Vratislav Podzimek: 1. ročník navazujícího magisterského studia
05.12.2011

PARSE.Insight

Permanent Access to the Records of Science in Europe.Insight
`http://www.parse-insight.eu/`[1]
Projekt Evropské unie jako součást 7. rámcového programu.
květen 2008 – únor 2010

Charakteristika

Jak už hlavní část samotného názvu (trvalý přístup k záznamům vědy v Evropě) napovídá, byl tento projekt zaměřen na problematiku permanentního přístupu k vědeckým datům. Množství dat z výzkumů neustále narůstá a každým rokem je tento růst rychlejší než v letech předchozích. A protože jedním ze základních principů umožňujícím rychlý vývoj vědy je to, aby každá další generace vědců mohla stavět na výsledcích svých předchůdců, je nutné tato data dlouhodobě uchovávat. A to jak surová data získaná přímo z měření a průzkumů, tak i závěry vyvozené na základě jejich zpracování. Jinak hrozí riziko, že budou tyto výsledky ztraceny nebo přinejmenším stěží verifikovatelné.

Druhá část názvu – Insight (neboli „vhled“) zase poukazuje na to, že hlavními postupy využívanými v projektu byly průzkumy mezi zástupci stran zainteresovaných v dané problematice – tzn. vědeckých organizací, organizací zajišťujících archivaci dat, digitálních knihoven a dalších. Tento projekt probíhal pod záštitou *Alliance for Permanent Access to the Records of Science* (aliance pro trvalý přístup k vědeckým záznamům) a má silnou návaznost na ostatních projekty této aliance – CASPAR, ODE, APARSEN, SCIDIP-ES (viz [2])

Cíle

Projekt PARSE.Insight byl z pohledu cílů poněkud netradičním. I když na tento projekt bývá velmi často odkazováno pouze zkratkou PARSE, právě druhá část názvu poukazuje na skutečné cíle projektu. Těmi totiž nebylo

zajištění permanentního přístupu k záznamům vědy v Evropě, ale získání „vhledu“ do problematiky. Tedy analýza současného stavu této oblasti, označení nejproblematičtějších bodů a návrh všeobecných postupů, kterými by se dala situace vylepšit.

Hlavním výstupem tohoto projektu je plán na vytvoření infrastruktury pro zachovávání a zpřístupňování vědeckých dat, na který by ostatní projekty měly navázat implementací a následným nasazením v praxi. Tento plán sdružuje pohledy na danou problematiku z hlediska zájmů národních, evropských i celosvětových a jeho porovnáním se současným stavem byly zjištěny nejproblematičtější oblasti, jež je třeba vylepšit. Projekt však počítá i s dopady permanentního přístupu k vědeckým datům, a proto bylo dalším cílem specifikovat nástroje, kterými bude možné míru těchto dopadů posuzovat. Posledním výstupem projektu pak byla metodika na posuzovaní správnosti postupů jednotlivých organizací na základě výkonnosti jejich repozitářů.

Výsledky projektu PARSE.Insight by měly sloužit jako návod pro strategii Evropské komise v oblasti infrastruktury výzkumu.

Popis projektu a jeho výsledků

V první fázi projektu vznikla první předběžná verze plánu pro vytvoření infrastruktury a byly vybrány tyto oblasti (a organizace) pro průzkumy a konzultace – případové studie:

- fyzika vysokých energií (částicová fyzika) – CERN,
- pozorování Země – ESA,
- oblast sociálních a humanitních věd.

V následující fázi probíhaly průzkumy mezi členy výše uvedených organizací a dalšími jedinci působícími v daných oblastech zaměřené na problematiku permanentního uchování dat, jejich přístupnosti a otevřenosti tohoto přístupu. Navíc začaly aktivity spojené s vytvořením metrik pro měření dopadu permanentního přístupu.

Podívejme se nyní pro příklad na některé konkrétní výsledky průzkumu mezi fyziky z organizace CERN. Je patrné, že naprostá většina z nich připisuje velký význam permanentnímu uchování dat. Jak teoretičtí, tak i experimentální fyzikové se shodují (i když někdy v různé míře) na důležitosti takových dat pro budoucí verifikaci utvořených závěrů, možnosti jejich kombinace s daty získanými při budoucích měřeních a verifikaci budoucích teorií

pomocí těchto dat. Asi polovina z dotázaných je přesvědčena, že by jim již dříve získaná data pomohla při jejich výzkumu, a ještě větší počet z nich si myslí, že v oblasti fyziky vysokých energií již byla některá podstatná data ztracena. Podíváme-li se na výsledky k dotazům na druhy dat, která by se měla uchovávat, je zřejmé, že problém správné interpretace a znovupoužitelnosti surových dat způsobuje nižší důležitost uchovávání takových dat a to i v případě jejich doplnění o nástroje pro jejich interpretaci. Naopak je tomu např. u publikovaných dat a podkladových dat pro publikace. Za poněkud problematický závěr se dá považovat to, že jen minimální počet fyziků by data chránil proti ztrátě okamžitě po jejich získání. Naproti tomu za ideální okamžik považují publikování článků vycházejících z těchto dat, případně konce projektů, ve kterých byla data získána. Tyto výsledky se téměř přesně shodují s výsledky pro dotaz, kdy by byli vědci ochotni data uvolnit. Dále se ve velké míře shodují na tom, že by je úprava dat pro trvalé uchování stála asi 10–50 % úsilí navíc ve srovnání s jejich pořízením a zpracováním. V oblasti autorství pak fyzikové vidí největší problémy v možnosti neodkazování na původní autory. Silné zastoupení mají obavy z nesprávného používání a nárůstu počtu nekorektních závěrů a výsledků. Zatímco fyzikové mají v mnoha případech zájem o data z jiných výzkumů a vidí problém ve ztrátě takových dat, jejich ochota k přispění k permanentnímu uchování dat je mnohem nižší. Často je to však způsobeno tím, že nemají představu o prostředcích, které současný stav technologií, práva a infrastruktury nabízí. Na základě takových zjištění bylo usouzeno, že je nutné připravit materiály pro školení vědců v oblasti prostředků pro permanentní uchování vědeckých dat. [3]

Ve třetí fázi probíhala analýza dopadu permanentního uchování dat, počáteční analýza rozdílů mezi plánem a současným stavem a úprava prvotního plánu. Poslední dvě fáze byly zaměřeny na posouzení rozdílů mezi plánem a současným stavem, specifikaci nástroje pro analýzu dopadu permanentního uchování dat a testování této analýzy. V závěru projektu byly vydány zprávy o jednotlivých oblastech, které jsou nyní dostupné na webu [4] společně s dílčími zprávami z jednotlivých průzkumů.

Výsledný plán identifikuje největší hrozby a zahrnuje závěrečná doporučení, [3] ve kterých oblastech a co je třeba udělat pro to, aby pokrok ve vědě mohl i nadále vycházet z principu, kdy nová generace výzkumníků staví na poznatcích a závěrech svých předchůdců. A to tak, že budou schopni jejich závěry verifikovat, upravovat a v případě, že budou ve shodě s dalšími novými objevy, z nich dále vycházet.

Podívejme se na příklady největších hrozeb a doporučení, jak je potlačit. Chceme-li zamezit tomu, aby byla data v budoucnu nepřístupná kvůli neudržování potřebného hardwarového a softwarového vybavení, musíme zajistit sdílení informací o jeho dostupnosti nebo o jeho případných náhradách. Dále budeme muset zajistit možnost spojování důkazů o autenticitě digitálních dat z různých zdrojů, protože určitě dojde k přerušení řetězce důkazů autenticity. Klíčové bude také správné nastavení tzv. digitálních práv (Digital Rights), abychom v budoucnu zabránili porušování omezení přístupu stanovených autorem. Abychom věděli, kterým institucím důvěřovat, že digitální data opravdu budou udržovat a nedovolí jejich zánik, bude muset vzniknout certifikační infrastruktura, která bude po důkladné kontrole důvěryhodným institucím udělovat certifikáty. Pro kompletní přehled viz tabulka *Threat/Requirement for solution.* [3]

Dále se výsledný plán věnuje problematice financování, neboť trvalé udržování dat, stejně jako jejich příprava do podoby vhodné pro tyto účely, přinese zvýšení finančních nároků každého projektu. Jedním z navrhovaných řešení je využití reklamy a to především u dat, která budou atraktivní pro velký počet uživatelů (např. snímky Země). Na problematiku financí navazuje i další část závěrečného plánu, jež poukazuje na nutnost dostatečné motivace vědců a vědeckých organizací k trvalému udržování a zpřístupňování dat. Jak totiž průzkumy ukázaly, mnoho vědců by při své práci rádo využilo již dříve získaná data, avšak ta jsou často buď ztracená, nebo z jiných důvodů nedostupná. Na druhou stranu ze stejných průzkumů vyplývá skutečnost, že vědci nejsou příliš ochotni svá data poskytovat ostatním či je připravovat k permanentní dostupnosti. Příčin tohoto stavu je hned několik – obava ze špatného použití dat, právní problémy (zneužití anonymních průzkumů apod.), technické problémy a další. Se všemi těmito příčinami bude nutné se vypořádat, chceme-li, aby vědci svá data zpřístupňovali ostatním a snažili se o jejich trvalé uchování. Stále to však bude znamenat vynaložení dodatečného úsilí, a proto se neobejdeme ani bez nějakých striktnějších motivačních (v tomto případě spíše donucovacích) prostředků. Jako např. granty podmíněné právě zveřejněním dat a zajištěním trvalého přístupu k nim. Všechny zainteresované strany kromě vědců však budou muset přejít k praktikám, kdy budou moci spolupracujícím vědcům nabídnout výhodné podmínky v případě jejich spolupráce, a naopak nevýhodné podmínky v případě, kdy vědci ochotni spolupracovat nebudou. Pro přehled jednotlivých stran a jejich výhodných a nevýhodných podmínek viz tabulka *Stakeholder/Can offer carrots/Can offer sticks ...* [3]

Ani těmito doporučeními výsledný plán projektu PARSE.Insight nekončí, ale bylo by zbytečné probírat zde všechna. I na výše uvedených příkladech je dobře patrné, že projekt splnil svůj cíl a na základě průzkumů současné situace vytyčil největší problémy, k nimž nabídl principy jejich řešení.

Vlastní zhodnocení

Původním smyslem znepřístupňování vědeckých dat byla podpora vynalézání a vůbec tvorby něčeho nového. Když se však zamyslíme nad současným stavem světa vědy (a to jak přírodních věd, tak i humanitních) a techniky, snadno dospějeme ke zjištění, že drtivá většina nových objevů a vynálezů vzniká díky práci mnohočlenných týmů. V nezanedbatelném počtu případů se dokonce jedná o celé komunity, které se pohybují i v řádech několika set tisíc až miliónů jedinců z celého světa. Příkladem zde budiž veškerý software vyvíjený podle stanov *Free Software Foundation* [5] a mnoho projektů přenášejících tyto principy do jiných oblastí. Vzniká tak například automobil budoucnosti, na jehož vývoji se podílí dobrovolníci z celého světa. [6] Stejného přístupu využívá i mnoho tzv. distribuovaných výpočetních projektů, které umožňují dobrovolníkům z celého světa připojit své počítače do velké sítě strojů, jejichž výpočetní výkon je následně využíván k vědeckým výpočtům. Výsledky takových výpočtů jsou pak použity např. při výzkumu fyzikálních zákonů nebo v boji proti různým chorobám. Tím se dostávám k další zajímavé skutečnosti. Např. projekt *Foldit* [7] totiž využívá kromě distribuovaného počítání na strojích velkého počtu dobrovolníků i tzv. „crowdsourcing“ (do češtiny by se tento pojem snad dal přeložit jako „využívání síly davu“). Jedná se o metodu, kdy je úloha (např. vyřešení nějakého problému) zprostředkována velkému množství lidí. V případě Foldit se jedná o hru, která je simulací problému skládání molekul bílkovin a kterou mohou hrát lidé na celém světě. Je pozoruhodné, že prostřednictvím této hry dokázali lidé během 3 týdnů vyřešit problém, který se vědcům po celém světě nepodařilo vyřešit za 15 let a to ani pomocí nejvýkonnějších počítačů. [8]

A proč zde tyto přístupy a jejich dopady uvádím? Velmi dobře totiž vystihují, že díky moderním technologiím jsme dnes schopni využívat otevřený přístup k výzkumu (a v něm shromažďovaným datům) k dosažení značného pokroku vědy a techniky. Kdyby totiž na problému skládání molekul bílkovin pracoval jeden odborník, jen stěží by se mu podařilo najít řešení během jeho života. Vezmeme-li v potaz větší tým vědců, stále je jen mizivá šance, že by se jim podařilo najít řešení dříve, než by tomuto výzkumu vypršely granty a

další nezbytné podklady. I když uvážíme obrovské množství počítačů po celém světě (a tedy výpočetní výkon daleko přesahující nejlepší superpočítače), stále trvá vyřešení problému velmi dlouhou dobu, protože umělá inteligence je nesrovnatelná s inteligencí lidskou, jež zahrnuje i intuici a další jedinečné principy. A tak se nyní dostáváme do stavu, kdy se objevují zcela nové otázky, např.: Jak určíme autorství u objevu, jenž byl učiněn na základě snahy tisíců lidí po celém světě, když navíc neznáme jejich identitu? A byl by tento objev vůbec možný, kdyby např. autorská práva zamezila přístupu těchto lidí k potřebným datům?

Nemusí to však být pouze autorská práva, která dokáží zamezit přístupu k datům. Žijeme v době, kdy se technologie mění tak rychle, že nejsme schopni získat a/nebo správně interpretovat data získaná před několika málo desetiletími, a vše nasvědčuje tomu, že se tento trend bude dále rozvíjet. Je tedy jasné, že pouhé zpřístupnění dat (z hlediska autorsko-právního) nestačí. Není možné všechna data zpracovávat tak rychle, jak vznikají, a tedy pokud bychom chtěli využít např. „crowdsourcingu“, museli bychom zajistit, aby byla data dostupná ještě dlouho po jejich vzniku.

Proto je určitě rozumné podporovat snahy v oblasti problematiky trvalého přístupu k vědeckým datům. Jednou z takových snah byl právě i projekt PARSE.Insight a i když se na první pohled může zdát, že výstupy tohoto projektu téměř nemohou být prospěšné, je třeba si uvědomit, kdy tento projekt vznikl a probíhal, a podívat se na projekty po něm následující a stavějící na jeho závěrech. Mohli bychom i namítat, že projekt trval velmi dlouho vzhledem ke skutečnosti, že závěry, které přinesl, nejsou příliš překvapující a většinu z nich bylo možné odhadnout předem. Pokud to ale se snahami o zajištění trvalého přístupu k vědeckým datům myslíme vážně, musíme znát spolehlivé údaje o současném stavu, stejně tak jako je nutné přesně znát cíle a nástrahy, které stojí v cestě k těmto cílům.

Reference

[1] PARSE.Insight Project. *Domovská stránka projektu PARSE.Insight* [online], [cit. 05.11.2011]. Dostupné z:

<http://www.parse-insight.eu/>

[2] členové Alliance for Permanent Access. *Projekty spadající pod Alliance for Permanent Access* [online], [cit. 05.11.2011]. Dostupné z:

<http://www.alliancepermanentaccess.org/index.php/current-projects/>

[3] PARSE.Insight consortium. *Závěrečná podoba plánu z projektu PARSE.Insight* [PDF] 05.06.2010, [cit. 05.11.2011]. Dostupné z:

<http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf>

[4] PARSE.Insight Project. *Publikace projektu PARSE.Insight* [online], [cit. 07.11.2011]. Dostupné z:

<http://www.parse-insight.eu/publications.php>

[5] Free Software Foundation, Inc. *Domovské stránky Free Software Foundation* [online], [cit. 07.11.2011]. Dostupné z:

<http://www.fsf.org/>

[6] OScar Project. *Domovské stránky projektu OScar* [online], [cit. 07.11.2011]. Dostupné z:

<http://www.theoscarproject.org/>

[7] Foldit Project. *Domovské stránky projektu Foldit* [online], [cit. 07.11.2011]. Dostupné z:

<http://fold.it/portal/>

[8] Coren, Michael J., Fast Company. *Článek v časopisu Scientific American o projektu Foldit* [online] 20.08.2011, [cit. 07.11.2011]. Dostupné z:

<http://www.scientificamerican.com/article.cfm?id=foldit-gamers-solve-riddle>