LDCX nem-konferencia a Stanfordon 2018. március 26-28.

Király Péter

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Göttingen, Germany
peter.kiraly@gwdg.de

2018. április 1.

A kaliforniai Stanford Egyetem 2010 óta évről-évre megrendezi az LDCX nevű nem-konferenciát (unconference). Az első időkben az LDCX a Könyvtári fejlesztők ... konferenciáját jelentette (az X tetszőleges technológiát jelentett), később a tematika kibővült a teljes, LAM-nak (Library, Archive, Museum) rövidített közgyűjteményi intézményhálózatra. A résztvevők száma korlátos, 100 körüli és meghívásos alapon működik. A legnagyobb kört a Stanford és más kaliforniai egyetemek (UC Berkeley, UC San Diego, UC Santa Barbara) munkatársai alkotják, de szinte valamennyi jelentős amerikai egyetemről (Harvard, Yale, Cornell, Princeton, Penn State stb.) érkeztek szakértők, sőt, idén többen is (nyolcan) Európából (Oxford, British Library, Norvég Nemzeti Könyvtár, Vlaamse Kunstcollectie stb.). A résztvevők köre a könyvtári, levéltári, múzeumi szféra vezető technológusai közül kerül ki. A rendezvény fő célja az eszmecsere, vita, ismerkedés. Hagyományos, prezentációra, netán tanulmányra épülő előadások nincsenek is, a jellemző formátum a mederben tartott beszélgetés.

Idén három ülésforma volt jelen: 5 perces villámprezentációk, ad-hoc (a konferencia első óráiban megválasztott) tematikus megbeszélések, és előre meghatározott témacsoportokra alapuló, az egész konferencián végigvitt, hosszú megbeszélések: mesterséges intelligencia/gépi tanulás, adat, digitális állagmegóvás, felhasználói felület (a kezdetben tervezett ötödik csoport, a webarchiválás résztvevői az első ülésükön úgy határoztak, hogy a témát összevonják a digitális állagmegóvással és inkább közösen tanácskoznak). Aki szerette volna napirendre tűzni a saját monomániáját (az én esetemben a metaadatok minőségének mérését), annak egy pár mondatos meggyőző szónoklatot (néhányaknak bizonyára ismerős a start-up találkozókon népszerű, a befektetők meggyőzését szolgáló "pitch" vagy "elevator speech") kellett tartania, és ajánlatát kifüggesztenie egy darab papírra, amire a résztvevők ráragasztható pöttyökkel szavaztak (kék pötty: érdekel a téma, támogatom, piros pötty: nekem is van erről mondandóm, szeretnék részt venni).

Az egyes ülések egy-másfél óra hosszúak voltak. Hogy ne essen szét az idő parttalanná váló vitákkal, minden beszélgetésnek volt négy kulcsszereplője: a beszélgetés vezetője (a témagazda, aki az adott tárgy felvételét kezdeményezte), egy jegyzőkönyvvezető (jegyzőkönyvre közösen szerkeszthető Google dokumentumokat használtunk), egy időmérő aki figyelt arra, hogy mennyi perc van még, és egy "kapuőr" (Gatekeeper), aki viszont arra figyelt, hogy a beszélgetés sokszereplős maradjon: egyrészt senki se uralja el a vitát, másrészt a csendeseket is időről-időre megszólítsa. A rendezvény honlapján világossá tették, hogy az aktivitás követelmény: „Ha azért jöttél, hogy megfigyelő legyél, vagy az esemény egész ideje alatt te legyél a légy a falon, ez számodra talán nem a legalkalmasabb hely vagy év”. Ez a proaktivitás ebben a környezetben egyébként kevésbé magától értetődő, ahogy egy résztvevő megjegyezte: azért

lássuk be, a legtöbb informatikus introvertált, és nem csap le azonnal azokra az alkalmakra, amiről a szervezők azt gondolják, hogy remek idő az ismerkedésre, networkingre. Azt hiszem mindazonáltal, hogy a formátum alapvetően segítette, hogy a konferencia rövid ideje alatt (két és fél nap) mindenki számára létrejöjjön valamilyen új szakmai kapcsolat, és ha nem is a legnagyobb plénum előtt, hanem egy kisebb csoportban, vagy akár csak a szünetekben, mindenki találjon módot arra, hogy kifejthesse szakmai véleményét.

A terítékre kerülő témák a következőek voltak:

- Metaadat-aggregáló rendszerek
- Mikroszerviz-architektúra
- Változatos és befogadó (szakmai) kultúra
- IIIF Presentation 3.0[1]
- Metaadatok minőségmérése
- Valkyrie[2] (a Samvera[3] nyílt forráskódú digitális repozitórium adattárolási API-ja)
- Teljesítménymérés + alkalmazástuning
- IIIF digitális bölcsészeti oktatási csomag[4]
- A gyűjtemény mint adata (A "Collections As Data" egy 2016-ban indult nagy esernyőprojekt[5])
- Hyrax[6] fenntarthatóság / adatmigrálás
- Nagy kockázatú személyes adatok a repozitóriumokban
- Metaadatok verziói
- Mirador 3[7] tervezés
- Besorolási állományok elérése a saját alkalmazásodban
- Az AWS-t (Amazon webszervizek) érintő legjobb gyakorlatok

Az összes ülés jegyzőkönyve elérhető a `https://library.stanford.edu/projects/ldcx/2018-conference/agenda` oldalról.

A nap végén mind a négy nagy téma csoportja beszámolt a többieknek arról, hogy mit végeztek. A konferencia zárása pedig egy „Plus/Delta" visszajelző ülés volt, ahol mindenki elmondhatta, hogy mi tetszett (plus) és az amin jövőre javítani lehetne (delta).

Az ad-hoc témák közül a metaadat-aggregálási illetve besorolási állományokkal foglalkozóakon vettem részt, illetve vezettem az általam javasolt (és végül elfogadott) minőségmérési csoportot. Számos nagy egyetem (Harvard, Stanford, Oxford) egyetemi könyvtárát nem egy intézményként, hanem intézményi hálózatként kell elképzelni, akik az utóbbi évtizedben nagy hangsúlyt fektetnek az informatikai együttműködésre. A Code4Lib-be tavaly Dave Mayo egy részletes cikkben ismertette 35 harvardi levéltár közös katalógusának kialakulását, és mint most megtudtam, hasonló aggregálási-normalizálási folyamat zajlik több más intézményben is. Az aggregálás során

[1] `http://prezi3.iiif.io/api/presentation/3.0/`
[2] `https://github.com/samvera-labs/valkyrie/`
[3] `https://en.wikipedia.org/wiki/Samvera`
[4] `https://docs.google.com/document/d/1RvVmNWbEtJ_ftGE3SpWFMYi4bJa2PUBIVANzhP_Ia2o/edit`
[5] `https://collectionsasdata.github.io/`
[6] `http://hyr.ax/` A Hyrax a Samvera felhasználói felülete
[7] `http://projectmirador.org/` A Mirador egy IIIF-re alapuló webes képnézegető.

számot kell vetni az intézmények eltérő katalogizálási szokásaival, esetenként ezeknek véget kell vetni és közösen áttérni egy egységes rendszerbe. A besorolási állományok használatánál az egyik nagy informatikai téma az adatok gyorstárazása és ennek szinkronizálása. Amikor egyszerre sok rekordon végeznek adatdúsítási (enrichment) eljárást (vagyis a meglevő közgyűjteményi rekord egyes adatelemeihez automatikusan máshonnan vett adatokat rendelnek), akkor az a lekérdezett rendszert nagyban megterhelheti – a gyorstárazás egy köztes adatréteget állít ebbe a folyamatba, hogy a külső szolgáltatást adott időszakon belül legfeljebb egyszer kelljen elérni. Az LD4L keretében a Cornell egyetem épit ki egy szolgáltatást, ami egységes felületet biztosít a legfontosabb authority szolgáltatásokhoz (ezek jelenleg: Library of Congress Subject, Names, Genre/Form állományok, Geonames, DBPedia, National Agricultural Library Thesaurus, Agrovoc és az OCLC Fast, a Getty TGM/AAT és a Wikidata integrálása folyamatban van).

A nagy témacsoportok közül az adattal foglalkozóban vettem részt (már csak azért is, mert ennek a vezetője, Christina Harlow hívott meg az LDCX-re). Az első ülésen rüvd bemutatkozás után meghatároztuk, hogy miről szeretnénk (és miről nem szeretnénk) beszélni a következőkben. A választott témák: adatokkal kapcsolatos bukások, adatkezelési esettanulmányok, cloud adattárolás és a serverless technológia, egy adott rendszer (az Islandora) belső adatforgalmának elemzése, végül három kisebb csoportra bontva elkészítettünk egy fiktív könyvtári rendszer adat-architektúráját. Több beszélgetésben is előkerült annak a kérdése, hogy mik az előnyei és hátrányai a cloud adattárolásnak (gazdaságosság versus adatbiztonság). Ami számomra újdonság volt az a serverless technológia

Fail4Data avagy a rossz megoldások elemzése. Ebben az részben olyan esettanulmányokat vázoltunk fel, melyek sikertelenek maradtak. A siker elmaradásában szinte minden esetben közrejátszott az, hogy a papíron eltervezett megoldások életszerűtlennek bizonyultak, például a program olyan megoldásokat szolgáltatott, amelyekre nem volt igény, vagy olyan adatokat várt el, melyek nem álltak rendelkezésre, ezzel szemben nem nyújtottak megoldásokat a tényleges igényekre, és nem kezelték a valós adatokat. A projektről sokszor menet közben derül ez ki, aminek a következménye: csúszás, félkész megoldások, extra erőforrások bevonásának követelménye. Sokszor probléma, hogy az adatcsere- vagy adatmegjelenési formátum és a tárolt adat szerkezete, mennyisége, különféle tulajdonságai jelentősen különböznek, ami kompatibilitási, teljesítménybeli vagy skálázási problémákhoz vezetnek. Megerősítve érzem azt az alapelvet, hogy a rendszertervezésnek (egyebek mellett) a valós adatok elemzésén kell alapulnia. Több kolléga explicit módon ki is mondta az ismert adatelemzési szlogent: ismerd meg az adataidat („know your data"). További hasznos tanulságok: az adatok verziókezelése, a régi változatok archiválása biztonságossá teszi az „adatmasszázs" (data wrangling, data munging, data massage) vagy más néven az ETL (extract, transform, load – azaz adatkinyerés, transzformáció, betöltés) folyamatát. További problémákat okoz ha az adatok csak részben jeleníthetőek meg, pl. a TEI-ben kódolt elektronikus szövegeket megjelenítő szoftverek egy ideig csak a kódolt információ egy részét tudták megjeleníteni, ahogyan a könyvtári rendszerek sem jelenítik meg a teljes MARC struktúrát - következésképpen a nem megjelenített részekben szereplő hibák is sokáig rejtve maradnak. Az egyik résztvevő megjegyezte, hogy érdemes az adatokra vonatkozó követelményeket és esetleírásokat átvételi ellenőrzésekben (acceptance tests) is rögzíteni, így a projekt során azok mindig kontrollálhatóak és mindenki számára világosak lesznek.

A következő ülésszakban három adatkezeléssel kapcsolatos projektismertető hangzott el. Én beszámoltam az Europeana adatminőségmérő eszközéről, majd Matthias Vandermaesen működés közben bemutatta a közgyűjteményi adatok parancssori ETL manipulálására szolgáló Perl alapú Catmandu-t[8] és a metaadat-aggregáló PHP alapú Datahub szoftvert[9]. Zárásul Esmé Cowles és Trey Pendragon ismertette a Princeton digitális repozitóriumát, a Rails alapú

[8] `http://librecat.org/`

[9] `https://thedatahub.github.io/`

Figgy-t[10]. A Figgy számos Samvera komponenst felhasznál, és együttműködik a helyi levéltári és könyvtári szoftverekkel, továbbá IIIF eszközökkel.

A felhő alapú infrastruktúrával foglalkozó ülésszakban először Alexander Kessinger vázolta a Bepress váltásainak történetét. A fő cél, amit a felhő-technológiától vártak az, hogy a szolgáltatás hardver-szintű üzemeltetése helyett azzal foglalkozzanak, ami a fő profiljuk, a szolgáltatás-fejlesztés. Az első, három évig tartó és meglehetősen kimerítő szakaszban a korábban hosztolt szervereken futó alkalmazásokat tették felhő-alapúvá (az Amazon AWS technológiáját és infrastruktúráját használva). Ebben az időszakban még nem sikerült kihasználni a skálázhatóság számos jó oldalát. A rendszer ugyan stabil lett, de sok energiába került, és főleg a konfigurációkezelés bizonyult szűk keretmetszetnek. Jelenleg a második fázis zajlik, amiben a fő célkitűzések a szerverek automatikus skálázhatóságának bevezetése, zavartalan cseréje és a biztonsági vagy hibajavítások alkalmazása a teljes szerverállományon egy órán belül. Mivel a szerverek kezelése még ezen a szinten is macerásabb a kelleténél (egyrészt a kezelőeszközök bonyolultak, másrészt az Amazon által nyújtott szolgáltatások nem elég precízen definiáltak - ez utóbbi problémát egyébként mások is felhozták), jelenleg a 3.0 változat előkészítése folyik, amiben a szolgáltatások menedzselését Kubernetes[11]. Ezután Erin Fahy (Stanford) a felhő-infrastruktúra egy eddig általam nem ismert technológiai részének, a Serverless alkalmazásoknak menedzsmentjéről beszélt. A serverless technológia lényege, hogy a szolgáltatás egy stabil, állandóan elérhető backend helyett külső felek szolgáltatásaira (Backend as a Service - "BaaS") illetve efemer konténerekben futó egyedi kódokra (Function as a Service - "FaaS", pl. AWS Lambda) valamint elsősorban front-end technológiákra épül. Az ilyen módon felépített szolgáltatásnak jóval alacsonyabbak lehetnek a működtetési költségei, ezzel szemben erősebben függ a külső szolgáltatásoktól. Az előadó azonban nem a serverless alapú szolgáltatás fejlesztéséről, hanem annak menedzsmentjéről beszélt. Az általuk választott szoftver a Terraform[12], ami egyrészt ismeri és együttműködik a népszerű felhő-infrastruktúrákkal, másrészt a beállításokat magas-szintű deklaratív kódként tárolja.

Az adatáramlásról (dataflow) szóló részben Diego Pino Navarro (The Metropolitan New York Library Council) mutatta be az Islandora-CLAW-ot[13] az Islandora digitális repozitórium következő generációs változatát. Az Islandora fő építőelemei a Fedora, a Drupal és a Solr, a CLAW ezen felül erősen épít az üzenetkezelés (message queue) és folyamat-feldolgozás (stream processing) technológiákra, így a teljes technológiai csomagban a fenti főbb összetevők mellett tucatnyi egyéb komponens is megtalálható (Apache Karaf, ActiveMQ, Apache Camel). Navarro az adatok útját elemezte a rendszeren belül: mikor milyen események zajlanak és ezek hogyan módosítják a rendszer egyes elemeit, azok hogyan kommunikálnak egymással, például egy file attribútumainak a felhasználói felületen történő módosítására mi is történik pontosan a rendszerben, hogyan lesznek az adatok módosítva a Fedorában, a Drupalban és a Solrban.

Manapság bevett gyakorlat, hogy a konferenciának van hivatalos hashtagje, amit a közösségi médiában lehet és érdemes is figyelni. Az esemény #ldcx tagjét azonban itt alig használta valaki, ami annak a bizonyítéka, hogy a résztvevők valóban folyamatosan benne voltak a beszélgetésekben. Mark „Anarchivist" Matienzo az első napon azt mondta, hogy nem tudni hogyan, de a konferencia végére a varázslat megtörténik. A konferencia formátuma, a szervezők készsége, nyitottsága, és valószínűleg az egyetemi kampusz különös jellege, a Stanford szelleme is hozzájárult, hogy a különféle hátterű emberek a végére együttműködő csapatokká szerveződtek, és bizonyára a konferencia után is fennt fogják tartani a szakmai kapcsolatokat. Én – a saját kutatásom kapcsán – bizonyosan.

[10] `https://github.com/pulibrary/figgy`

[11] `https://en.wikipedia.org/wiki/Kubernetes`, illetve The Illustrated Children's Guide to Kubernetes `http://blog.kubernetes.io/2016/06/illustrated-childrens-guide-to-kubernetes.html`

[12] `https://www.terraform.io/`

[13] `https://islandora.ca/CLAW`