS se používalo řešení datace, které spočívalo v rozdělení časového období na několik vrstev (A, B, C, D, E, F, G) a rámci každé vrstvy měl každý pramen přiřazeno pořadové číslo. (Toto řešení vyžaduje, aby byl předem znám repertoár pramenů a ke každému pak „ručně“ přiřazena vrstva a pořadí v ní.)
     1. To se mi nezdá jako vhodný přístup pro stále se rozšiřující kolekci textů. Nemůžu po editorech nebo správcích chtít, aby při přidávání nebo úpravě zdrojů přečíslovávali jejich pořadí, protože si už nemusejí pamatovat, jak v minulosti pořadí (arbitrárně?) určili.
     2. Výše uvedená kritéria pro řazení žádná další rozhodnutí ani údržbu nevyžadují.