Ukládání a sdílení vědeckých dat

Luděk Matyska, ÚVT MU

Výsledky vědecké práce jsou publikovány především formou monografií, článků v časopisech a příspěvků na konferencích. Je proto přirozené, že aktivity *otevřeného přístupu* jsou primárně zaměřeny tímto směrem. Články ovšem představují jen určité završení etapy odborné práce vědců, postavené nejen na předcházejících publikacích, ale využívající také další zdroje, včetně experimentálních měření, výsledků výpočtů, statistických analýz apod. Ověření správnosti publikovaných výsledků se standardně provádí opakováním (schopnost opakovat experiment a dojít ke stejným závěrům, tzv. *reproducibility*, je základním kamenem soudobé vědy), to však není možné bez přístupu ke stejným datům a měřením, jaké použili autoři publikované práce.

Při vědomí této souvislosti jsou odborné články zejména v renomovaných časopisech doprovázeny tzv. doplňkovým materiálem, který posuzovatelům a následně čtenářům přijatých článků umožňuje nahlédnout do původních dat, provést případnou nezávislou analýzu, a tak potvrdit (nebo vyvrátit) závěry publikované práce. Bohužel tento přístup již nedostačuje, protože autoři článku mohli do publikované kolekce primárních dat vybrat pouze ty údaje, které podporují jejich tvrzení, a ostatní údaje prostě „pominout". Existence těchto „pominutých" dat nemusí být snadno odhalitelná, posuzovatel či čtenář by museli zopakovat kompletně všechny experimenty, což samozřejmě není praktické a v řadě případů ani možné (např. při využití unikátních experimentálních zařízení). Potřeba přístupu k *úplným* sadám experimentálních dat, případně výsledkům rozsáhlých výpočtů se tak stává prakticky nezbytnou podmínkou pro udržení a další růst kvality publikovaných prací, zejména v prostředí standardní kontroly kvality formou peer review procesu.

Pro volný přístup k primárním experimentálním datům, laboratorním deníkům, výsledkům simulací a jiných výpočtů však hovoří další důvody. Řada vědeckých oblastí je ve stále větší míře závislá na unikátních, zpravidla extrémně drahých přístrojích. Příkladem mohou být velké urychlovače v CERNu, Fermilabu a dalších podobných pracovištích, velké genove sekvenátory schopné rychle analyzovat velký objem genetického materiálu a poskytnout sekvence nukleových kyselin, teleskopy a radioastronomická zařízení (např. Hubbleův teleskop) apod. Jejich použití vždy generuje enormní objemy dat (např. jeden běh sekvenátoru může generovat až desítky TB v jediném dni), přitom v drtivé většině případů je z těchto dat primárním vědeckým týmem, který inicioval měření, využit jen nepatrný zlomek. Pěkným příkladem může být sledování oblohy teleskopem. Přestože teleskop je například zaměřen pouze na jeden konkrétní sledovaný objekt (např. vzdálený kvasar), automaticky (jaksi „mimochodem") zabírá určitý výřez oblohy a během pozorování shromáždí data o velkém množství dalších objektů. Tato data však primárním týmem (který zaměřil teleskop na kvasar) nebudou zpracována a bez vhodného přístupu k uchování a následnému zpřístupnění toto přesné pozorování určitého úseku oblohy nebude nikým nijak využito.

Jiným příkladem mohou být rozsáhlé klimatologické studie, které není možné udělat, pokud objem dat, které má konkrétní výzkumný tým k dispozici, nepřekročí určitou kritickou velikost (i numerické simulace je nutné kalibrovat s využitím historických záznamů). Jak ukazuje aktuální únik e-mailů z britského centra výzkumu klimatu, jsou přístup primárním datům a jejich interpretace umožněny zpravidla pouze poměrně úzké skupině vědců, kteří pak mohou – často v zájmu „dobré věci" – podlehnout snaze ze svých studií vyloučit měření, která do jejich obrazu světa nezapadají a publikovat jen podpůrná data[1],

Vedle otevřeného přístupu k publikacím a je bezprostředně doprovázejícím materiálům roste

[1] Abychom byli správně pochopeni: většina vědecké práce v experimentálních oborech spočívá ve schopnosti najít ve zdánlivě neuchopitelných a chaotických primárních datech nějaký řád, a ten pak dokázat. Je však nezbytné umožnit jiným studium stejných primárních dat s možností, že najdou jiný řád a uspořádání. Špatné není ignorování některých primárních dat, špatná je pouze manipulace, která jiným vědcům znemožní přijít k jiným výsledkům.

proto zájem o zpřístupnění primárních (nezpracovaných, a tedy ani nemanipulovaných) vědeckých dat. Tento zájem především vědecké komunity nalézá pozitivní ohlas i ve veřejných agenturách, které vědecký výzkum financují – neomezený přístup k výsledkům měření jiných snižuje náklady odstraněním duplicit a zvyšuje současně efektivitu vynaložených prostředků – výsledky měření prováděných za jedním konkrétním účelem mohou být využity pro úplně jiné studie. Zde je opět exemplárním příkladem astronomie, konkrétně Sloan Digital Sky Survey (`http://www.sdss.org`). V rámci této aktivity jsou dlouhodobě (od roku 2000) shromažďovány snímky oblohy ve vysokém rozlišení, které jsou zpřístupněny široké (nejen astronomické) veřejnosti. Jak ukázala analýza publikací za rok 2004, digitální archiv SDSS představuje fakticky nejúspěšnější světovou observatoř – články publikované s využitím dat ze SDSS byly citovány o 25% více než články využívající data Evropské jižní observatoře a o 60% více než články s daty z Hubbleova teleskopu. SDSS umožňuje nalézat velmi vzácné objekty, neboť obsahuje snímky stejných částí oblohy z různých let –data, která by jinak byla uložena v nepřístupných archivech nebo dokonce smazána poté, co je primární tým využije pro své původní záměry.

V posledních letech se stále více pro činnosti související s ukládáním a zpřístupněním primárních i odvozených vědeckých dat používá termín *Data Curation* či *Digital Curation*[2]. Data Curation se věnuje udržení a zvyšování hodnoty důvěryhodné sady digitálně uložené informace (dat) tak, aby byly použitelné v současnosti i v budoucnosti [1].

Podstatnou součástí uvedené definice je pojem „důvěryhodná sada dat", protože ta jasně předpokládá kontrolu (nejlépe opět veřejnou, tj. primárně vědeckou komunitou) kvality dat. Zpřístupnění v budoucnosti pak vyžaduje, aby primární data doprovázel jejich popis ve formě metadat, která často zahrnují i *provenanci*, tedy původ dat a veškeré předcházející manipulace s nimi. Data Curation také musí zajistit průběžnou transformaci dat do nových formátů, aby data byla přístupná (čitelná) bez ohledu na vývoj informačních technologií (přitom musí být zaručeno, že tyto transformace nemění původní obsah – riziko vzniku artefaktů může nastat např. u obrazové informace při nevhodné volbě nových kompresních formátů).

Zajímavým problémem je rovněž anonymizace dat, zvlášť důležitá v lékařských či sociologických výzkumech – primární data nemohou být zpřístupněna sama o sobě, ale až poté, co jsou z nich odstraněny odkazy, které by mohly vést ke zcela konkrétním osobám. Nedílnou součástí Data Curation je i rozhodování o tom, kdy data ztrácejí hodnotu a je možné je nenávratně odstranit. Na rozdíl od knih, časopisů a dalších vědeckých publikací, které se snažíme uchovat trvale, je zde schopnost „ořezávání" (pruning) kritickou a nezbytnou součástí práce s vědeckými daty. Posouzení dlouhodobé vědecké hodnoty uložených dat se musí postupně stát součástí přípravy nových vědců, protože jinak hrozí na jedné straně předčasná ztráta dat (současná situace ve velké většině disciplín), na druhé straně zavalení irelevantními a dále nepoužitelnými daty.

Primární data a procesy data curation budou stále významněji ovlivňovat budoucí vědeckou práci. Umožní její lepší kontrolu, přispějí je zvýšení efektivity prostředků vložených do výzkumu (řada zemí začíná požadovat plné zpřístupnění dat, které byly získány v rámci realizace projektů hrazených z veřejných rozpočtů) a povedou k rozvoji úplně nových metod vědecké práce (např. data mining dlouhých časových řad, jejichž shromáždění jde nad možnosti i zájem jednotlivých úzce zaměřených výzkumných týmů). Vedle otevřeného přístupu k publikacím se tedy musíme připravit i na otevřený přístup k našim primárním datům.

A na závěr jedno ohlédnutí: Protokol http, základ světového webu, byl původně vynalezen v CERNu jako technický nástroj pro sdílení publikací, primárních experimentálních dat i výsledků výpočtů co nejširší vědeckou komunitou částicových fyziků. Současný Internet tak za podstatnou část svého úspěchu vděčí potřebě vědecké komunity sdílet data a současně nežádat ochranu intelek-

[2] Viz. např. časopis *International Journal of Digital Curation*, UKOLN„ University of Bath, UK, `http://ijdc.net`

tuálních práv na software, který sdílení umožňuje.

Literatura

[1] N. Beagrie, *Digital Curation for Science, Digital Libraries, anmd Individuals*, IJDC, Vol 1(1), 3–16, 2006, http://ijdc.net/index.php/ijdc/article/viewFile/6/2. □