Národní knihovna
České republiky
National Library
of the Czech Republic

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	**Mgr. Zdenko Vozár**
Pracoviště – dle organizační struktury	2.1.3.
Pracoviště – zařazení	vedoucí Odd. podpory aplikací NK - NDK
Důvod cesty	1. Návšteva a assesment využitia HDFS v praxi Britského Webarchívu 2. Obecné zoznámenie sa s jeho backendom, sklizňami, prezentačnými vrstvami a zároveň manažmentom netechnických požiadaviek 3. Universal Viewer 4. Zázemie British Library
Místo – město	**York - Boston Spa**
Místo – země	Veľká Británia
Datum (od-do)	**21.10. - 25.10.2018**
Podrobný časový harmonogram	21.10 Praha - Viedeň - Manchester - York - 22.10 -24.10. Boston Spa (denne z Yorku) - York - Manchester - Brusel - Praha -25.10.2018
Spolucestující z NK	nikto
Finanční zajištění	NAKI
Cíl cesty	British Library
Plnění cílů cesty (konkrétně)	1. Nadviazianie priateľského vzťahu s jedným z popredných archívov konzorcia IIPC 2. Využitie ich skúsenosti s HDFS a návrh implementácie v NK ČR 3. Možnosti implementácie tejto technológie v iných segmentoch BL 4. SW využívaný BL WA a jeho ekosystém 4. Iné technológie v BL (Zobrazovače, LTP a virtuálna emulácia, manažment procesov)
Program a další podrobnější informace	podrobná správa ako Príloha č. 1
Přivezené materiály	bez materiálov
Datum předložení zprávy	20.11.2018 (ústne pred DAT 30.10.2018)
Podpis předkladatele zprávy	20.11.2018
Podpis nadřízeného	20.11.2018
Vloženo na Intranet	
Přijato v mezinárodním oddělení	

Príloha č. 1. Zápis z cesty do Britskej knižnice
Oddelenie podpory aplikácií NK-NDK

Dátum, miesto konania 21.-25. 10. 2018, British Library, Boston Spa

Účastníci: Mgr. Zdenko Vozár

1. Úvod a základný nárys systémovo aplikačného rámca

Britský Webarchív, ďalej BWA, sprístupňuje dáta na základe partnerstva šietich britských depozitných knihovní, v 6 bodoch Veľkej Británie (Londýn, York, Wales, Edinburg, Cambridge, Bodleian) a 1 v Írsku (Trinity College, Dublin). Webarchív disponuje v základe 75 000 frekventovaných seedov (a subseedov) nerátajúc celoplošné zbery a 600 TB dát, tzn. warcov celkovo, uložených na HDFS.

Kurátorský tím a aj technická podpora vychádzajú zvnútra knihovne. V malej miere začali už v roku 2003. Prava legislatívy trvá, podobne ako u nás, prístup bol značne obmedzený do roku 2013. Pomocou zazmluvnenia a public webov, môžu sprístupňovať pomerne dosť, zvyšok nahrádzajú pripravenými datasetmi. Dodnes sa však u verejne neprístupných stránok jedná o model simulujúci fyzický depozitný kus, kde sa BWA konkrétna archivovaná stránka prezerala v knižnici, napriek šiestim depozitom, locknutá na čítanie len pre jeden konkrétny dotaz. Používajú heritrix pre crawlovanie s cielením na doménu .uk a s podmienkou, ak je Geo IP v poriadku, prípadne mimo doménu využívajú GeoIP pre identifikáciu ďalších stránok. Briti využívajú denné kontinuálne crawlovanie, skrz stream json konfigurácie, nazývané pravidelné crawlovanie, ktoré je silne administrované. Kurátory majú možnosť zmeniť frekvenciu a ďalšie pravidlá.

Naproti nášmu workflowu, kde prevažujú kurátori je tu tým rozdelený pol na pol, kurátori a tech. podpora. Každý má ale vlastného vedúceho, pričom kurátori patria do úseku Kolekcií a TP do Inžinieringu. Na technickú podporu, ktorá siaha od hardwaru až po vývvojársky software a systémové administrovanie majú technickú podporu štyroch ľudí (javista je ale dostupný pre celú organizáciu). Konkrétne disponujú jedným človekom hlavne pre hardware, operačné systémy, systémovým aministrátorom (HW + Sysadmin); jedným juniorom pod ním (ale

bez HW ale je zároveň datovým analytikom), jedným externým Javistom, jedným technickým vedúcim (Andrew Jackson, DevOps, Pythonizácia BWA) a troma kuratormi z iného oddelenia.

Vo Webarchíve z Collections dôležité postavy:

- Ian Cooke- head of contemporary british collections
- Nicole - lead of webarchive. There from 2005, head from 2016. Before Helen Hopps, when she left, they split WA team and give curators to collections and more bbizinis side, stakeholders, but silos, lot of different and also digital format
- Carlos - curator, previously web dev, new user interfencis

Vo Webarchíve z Technology dôežité postavy:

- Andy - technical lead o technical webarchive, 7 years
- Gill Senior - webarchiving enginneer
- Lee Web - archiving engineer
- Mindy - software development, but architectural team

Third party contractor - QA by users, different human profiles, but library has testing department automatically

Zber dát je automatizovaný proces. Štartuje sa pre výberovky z W3Act, ktorá má jednoduché API. BWA má crawly spúštané denne, mesačne, kvartálne či polročne. Crawly dnes ročne vytvoria ca. 60 TB dát bez celoplošky.

Ročný celoplošný zber nie je automatizovaný, pretože sa tam vyskytujú problémy. Frontier heritrix býva spontánne prerušovaný, čo je problém, ktorý sme zaznamenali aj u nás. Pre zber majú od providera seedy .uk, GEOIPs majú z verejnej databáze. Výsledne má ca. 70 TB, predtým to bývalo 30 TB. Zber sa ukladá do staršieho systému operovaného Apache Hadoop, release 0.20, s map. redukce, ktorý pôvodne slúžil ako dočasný systém, pred proipravením do preliatia našej obdoby LTP: Digital Libary System (štvornásobná distribúcia po Veľkej Británii). Digital Libary System funguje ako naša verzia, avšak zatiaľ tam majú iba cez 150 TB dát z webarchívu. Na indexáciu používajú austrálsky OutbackCDX, ktorý funguje ako servica, indexuje a zároveň to ukladá. Dotazujú sa jej Waybacky a Pywaybacky (sú rôzne dimenzované pre rôzne (6x) Reading Room services v rámci endpoinotv a pre rôzne Open Access data). Rovnako nad dátami maju vybrané verejne prístupné datasety, SOLR index a snažia sa sprístupniť všetko, čo ide, napriek nevýhodne postavenému zákonu. Pre CI používajú PythonLuigi, ktorý zanecháva správu o každej dokonanej činnosti.

Andrew má predstavu, že postupne tým pokryje celý tento ekosystém. Ohľadom strojov na zber používajú crawler 1 a pre celú doménu používajú ešte ďalšie dva, nejde ale vyslovene o snapshot, napriek ich šťavnatému založeniu. Napriek malému množstvu sú celkom výkonné, príkladom jeden má 64 jadier.

Ako disky používajú pre ukladanie dát momentálne Hitachi, 12 TB helium. 7200 RPM, pôvodne používali 5400 RPM. Disk je schopný poňať tak veľa, ako je možné, ale pre platformu výpočtového výzkumu, ako je ten náš, 8 TB je výhodnejšia z dôvodu jej dynamickej štruktúry. Overhaul rezerva nemusi byt až tak vysoká pre generovanie dát smerom von. V prípade, že sa tam niečo používa Spark, odporúčajú použiť 20%. BWA staščí skôr smerovať k 10%, je to v poriadku, ale nesmie sa táto hranica prekročiť, inak riskujú významné problémy. Spektrálny prístup od obnovenia jednotlivých položiek po plný výpočtový prístup.

Čo sa týka komplikovanejších otázok na webarchív, v rámci podpory fazetového vyhľadávania, pre solarizáciu treba použiť kvanitutu na SSD odpovedajúcu cca. 10 percent z celkového objemu dát. Oni majú dva SOLR servery, ktoré sú ale veľmi výkonne, kapacitne i výpočtovo.

V prípade výzkumníkov s ich vlastnými otázkami, ktorí však nepotrebujú príliž stravujúce data a postačia im sekundárne, používajú predprodukované datasety alebo v prípade výzkumníkov využívajúcich vzdialené dotazy by chceli využívať službu iunder Jupyter NB, avšak ešte nemajú na to postačujúcu infraštruktúru. Turingov Inšitút, zaoberajúci sa linguistickou analýzou v rokoch 1996 - 2004, má záujem na ich datách. Spolupráca by šla rozchodiť cez Azure.

Čo sa týka dotazov on-site na celú dátovú sadu BWA, týmito kapacitami zatiaľ nedisponujú.

Pre túto chvíľu Britskej knižnici stačí Spark, respektíve MapReduce Vanilla a SOLR. Ak ľudia chcú užívať R, môžu si ho previesť. Jediné, čo k tomu treba, je trocha technológii a bezpečná vykonová platforma. Toto by zvládal Kubernetes sám, tento však nemajú, ale chceli by v tom pokračovať.

Využívajú hlavne Python v roli orchestrácie. Technológie naviazané na IO nepotrebujú. Hlavné kompotenty sú založené na Jave, teda výnimkou Python Waybacku.

2. Otázky všeobecne súvisiace s webarchívom

2.1 Ako fungujú ukwa/ukwa-ingest-service?

- Prevedenie Heritrixu a okolných ekosystémov do kontajnerov Docker.
- Prečo práve kontajnery na Dockery?
- Využívajú cluster?
- Spravujú kontajnery pomocou Kubernetes?

British Library používa Docker Swarm compound, ktorý sa podobá práve Kubernetes. Inak využívajú ešte Zookeeper, Kafka, CDXservice, Ukwa ingest services, deploy, Crawl engines, zvažujú do budúcnosti prechod na Ansible. Docker im vyhovuje práve preto, že je veľmi vhodný a jednoduchý. Je to hlavne krátkodobé taktické rozhodnutie, ktoré sa im ale vyplatilo.

2.2 Ako riešia metriky archívu, zberu, atp?

- Ako riešia projekt ukwa/monitrix, prípadne ELK stack alebo celkovo?
- Majú nejaké informácie či výstupy z nového projektu ukwa/ukwa-monitor?

Zápasiac s rôznymi spôsobmi manažovania, monitorix by indexoval v ELK každú jednu udalosť. Logstash. Dosť ťažké je to ale vylepšiť, ani nikdy netreba tak strašne moc detailov. Vykazuje veľmi špecfifickú maticu. Podávajú to do Promethea, aby vyextrahovali a interpretovali špecificky užitočnú dataset z rôznych systémov pre orchestráciu práce. Chcú do budúcnosti viac podrobností, zatiaľ im to stačí. Problém je dotazovacia syntax, ktorá je strašná, aj keď to dlho vyzeralo, že to bude fungovať v implementácii.

2.3 Ako sa im osvedčil vlastný task management ukwa-manage

- Prečo akurát vedú workflow pomocou frameworku spotify/luigi?

Nagios sa používa, keď veľa strojov má problém agregovať výsledky, jedná sa ale skôr o cloud a celkom nevýhodný. Luigi, framework pre riadenie úloh/ taskov. Ak skončí správne, robí pit managment, dáva rovnaké flags, ponáha s analýzou logov Hadoopu, zbiera vstupy do SIP. Monitoruje a kontroluje kód. Automatizácia musí byť dobre zomknutá a kontrolované. Úlohy sa opätovne spúšťajú, ak zlyhajú. Luigi je pre tieto účely najlepší. Reťazí vstupy k výstupom. Pomáha v správnom výsledku. Vždy, keď zlyhá, pošle správu Prometheovi. Ich implementácia, je ale dlhodobá práca. V BWA, je konečne pridávať stroje jednoduché. Stačí len pridať súbory na file systém.

2.4 Ako sa im osvedčil managemet oprávnení?

- Aké majú skúsenosti s ich vlastným riešením ukwa-access-services ?

CDX service a veci založené na prefixoch generujú Whitelist a vylučujú veci do webarchívu. Luigi generuje curate z datasetov, spája ich so statickým zoznamom a potom ich načíta na server. Generuje to každý deň.

2.5 PYWB – prístupová kontrola UKWA a PYWB

The BL tried to support open Wayback device throughout past years. It was very slow and had few users with new code. Javascript was a heavy solution, then PYWB came along and Ilja got a lot of movement. It has clean implementation in python form: it has intensive quality of playback, better quality, raiser is paid to adjust it or to try properly curators verification. There is no example that is worse than with Wayback and myriad examples that were better. Javascript resources have better playback. Python is good also for a team. A new website deploys Python Wayback. Portuguese, UK Parliamentary, National Archive got a good shift with this. Original developers of Openwayback moved away from IIIPC (problem). IA implemented also own version of PYWB (it is a closed source), not Iljas. New website is probably ready, but later this week went online.

Povoľujú stránky a všetko ostatné zakazujú. K spúšťaniu PYWB používajú Rhizome Extensions one CDX service so všetkým v tom. IP adresy su stále. UKWA PYXB, ak to je vblock, predstiera sa, že nie je. Majú šesť Wayback inštancií. Pre každú fyzickú lokáciu sa pripája cez VPN. Už v tomto momente majú právne vkladové právo. Pre vzdialený prístup používajú Ericom Software, ktorý funguje, ale úplne nadšený z neho nie sú. Pre digitálne edície kníh, nie webarchív, používajú Adobe Digital Editions pod DRM. Ak je to *allow,* vracajú späť pre open access Wayback. Ak je to not *allowed* a zároveň, ak to nie je *block,* odporúčajú prístup k Legal deposit Library Reading Rooms. Každá Wayback inštancia má inú konfiguráciu a jedinečný súbežný zámok (lockujúci prehliadané stránky). Je to konfigurované hlavne na britské právne problémy.

2.6 Ako balíčkujú WARC?

- Ako vyzerajú kurátorské veci? Ako to majú, ako katalogizujú, ako často harvestujú? Ako to vyzerá u nich so social media a novinami?

Problém heritrixu je, že je veľmi zviazaný frontier s jobmi a vlákami. Ak zaistíme kópiu všetkého, od heritrixu zvonka, môžu používať python scripty pre kontrolu toho, čo sa deje. Kafka riadi kompresiu a všetko, disponuje datami ca. 4TB. 16 strojov posiela dokopy 5 000 správ za sekundu a má to previazané s Prometheom. Každý zapisuje naraz asi 10 balíčkov. Limit na balíček je jeden GB. Pridali bean, ktorý vystavuje kritické správy a ktorého vlákna sú odchytávané.

Site to Warc balíčkovanie by kurátori BWA preferovali, ale každý crawler to narozdiel napr. od Slovenského WA, kvôli technologickým podmienkam mixuje dokopy. Existuje higj frequency crawler and doménový crawl server. Doménový crawler pracuje tri mesiace. Obsahuje dve služby so 16 heritrixami. Má to dokerizované a zapúzdrené do kontajnerov. Nebolo to bez problémov. Bolo potrebné špecifikovať environmentálne premenné zvonka, s tým že všetky url šli cez Kafku a log systém. Kafaka má všetky url. A keď sa zmení počet strojov, natiahnu si to z Kafky, a preto je to viac dynamické. Zámer je viac kontrolovať crawlery. Tento rok prvý krát, každá jedna URL padla vo fronte, dokonca keď crawlovala, bola duplicitná. Získali 20 miliard URL adries. Keď treba repassnúť, je to problém, lebo sa musí ísť na začiatok fronty, takze buduci rok budú optimalizovať načítanie.

2.7 How does British National Library work with CXML?

They use one bean configuration, but the usage is unconfortable. The number of files, 500 MB, selectively changed and the crawl is up to 2 GB, or unlimited. Now too many people can change it, they need to adjust. By default they obey robosts.txt, except for screenshots. Currently now not per site, but whole crawl, than they want specify this rule per site. Interesting interpretation of budget, they prefere rather in bytes than in number of elements. They have polite setting for hostname, not IP adress archiving. They have blacklist for some intranet sites, ca. more than 50, when something illegal was published, which infringed data protection, no offensive data. Right now they do´nt need to delete it.

They automatically query W3Act, rsp. what should they do. He will give them complex answer to Kafka (through cron job on luigi level orchestration), and from that point onwards, they formed in dependencies on job.

2.7 How do Catalogue integrations work?

Catalogue records are given into collections and into Aleph, but they are not integrated, which causes problems. Document harvester, postprocessing, like pdf from crawl, prepare metadata form and submit them to library service, which individually put into Aleph.

(Instead for GDPR, they have British Library Act.)

2.8 WARCLIGHT

In 2005 - 2013 the BNL switched to permission-based crawling and now they have about 20 000 seeds. In system they have about 50 000. We have sent approximately 45 - 50, 000 requests for permission in total, so our success rate is about 25%. (info via Nikola).

3. Software architects meeting

Využívajú Aleph a Primo a discovery services. Zvažujú prechod na niečo, ale nie momentálne. Na Almu ich tlačia, ale do 3-4 rokov sa pokúsia prejsť na Folio podľa jeho výkonnosti.

Zmena hardwaru. Prešli na cloud, hlavne s hot services, cold si asi nechaju. Je mozn, že sa pripoja k statom pripravovanemu usiliu, ministerstvo armady nieco take už má.

Čo sa týka SW architektov tímu,, vedúci informačných zdrojov a služieb majú polovičku SW architektov a polovičku bussiness architektov (v podstate ľudí tvoriacich diagramy).

Čo sa týka prezentačnej vrstvy, disponujú Rádio nahravaním. Majú 50 radií pre nonstop nahravanie. a chystajú sa rozširovať. Radi by v tomto s nami kooperovali.

4. Digital preservation

4.1 Tímy

Majú niekoľko tímov. **Technical analysis and development** zabezpečuje uchovanie štandardov. Je vytovrený hlavne pre vytvorenie, prípravu a podporu užívateľov, hlavne pre ich prístup k úložiskám dát. Ale využíva sa aj pre operačnú prezerváciu. Zhromažďujú a opravujú tiež prístroje z 80´tok, aby mohli kontrolovať úspešnosť emulovaného prostredia, oproti originálnym prístrojom,

čítajúcim dané formáty a rovnako majú onsite slušnú kolekciu technologických časopisov z danej doby.

Metadata services team coordinates with standard production bodies of standards and reccomendation for inner development of digital repository. It is very old, 2003 they began. Aiming to complexity for big states, not pushing standards on other people, but on themselves, converting themselves. Now they are going for commercial platform, massive undertaking, DLA over petabyte in repositories and 2-3 petabytes waiting for content ingest. Substantial amount of legacy digital content from 70´ties and 80´ties. There are still pockets of legacy data. There are problems of modeling new DLS, standard streams and also for network shares. Dangers dwell in network shares, collection and personal files, mostly with deleting. No onae knows fors sure what is there, without qualitative metadata. Digital processing there is to support ingest. They work with acquisition colleagues to ensure best practices and investigate reading rooms, when their presentation of sources is not working. Originally, their current system was not desiged to present. When they change to new system, their role will be much more focused on preservation.

5. BRITISH LIBRARY Universal viewer IIIF (Andy Irving)

The British Library has 35 viewing viewers-apps, very complex enviroment. They introducet Universal Viewer. I further investigated its performance, according to problems of our Kramerius. They currently use 3 image servers, now expanded to ten, half million books digitalised by google, transformed via 11 server to use them. They use open source, bodleian IIP image server, using C, linux. Frontend is JS, not much server complexity, need a shim between application and solr, mongo FDB json docs, not generating docs, reversed export, IIP imaging, contSecureAuthentificaiomDesicomAMaker. Pretty light solution, 20 miliseconds response, now in production, load depends - search is easy 5 %, but images requests are very high, generally 40 requests per second, Bodleian have 16 images servers, no special configuration, cross site avaliability. Load balancer is actually done via Nginx, caching proxy has less than 1TB in front of apps, quad CPU 16GB ram, kakadu based streaming. They run from master, we have seperate master image secondary, caching in Nginx, they have long scolls, like 50m (5Gb), load image tiled and send then, they could cut them into smaller images and joint them by stitching, great with IIPS. Good also to implement Mirador viewers for scholars. They generate pdf on fly, service possible to take

the whole file, but then they view in universal viever and pdf combined, in couple with OCR. They limit degree of paralelism and ligthweight jobs system.

They are throttling through nginx. Using fedora for open access system. Wales has problem with resolving identifiers with fedora, keep access and ingest apart! They use redis as memeory cache think, locks paths and needs lot of memory. MyCycle server is better with it. Access stack is all linux here. Redis has lockup, system resolvers behind as well for access control. It is also used for representation between. It is important that they have different resolvers for different access layers, not using flags.

6. Webarchive Discovery Stack - implementation, workflow

Part of **warc indexer** (took **apachae tikka** and ability to open warc containers and takes also human metadata from **w3act**, collections and so on), create **jsons documents** sent to **solr,** that what used Danes to do with one really meaty server. UKW webarchive discovery, it is on maven likwa.discovery:warc-indexer, you need. The BL uses **warc hadoop indixer** - it is **warc indixer**, actually using the same code. Wrapped in hadoop record reader, each mapper gets a stream of warcs, wrapping of indexer with hadoop specific **map reduction.**

MAP device uses warc to solr json. The **REDUCE device** just posts hundreds of solr documents to solr with schemer. Numbers of warc **mappers** could be one hundred, their output is taken by **Reduce** device. It is easy to controll number of reducers and usage of **solr.** The use ten, not so much. **M a R** are independent from each other, the number can be tuned up. **Infrastructure is I/O bound, SSD defines speed limit.**

The SOLR is used for language identification, etc. They use **vanilla solr** plus of **standard schema for webarchives,** which makes possible to deal with links etc. They use it for building collections, they could flip alias of collection and client will see other collections. It is great for management of collections and clients.

7. HADOOP

They have cluster with 66 nodes. Tikka is intensive, configuring how many each could run, 2 gigs per jobs, for N:LP fancy stuff, more ram will be neeeded. This only extracts json, and posts it to meaty servers 1/4 and 1/2 TB of ram, ssds, and decent cpus, due to advances in solr one does not need so much loading.

Also if you have many facets, query complexity is not so much loading, but with much RAM storage, it could efficiently cache critical cache size. **SOLR servers are most expensive part of kit.** It is strongly recommended only for particular collections. As a rule of thumb, ca. 10 percent of collection will index SSD.

The best it would be to give there contract collection, but with one hop off policy. Maybe it would be even good to put it on SOLR and erase it. Advantage for general audiencie and high technical audience. The majority of SOLR content is made up texts. It is needed to think through how to put down the weight of warcs into texts. Unless is it optimized index like SOLR. To get into what we need to tab seperated useful intermediary, this will do great stuff. It depends on research of questions. There is a need of stakeholders to convince.

What concernns configuration of hadoop, the use ansible magic here. They tune configuration and restart everything use for it, good for ansible. Yarn better conf platform. They do not have yarn, that is why they have static deployment.

HDFS Eco

It is structured by Name node and Backup Name nod.
plus grunt (data node)
DN - DN DN DN
Since they do not have yarn, they have job tracker on the level of NN.
grunt is DN and TAskTracker
MAP generates data and Reducers gather it.
MAP uses 1 task per warc.
Reduce its spread everywehere. There is a lot of redistribution of data. Each node is different than orders. It is led by common key.

Main outputs are extracted by Reducers. CDX index [remote services], SOLR [remote service], links most common formats - put in file format, links, how they generate datasets for further processing [and shuffle back new files as add to cluster.] Preference is not to produce WAT and others but line generated files, which are more lighter lightways for further processing and to have them. And also problem to store them.

Under old hadoop it is hard to mix java and python, our only capabilitiy to parse is by warcs in java First step is map-reduce to parse the warc, with some text format and then carry on in python, through python streaming. They wanted to change it in time to Spark, but they are still not there.

For aggregation in hadoop, you need to generate some intermediary file format with critical information. So many of warcs (10 000 000) job tracker will kill it.

You need to do it in batches. Problem is based on mapping from warcs to intermediary text (good model will be 10 000 warc to one line oriented file). Hadoop could do magic with line orientated text. Then you could sort, analyze and aggregate them. Aggregate frequency analyses from units to aggregated file and from this on goes analysis.

Runnning SOLR drive and have many shards on separate devices it is still fine, we do not need to have it on separate SOLR SSD server. The idea is to run it fromSSD.

Research question is what you want you to do first? Do they need texts? Do they need some subste of texts? There will be way more records in disk space then.

8. SPARK

ukwa.webarchive-discovert Front ends.

How do they prepare **dataset?**
They use mainly human strategy, BL Collections development policy. Support comes also from other curators, thematic collections or events. Datastes, on their pages, are research-led and help them to refine. New legislative - right to read is right to mine. It will take a while.

9. Financing - EU projects

Web-archiving salaries is by 100 percent supported by the British Library. Other endpoints partly contribute by 1/3, lot of project funding for spent specific projects, to develop tools, other members of the stuff, more ad hoc bases, maybe had some PhD students on short term (3-12 months), cosupervising PhD or host. European projects that are used are Cleopatra or RESAW. They use rather smaller IIPC projects more likley.

10. Instant archiving of newspapers and social media

Every day the do snapshots on the British level, smaller they do every weekend. They have sensitive relations with newspaper publishing. Social media dont do so often. It is hard to put the limit, what comes from UK, curators are needed for that. Social media tend to be international. From social media screened are high-profile public profiles, politicians. Looking into alternative tools, Webrecorder. Social feed is a good tool, how to integrate those warcs outside workflow.

Toto je interní dokument Oddělena podpory aplikací

Zdenko Vozar
Dne 25. 10. 2018