Masarykova Univerzita

Projekt LOCKSS

Tomáš Svoboda, 396475
3. semestr NMgr.

URL projektu
http://www.lockss.org/

9. prosince 2015

Úvod

Program LOCKSS - Lots Of Copies Keep Stuff Safe, neboli mnoho kopií udrží věci v bezpečí - je open source projekt vytvořený na Stanfordské univerzitě. Hlavním účelem tohoto projektu je dlouhodobé uchovávání a poskytování digitálního obsahu. Tento projekt je postavený na myšlence redundantních dat, tj. každý digitální dokument je uložen v několika separátních repozitářích a v případě, že některý z nich není dostupný, obsah je čtenáři předán ze zbývajících dostupných zdrojů. K zabezpečení redundance a dostupnosti digitálního obsahu využívá peer-to-peer síť.

Projekt LOCKSS byl založen v roce 1999 a do roku 2002 byla úspěšně nasazena jeho zkušební verze v 50 knihovnách po celém světě. V roce 2004 byla uvedena na trh jeho finální podoba, na které se podílely vedle Standfordské univerzity i veřejná knihovna v New Yorku a knihovny z dalších univerzit. Za léta svého působení získal projekt LOCKSS několik ocenění od nejrůznějších organizací. V současnosti se na jeho financování a podpoře podílí nespočet organizací, knihoven a vědeckých pracovišť, jako např. Harvardská univerzita, které jsou zapojeny do tzv. aliance LOCKSS. Do projektu se k dnešnímu dni zapojilo více než 500 vydavatelů a obsahuje přes 9000 digitálních dokumentů. Minimální nároky systému na hardware nejsou nikterak velké a běžný počítač je spolehlivě splní. Asi nejzásadnější komponentou je diskové pole. Je doporučováno, aby mělo minimální kapacitu 4TB. Takováto sestava je dostupná přibližně za 2000 liber.

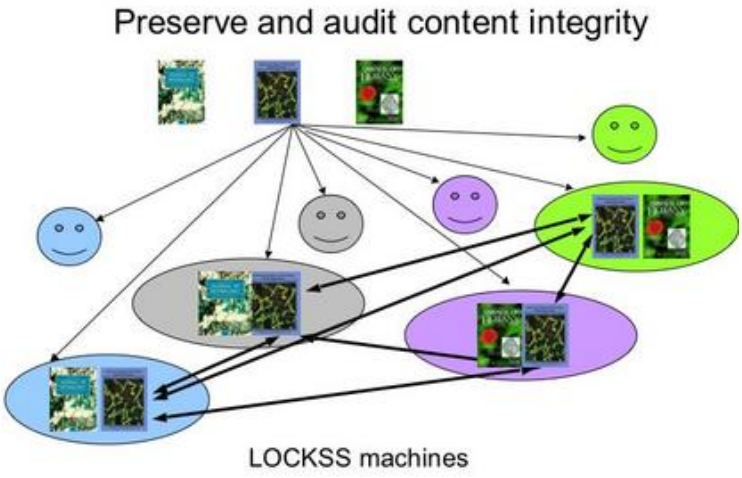
1 Cíle projektu

Projekt LOCKSS si klade za cíl uchovávat a následně zprostředkovávat uložené dokumenty v elektronické podobě po velmi dlouhou dobu. Tento cíl vychází z klasického knihovního modelu, kdy si knihovna zakoupí publikaci v tištěné podobě a poté ji má k dispozici na neurčito. Jelikož je pravděpodobné, že si tu stejnou publikaci zakoupí i další knihovny, je patrné, že bude současně existovat několik kopií té stejné publikace a tedy je i velká šance, že i po velmi dlouhé době bude v některé knihovně existovat její výtisk a nebude tak ztracena. S tím souvisí i druhý cíl, kdy má být každá taková elektronická publikace dostupná i v případě, že již není z nějakého důvodu dostupná na straně vydavatele - např. že vydavatel zanikl nebo vypršelo předplatné dané publikace.

2 Principy uchování dokumentů

Projekt LOCKSS staví na myšlence, že žádná knihovna nemůže sama zabezpečit spolehlivé a dlouhodobé uchovávání dokumentů. Systém uchovávání dokumentů u tohoto projektu je postaven na několika principech, které byly vyvíjeny během dlouholetého výzkumu. Mezi tyto principy patří především decentralizované a distribuované uchovávání dokumentů. Toho je docíleno tím, že systém je vybudován na peer-to-peer síti, ve které se vyskytuje několik kopií stejného dokumentu na různých uzlech sítě. To zaručí spolehlivé uchování dokumentů. S touto vlastností systému souvisí ještě jeden fakt, a to nemožnost jakkoliv neoprávněně manipulovat a upravovat obsah těchto dokumentů, protože každý z nich je distribuován na uzlech po celém světě, které jsou navíc pod správou různých administrátorů.

Co tedy tvoří z velké části tyto distibuované uzly? Tvoří je právě lokální repozitáře - tzv. LOCKSS box - vlastněné knihovnami začleněnými do tohoto projektu. To zaručuje mimo jiné i kontrolovaný přístup k vlastněným publikacím a odpadá tak riziko spojené se ztrátou přístupu k článkům umístěných u cizích poskytovatelů. Tento princip vychází z tradičního modelu purchase-and-own, tj. zaplať a vlastni, který se využívá v knihovnách pro tištěné publikace.



Obrázek 1: Vztah mezi jednotlivými LOCKSS boxy při kontrole integrity.

Systém pro uchovávání obsahu je specifický tím, že čtenářům poskytuje originální publikaci, včetně značkování a vzhledu, který vytvořil její autor. Odpadá tak problém, kdy existuje několik podob jednoho dokumentu na různých stránkách. To, že dochází k vystavování originální publikace je i jistý způsob, jak může mít čtenář jistotu, že přistupuje k důvěryhodné a hlavně nejpřesnější verzi dokumentu. Vedle toho vydavatelé mají i možnost monitorování přístupu ke svým článkům, včetně zdroje, odkud čtenář k dokumentu přistupuje.

Velmi zásadním problémem, ohrožujícím jakýkoliv systém pro dlouhodobé uchovávání dokumentů, je ekonomika a tedy i financování poskytovaných služeb. LOCKSS tento problém řeší tím způsobem, že svoje služby poskytuje vydavatelům publikací zdarma a až knihovny platí nepatrný poplatek, který slouží k pokrytí provozních nákladů. Tyto poplatky jsou však natolik nízké, že využívání systému LOCKSS je méně finančně náročné, než provoz a údržba vlastního systému pro zálohu a monitorování dokumentů. Výše poplatků se dále dělí podle velikosti dané instituce a to přibližně v rozmezí 2000 až 12000 dolarů v USA nebo 1800 až 5000 liber v Británii ročně. V ostatních zemích je výše poplatků dostupná na vyžádání.

3 Jak funguje systém LOCKSS

Aby mohlo dojít k uložení publikace, musí se provést několik kroků. Nejprve musí dát vydavatel svolení se zařazením dokumentu do tohoto systému. Dále musí daná knihovna zřídit autorizovaný přístup pro LOCKSS box k tomuto dokumentu a v neposlední řadě musí být tento LOCKSS box registrovaný u jedné sítě LOCKSS aliance.

Jakým způsobem jsou data získávána? Na počátku musí správci projektu LOCKSS provést analýzu struktury URL obsahu, jeho dostupné formáty a způsob jeho získání ze stránek vydavatele. Po té analýze navrhnou plán provedení pro uchování obsahu, který je specifický pro každý spravovaný obsah. Samotné získávání dat poté probíhá pomocí speciálního webového crawleru. Samotný přístup k obsahu umožňuje vydavatel pomocí speciální webové stránky, obsahující tzv. manifest. Tento dokument obsahuje prohlášení o přístupu a slouží tak jako obdoba smlouvy mezi vydavatelem a systémem LOCKSS. Vedle prohlášení obsahuje i odkazy na jednotlivé části publikovaného obsahu. Pro usnadnění přístupu k obsahu jednotlivých vydavatelů využívá LOCKSS box moduly, které obsahují informaci o poloze jednotlivých manifestů. Aby nemohlo dojít k neautorizovanému přístupu k publikacím, na straně vydavatele dochází ke kontrole IP adres, ze kterých přistupují

jednotlivé LOCKSS boxy.

Poté, co je do systému začleněn v rámci předchozího kroku nový obsah, dochází u něj k pravidelnému monitorování změn, aby se zamezilo jeho poškození. Během tohoto monitorování dochází k porovnávání každé kopie dat mezi lokální instancí a ostatními LOCKSS boxy. Po každém dokončení této operace je zaručeno, že každá instance obsahuje správná data, která reprezentují původní obsah. Pokud došlo v průběhu monitorování ke zjištění, že některá data jsou poškozená, dojde k jejich opravě na základě dat uložených v některé vzdálené instanci LOCKSS boxu. Aby mohla být zaručena alespoň minimální integrita dokumentu a tedy i systému, je doporučováno, aby měl každý dokument alespoň 7 kopií. Pro srovnání - v roce 2013 obsahovala globální síť LOCKSS přes 150 distribuovaných uzlů a medián počtu kopií jednotlivých publikací se pohyboval okolo 25. Jak dosáhnout minimálního počtu kopií? U populárních publikací je to snadné, ale u méně populárních to může být problém. Jako první možnost se jeví přesvědčit nějakou další instituci zapojenou do projektu, aby si také pořídila danou publikaci, například pokud by spolu měli dohodu o vzájemné výpomoci v takovýchto situacích nebo se jedná o speciální publikaci, kterou mají povinnost na základě členství v alianci uložit. Pokud by se jim toto nepovedlo, mohou si buď individuálně nebo hromadně pořídit další repozitář, který by uchovával novou kopii dokumentu. Jelikož je ale minimální počet kopií tak nízký a počet distribuovaných uzlů tak vysoký, tato situace se objevuje jen zřítka a u velice specifických publikací.

Způsob servírování uchovávaného obsahu čtenářům je specifický tím, že představuje neviditelnou mezivrstvu mezi čtenářem a publikací. Pokud přijde požadavek systému na zobrazení dokumentu, systém se pokusí poskytnout obsah na jeho originálním umístění u vydavatele. Pokud je toto nemožné - např. kvůli přetížení serveru nebo vypršení předplatného, dojde k zobrazení kopie tohoto dokumentu, která je uložena a spravována systémem LOCKSS box.

V rámci systému existují tři hlavní způsoby, jak mohou čtenáři přistoupit k dokumentu. Prvním z nich je tzv. proxy. Při tomto způsobu jsou veškeré požadavky na zobrazení publikací v jejich originálnim umístění - u vydavatele - prováděny automaticky na pozadí a bez povšimnutí čtenáře. Výhodou tohoto řešení je ta vlastnost, že pokud není dostupný obsah u vydavatele, automaticky zobrazí lokální kopii téhož dokumentu čtenáři a chová se tedy jako mezipaměť. Druhý způsob se vyznačuje tím, že k dokumentům je přistupováno pomocí jejich lokálních URL, které ukazují na jejich umístění kopie. Při takovémto požadavku je nejprve zkontrolováno, zdali je možné poskytnout obsah z webu vydavatele a poté, když to není možné, poskytne

čtenáři svoji lokální kopii. Rozdíl oproti předchozímu řešení je ten, že se chová spíše než mezipaměť jako běžný server. Poslední způsob kombinuje obě předešlé řešení v podobě integrace systému LOCKSS Box do knihovního katalogu. Čtenář poté může přistupovat k jednotlivým publikacím na základě odkazu z katalogu, který získál z OpenURL překladače.

Velmi důležitou součástí systému je jeho správa. Ta je dostupná skrze webové rozhraní, ve kterém se nachází monitorování uchovávaného obsahu, včetně řízení přístupu k němu. Vedle monitorování obsahuje i službu, s jejíž pomocí mohou knihovny lehce přidávat nový obsah.

Jako poslední bod bych chtěl zmínit systém pro migraci formátu dokumentů. Ten je důležitý z toho důvodu, protože LOCKSS je schopen uchovávat nejrůznější podoby webového obsahu, jako například obrázky, zvuk, video nebo prostý text. Pokud by se měla s každým takovým záznamem uchovávat i jeho různá podoba, vyžadovalo by to velkou režii a hlavně velké množství prostoru. Proto se v rámci systému LOCKSS uchovává pouze originální podoba a v případě, že čtenář nemůže zobrazit požadovaný záznam v originální podobě, dojde za běhu k jeho přeformátování do dočasné kopie, kterou již může zobrazit.

4 Zhodnocení

Myslím si, že tento projekt je velmi zajímavý a svou koncepcí do jisté míry i unikátní. Jeho hlavní výhodou je to, že se řadí k open-source softwaru a využívá tak všech jeho vlastností, jako například otevřené standardy, transparentnost nebo nižší náklady v porovnání s komerčním řešením. Jako další velkou výhodu vidím v tom, že spolupracuje s mnohými velkými organizacemi, jako je Stanfordova univerzita, DSpace nebo Německá národní knihovna, což mimo jiné reprezentuje jisté záruky za kvalitu tohoto systému. Velkou výhodu představuje také způsob, jakým jsou dokumenty zpracovávány a uchovávány uvnitř systému, protože představuje jen nepatrnou režii pro vydavatele participující v projektu LOCKSS. Jako přínos tohoto projektu bych zvolil to, jakým způsobem se jim povedlo zapojit tak velké množství vydavatelů, knihoven a podobných organizací a to především díky jednoduché počáteční myšlence - dlouhodobé uchovávání obsahu skrze redundanci dat. V neposlední řadě bych rád vyzdvihnul přínos tohoto projektu v rámci projektu CLOCKSS, který si klade za cíl uchovávat vědecké publikace pro další generace.

Jako určitou nevýhodu tohoto řešení bych viděl počet kopií jednoho dokumentu v rámci celého systému. Mnohem úspornější řešení by bylo,

pokud by se v rámci systému vyskytovalo jen předem stanovené maximální množství kopií a poté by docházelo k jejich distribuci z ostatních repozitářů.

Reference

[1] *LOCKSS* [online]. 2015 [cit. 2015-12-01]. Dostupné z: https://library.stanford.edu/projects/lockss

[2] *Lots Of Copies Keep Stuff Safe* [online]. 2015 [cit. 2015-12-01]. Dostupné z: http://www.lockss.org/

[3] *Get Involved* [online]. 2015 [cit. 2015-12-03]. Dostupné z: http://www.lockssalliance.ac.uk/get-involved/

[4] *The LOCKSS Team* [online]. 2015 [cit. 2015-12-05]. Dostupné z: http://www.digitalpreservation.gov/series/pioneers/lockss.html

[5] *Meta Archive: How it works* [online]. 2015 [cit. 2015-12-05]. Dostupné z: http://www.metaarchive.org/how-it-works

[6] *New experimental features in LOCKSS* [online]. 2013 [cit. 2015-12-05]. Dostupné z: http://www.lockssalliance.ac.uk/2013/10/28/new-experimental-features-in-lockss/

[7] *From Prototype to Production: CLOCKSS Debuts at ALA* [online]. 2015 [cit. 2015-12-08]. Dostupné z: http://www.clockss.org/clockss/From_Prototype_to_Production

[8] *LOCKSS (Lots Of Copies Keep Stuff Safe) UK Alliance Membership* [online]. 2015 [cit. 2015-12-09]. Dostupné z: https://www.jisc-collections.ac.uk/Catalogue/FullDescription/index/879

[9] *Talk for "RDF Vocabulary Preservation"at iPres2013* [online]. 2013-9-03 [cit. 2015-12-09]. Dostupné z: http://dcevents.dublincore.org/public/special/dc2013/DavidRosenthal.pdf

Metadata

TITLE=Projekt LOCKSS
CREATOR=Tomáš Svoboda
DESCRIPTION.Abstract=Tato práce pojednává o open source projektu LOCKSS (Lots Of Copies Keep Stuff Safe), který slouží pro spolehlivé a dlouhodobé uchovávání digitálního obsahu. Projekt vznik na konci 90. let minulého století pod vedením knihovny Stanfordovy univerzity. Do dnešního dne se do projektu zapojilo více než 500 vydavatelů. Vedle globální sítě LOCKSS, která slouží pro běžné uchovávání digitálního obsahu, se postupem času začaly objevovat i soukromé sítě nebo například globální archív vědeckých publikací.
SUBJECT.Keywords=LOCKSS, knihovna, repozitář, redundance, kopie, dlouhodobé uchovávání, digitální obsah
DATE.Created=2015-12-9
LANGUAGE=czech
FORMAT.Medium=application/pdf