Masarykova univerzita v Brně

Filozofická fakulta

Ústav české literatury a knihovnictví

Kabinet informačních studií a knihovnictví

****

**SOFTWARE Heritage ARCHIVE**

 Esej k předmětu PV070

Autor: Jan Plíšek

UČO: 100488

Typ studia: Kombinované

Ročník: 3. Ročník

Počet znaků: 6538

**Brno 2018**

Software Heritage (Softwarové dědictví) reaguje na situaci, kdy na světě neustále vzniká nový a nový software, ale neřeší se otázka dostupnosti toho starého. Technologický pokrok jde stále dopředu a dokonalejší hardware vyžaduje sofistikovanější software. Stejně tak i služby poskytované ve virtuálním prostředí internetu musí odpovídat nárokům doby. Archivy Software Heritage se stávají možným zdrojem pro zkoumání starších volných kódů. Nearchivují se jen samotné finální programy ale i celá historie v různých verzích. Software Heritage je iniciativa s počátkem v roce 2015, která si klade za cíl zachránit a pro budoucnost uchovat nejen samotný software ale i původní zdrojové kódy. Zaměřuje se na takový software, který je free nebo open source. Open Source software (tzv. otevřený software) je definován jako program, který je poskytován s otevřeným zdrojovým kódem, ale stále pod nějakým druhem licence. Ta pak určuje, do jaké míry lze s produktem nakládat. Licence je důležitá, protože omezuje možnosti dalšího zpracování. Existují licence, které jsou naprosto neomezující (např. GNU – kód lze dále distribuovat i měnit), nicméně výběr licence je na autorovi. Je více druhů otevřeného přístupu podle toho, jakým způsobem jsou informace publikovány. Svobodný software je svobodný (volný, libre) z pohledu nakládání, ale nutně nemusí být zdarma.

Projekt Software Heritage začal v roce 2016 na půdě francouzského institutu pro výzkum informatiky a automatizace (Institut National de Recherche en Informatique et en Automatique neboli INRIA) a v úvodním prohlášení se jeho autoři označují jako „alexandrijská knihovna softwaru“ (INRIA, 2016). Institut poskytl základní infrastrukturu pro spuštění archivu a prvotní finance do rozpočtu na první tři roky ve výši 500.000,- EUR (Jost, 2016). Další finance by měly být získávány ze sbírek, darů nebo dotací. Počítá se také s podporou velkých firem z oblasti informačních technologií. Prvním hlavním partnerem byl na počátku Microsoft, který poskytl technické prostředky ve formě cloudového prostředí Microsoft Azure (Margaux, 2017). Ačkoliv projekt byl slavnostně prezentován právě v červnu 2016, již v této době disponovaly archivy projektu Software Heritage 22,8 milióny softwarových projektů a 2,7 miliardami verzí zdrojových kódů (Jost, 2016). Začátkem roku 2018 je již počet projektů 83,8 miliónů a 4,4 miliard zdrojových kódů (Cosmo, 2018). Hrubý aktuální stav a graf dosavadního vývoje znázorňuje obrázek číslo 1.



Obrázek č.1: Vývoj archivu Software Heritage v čase (Inria, 2018)

Důležitost projektu si uvědomují nejen národní vědecké instituce jako je francouzská INRIA, ale také giganti z oboru informačních technologií a z businessu v čele s bankou Societe Generale, softwarovou firmou Microsoft nebo producentem čipů Intel.

Na úvodní stránce https://www.softwareheritage.org/ si každý může otestovat, zda nějaký konkrétní software již v databázi je. Zároveň je u okna vyhledávače i online počítadlo s aktuálními počty záznamů. Stručný a názorný příklad ulehčuje použití vyhledávacího rozhraní. Základním pilířem archivu jsou sdílené databáze již existujících projektů a firem. Takovými spolupracovníky jsou například GitHub, GitLab, debian nebo i Google.

Samotné uchování softwaru by bylo pouze částečnou službou. Aby užitek byl maximální, musí být tento kód také sdílen. Miliardy verzí jsou tedy indexovány, organizovány a spravovány systémy kontroly verzí. Rozhraní pro vyhledávání je nastaveno na uživatelsky přívětivé úrovni a lze vyhledávat i podle částí kódů. Databáze mohou používat i vědci, kdy bez velkých překážek získají přístup nejen k volným vědeckým datům, ale také k softwaru, na kterém byla data zpracovávána. Opakovatelnost experimentů je důležitá pro věrohodnost jejich výsledků, a proto dostupnost původního softwaru výzkumnou práci usnadňuje. S delším časovým odstupem ale může uživatel narážet na hardwarová omezení, kdy novější počítače nejsou kompatibilní se starým programem.

Osvětová činnost Software Heritage se snaží o budování povědomí nejen v odborné komunitě o tom, že software není jen to, co nám spouští počítač, ovládá chytrý telefon nebo hlídá jídlo v ledničce. Je důležité, aby se zdrojové kódy uchovávaly pro budoucí generace. Projekt softwarového dědictví zachraňuje i kódy již zapomenuté, aktivně vyhledává kódy ještě nearchivované a vytváří podmínky pro jejich uložení. Podporovatelé poskytují prostor pro ukládání dat prostřednictvím zrcadlení na více místech, aby se snížilo nebezpečí ztráty dat.

V archivech jsou ukládány tyto typy dat (INRIA, 2018):

Obsah – konkrétní soubor s programem, který je uložen pod jménem vytvořeným na základě obsahového kontextu

Složka – ve složce jsou umístěny všechny atributy související se souborem a také samotný soubor

Revize – snapshot obsahu složky včetně souvisejících metadat vzniklých v čase vzniku snímku (například. logy, autor, timestamp)

Release – revize označovaná jako mimořádná s vlastním jménem (nebo jen specifickým pořadovým číslem) doplněná o vývojová metadata (např. timestamp nebo autor).

Zdroj – internetová cesta (URL), odkud byla všechna koherentní data objektu archivovaného v Software Heritage přijata (obsah, revize, release)

Návštěvy (visits) – mapované úspěšné pokusy o stažení doposud neuloženého softwaru, jsou to všechna potřebná data v okamžiku stažení (aktuální metadata, release apod.), každá návštěva ukládá nový záznam, ale starší se nemaže

Osoba (person) – záznam o autorovi nebo člověku, který prováděl změny, ve formě plného jména a případně i email

Určitě si autoři projektu na záchranu volného softwaru nedělají iluze, že by se jim podařilo obsáhnout veškerou světovou produkci, ale z jejich jednání je zřejmé, že se o to budou i nadále snažit. Nesměřují ale jen k pouhému skladu informací, postupně se doplňují a vylepšují vyhledávací a prohlížecí nástroje. Přesnější metadata dovolí lépe trackovat minulost softwaru daleko před čas vložení do archivu Software Heritage. Full textové vyhledávání pomáhá nalézat programy i podle částí zdrojového kódu. Veřejnost má tedy v ruce mocnou službu, která poskytuje podklady pro výuku, studium nebo výzkum volně použitelného softwaru. Pokud hledaný program chybí, může každý také navrhovat jiné potenciální zdroje k rozšíření databáze (Brown, 2016).

Software Heritage plní roli zachránce dědictví v podobě programů a uchovává je pro další využití. V dnešní informační době (společnosti) je nutné mít představu nejen o současných a používaných softwarech, ale je nutné vědět, odkud jsme vyšli, abychom snáze vytvářeli cesty nové.

**Zdroje:**

ABRAMATIC, Jean-Francois a spol., 2018. *Building the Universal Archive of Source Code* [online]. Communications of the ACM [cit. 27.11.2018]. Dostupné z: <https://doi.org/10.1145/3183558>

BROWN, Paul, 2016. Software Heritage: Creating a safe haven for software [online]. Boing Boing [cit. 29.11.2018]. Dostupné z: https://boingboing.net/2016/06/30/software-heritage-creating-a.html#

COSMO, Roberto di a spol., 2018. *Identifiers for Digital Objects: the Case of Software Source Code Preservation* [online]. HAL archives-ouvertes [cit. 27.11.2018]. Dostupné z: <https://hal.archives-ouvertes.fr/hal-01865790/file/main.pdf>

HAVLOVÁ, Jaroslava, MAREK, Jiří, 2003. *Otevřený přístup (k vědeckým informacím).* In: KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Praha : Národní knihovna ČR, [cit. 2018-11-27]. Dostupné z: <http://aleph.nkp.cz/F/?func=direct&doc_number=000015817&local_base=KTD>.

INRIA, 2016. *Collect, organise, preserve and share the Software Heritage of mankind*[online]. INRIA [cit.25.11.2018]. Dostupné z: <https://www.softwareheritage.org/wp-content/uploads/2016/06/PressReleasePressKit-2016-06-30.en_.pdf>

INRIA, 2018. *The Software Heritage Archive* [online]. INRIA [cit. 27.11.2018]. Dostupné z: <https://www.softwareheritage.org/archive/>

JOST, Clémence, 2016. *Open source : lancement de Software Heritage, la plus grande bibliothèque de codes source de la planète* [online]. Archmag [cit. 26.11.2018]. Dostupné z: <https://www.archimag.com/demat-cloud/2016/07/01/open-source-software-heritage-archive-codes-source>

Margaux, Allain, 2017. *Avec Software Heritage, Microsoft met son Cloud au service de la préservation du patrimoine logiciel* [online]. Hub Presse [cit. 30.11.2018]. Dostupné z: <https://news.microsoft.com/fr-fr/2017/04/03/projet-software-heritage-dont-microsoft-partenaire-fondateur-prend-nouvelle-dimension-signature-dune-convention-lunesco-faveur-de-preservation-p/>