torjunta kiinnostaa, niin kannattaa tutustua NIST 1.1 -viitekehykseen, tämäkään ei myyntitykkien puheista huolimatta ole rakettitiedettä, vaan jo perustoimenpiteillä ja järjen käytöllä saa varsin hyvän mielenrauhan.

Seuraavaksi olikin vuorossa Memory Labin uuden supertietokoneen Hipun esittely. Esittelyn aloitti Antti Liusjärvi Atealta ja allekirjoittanut jatkoi. Laitteistot hankittiin EARK-rahoituksella hyödyntäen Atean Tiera+ -verkkokauppasopimusta, jossa Xamk on mukana. Puheenvuoroni keskittyi pitkälti rautaan, koska ensimmäisiä yrityspilotteja ollaan vasta suunnittelemassa koordinoimassamme ”Memory Lab Digitaalisuuden ja Innovaatiokulttuurin ajurina Etelä-Savossa hankkeessa”. Jos hankkeen sisällöt kiinnostavat voi kurkata *xamk.fi/ml-aiko* osoitteeseen. Hipun laitteistot koostuvat Nvidia DGX-A100 -laskentayksiköstä, NetAppin AFF A400 -levyjärjestelmästä sekä virtualisointia ajavista palvelimista, verkkoinfraa unohtamatta. Laskentatehoja tuosta DGX-yksiköstä irtoaa noin 20 000 kannettavan tietokoneen verran ja Petaflopseina ilmoitettuna teoreettinen maksimi on 5 (vrt. CSC:n Lumi-supertietokoneen 375). Jos tuo Hippu-nimi herätti ihmetystä, niin se on peräisin nimikilpailusta, johon saimme noin 150 erinomaista ehdotusta. Hippu (gold nugget) kuvastaa siis TKI-toiminnan kullanhuuhdontaa ja lisäksi Hipussa on kullanvärinen etupaneeli.

Hipun etupaneeli Gimpillä tehostettuna. Kuva: Anssi Jääskeläinen.

Viimeisenä kävin kuuntelemassa With securen Antti Laatikaista, joka puhui NIS2-direktiivistä, joka suomessa on saanut hallituksen esitysluonnoksessa muodon ”Laki kyberturvallisuuden riskienhallinnasta”. Siinä missä direktiivissä on ”vain” 90 sivua, lakiesityksestä on Suomessa saatu leivottua 248 sivun mittainen eepos. Kyseinen teos on muuten Liikenne- ja viestintäministeriön toimesta lausuntokierroksella 29.11.2023 asti eli vielä on mahdollista vaikuttaa sisältöihin. Tällä hetkellä vaikuttaisi siltä, että koulutuslaitokset kuten Xamk saavat huokaista helpotuksesta, koska esitysluonnoksen mukaan ”lakia ei kuitenkaan sovellettaisi korkeakouluihin tai muihin opetus- ja koulutusalan laitoksiin”. Muuten lain piirissä tulisivat olemaan kaikki sellaiset tahot, joiden toiminta on yhteiskunnan toiminnan kannalta kriittistä. Lisäksi lain piirissä olisivat kaikki yli 50 työntekijän ja yli 10 m€ liikevaihtoa pyörittävät toimijat.

Anssi Jääskeläinen
Tutkimuspäällikkö
Xamk

Stina Westman
TKI-yksikön johtaja
Xamk

DLM Forumin Triennial konfenressi pidettiin lokakuussa Salamancassa Espanjassa, mutta heti alkuun on todettava, että kelit eivät olleet sitä, mitä Espanjalta saattaisi odottaa. Ei siis palanutta nahkaa eikä kylmiä juomia, vaan 3–4 päivää tiukkaa asiaa.

Konferenssin miljöö oli historiallinen, onhan Salamancan yliopisto perustettu jo vuonna 1218 ja kaupunkin itsessään Unescon maailmanperintökohde. Ohjelmassa saatiin tutustua myös espanjalaisiin arkistoihin eri näkökulmista. Arkistotoimi Espanjassa näyttäytyi vahvasti hajautettuna ja moniulotteisena. Maassa on yhteensä 36 486 arkistoa ja yli 100 lakia tai asetusta, joita arkistointiin sovelletaan.

Edellisestä Triennialesta oli vierähtänyt jo kuusi vuotta. Tuona aikana tekoäly on vallannut tilaa myös arkistoalalla. Tämä näkyi avauspuheenvuoron pitäneen DLM Forumin puheenjohtajan Anja Paulicin mukaan myös ehdotetuissa puheenvuoroissa ja konferenssin tee-

DLM Triennal Salamancassa: Arkistot digiyhteiskunnassa

maksi oli valittu "Keeping and connecting: data in the digital society". Avaus sisälsi luonnollisesti kiitokset sekä järjestäjille että konferenssin Platinum ja Gold sponsoreille (Artefactual, Libnova, Preservica, Odilo ja Xercode).

Data ja digitalisaatio arkistoissa

Konferenssin sisällöllisesti avannut Espanjan kansallisarkiston johtaja Ana López Cuadrado kuvasi tavoitteenamme olevan modernisoida arkistojen rooli ja toiminta digitalisoituvassa yhteiskunnassa. Hän näki digitaalisten aineistojen ja tekoälyn muuttavan myös arkistoasiantuntijoiden toimenkuvia ja työprosesseja. Myöhemmissä puheenvuoroissa toistui ajatus siitä, että digitaaliset aineistot ovat erilaisia kuin paperiaineistot, ja siksi niiden arkistojenkin pitää olla erilaisia.

Ajatuksia herätti myös puheenvuoro arkistoyhteisön ulkopuolelta. José Manuel Alonso Espanjasta on työskennelyt pitkään konsulttina ja asiantuntijana avoimen datan ja avoimen webin puolesta. Digitalisaatiossa huomionarvoista on tällä hetkellä raju kasvu datan määrässä sekä muutos siinä millaista dataa syntyy ja missä. Valtaosa verkkoliikenteestä syntyy videoiden striimauksesta ja botit pystyvät luomaan valtavia määriä uutta sisältöä lyhyes-

Plaza Mayor, Salamancan keskusaukio sateen jälkeen yöllä.

sä ajassa. Josén mukaan meneillään olevassa digimurroksessa arkistojen kannattaa hyödyntää se luottamus, joka kansalaisilla ja asiakkailla jo on niihin. Arkistoissa kannattaa toisaalta myös varautua siihen, että tiedonhallinta tulee olemaan vaikeampaa kuin ennen. Miten niin? Periaatteessahan esimerkiksi avoimuuden vaatimukset eivät ole arkistoille mitään uutta. Nyt kuitenkin arkistoidaan jo aiemmin avoimesti saatavilla olleita aineistoja, joiden pitäisi siis olla välittömästi (katkotta) saatavilla myös digiarkistosta. Oletus on siis, että sekä käyttäjien odotukset että aineistojen asettamat vaatimukset kasvavat.

Tekoälyn mahdollisuudet ja uhkakuvat

Tekoälytyöpajan nimellä pidettiin tavallista pidempi esittely siitä, mitä tekoälyllä voi tehdä. Laura Esparza Sainz (Espanjan kansallisarkisto) kertoi yhdessä Joan-Andreu Sánchezin ja Enrique Vidalin (Valencia Polytechnic University & tranSkriptorium AI SL) kanssa käsinkirjoitetun teksin prosessointitekniikoista. Jos sisältö aukesi oikein, he olivat opettaneet alusta asti itse oman mallinsa käsinkirjoituksen tunnistamiseen hyödyntäen PRIX (Probabilistic Indexing) -menetelmää. Tällä menetelmällä saatiin havainnoista ulos sellaista tilastollista tietoa, jota voitiin hyödyntää sanahypoteesien tekemiseen NER- ja NLP-tekniikoilla. Muuten tämä vaikutti oikein lupaavalta, mutta mieleen herää pakostakin kysymys, miksi taustalla ei oltu hyödynnetty jotakin olemassa olevaa käsinkirjoitetun tekstin foundation-mallia, joita löytyy mm. Hugginface-palvelusta.

Työpajan lisäksi tekoälystä puhuttiin myös Joséma Alonson keynotessa. Hänen esityksensä alkoi Coca-Colan Y3000-maulla, joka on siis suunniteltu yhteistyössä tekoälyn kanssa. Hän myös totesi kuivausrumpunsa hajonneen ja uudessa Samsungin laitteessa olevan AI Dry -toiminnon, josta hän ei osannut sanoa mitä se edes on. Pikainen käynti Samsungin sivuilla kertoo ”using AI Dry helps reduce wrinkles, while still being gentle on your clothes”, joka on ilmeisesti toteutettu erilaisten sensoreiden yhteistoiminnalla. Mutta emme siis edelleenkään tiedä mikä AI-malli on taustalla, miten se oikeasti toimii jne. Periaatteessa tilanne on sama minkä tahansa tekoälymallin kanssa, koska ne ovat mustia laatikoita, joissa on miljardeja yhteyksiä mallin oppimien asioiden välillä.

Generatiivinen tekoäly on tällä hetkellä Gartnerin Hype Cycle -listalla kohdassa Peak of Inflated Expectations kuten myös AI-Augmented. Jo aikaisemmin mainittu José kehotti kuitenkin pitämään mielessä tekoälyn potentiaalin ja mahdollisuuksien lisäksi siihen sisältyvät uhkakuvat, joista muutamina mainittakoon läpinäkyvyyden puute, vinoumat ja arvaamaton toiminta. Vastaavia listauksia löytyy netistä paljon, joten absoluuttisia totuuksiahan ne eivät toki ole. Kuitenkin monien ohjelmistojen taustalla on jo tekoälyä. Microsoftin tuotteisiin saa, tai on tulossa Copilot, Photoshopin monet kuvankäsittelyalgoritmit perustuvat tekoälyyn ja ChatGPT ja Bard osaavat luoda valmista koodia käytännössä ohjelmointikielestä riippumatta. Tekoälyn avulla on tehty myös monia KAM-sektorin alan pilotointeja USA:ssa liittyen mm. PII (Personnally Identifiable Information) -tunnistuksessa, kuvailevien metatietojen automaattisessa täyttämisessä ja tiivistelmien tekemisessä. Tarkemmat tiedot kokeiluista löytyvät NARAn sivuilta *https://www.archives.gov/data/ai-inventory*.

Konferenssissa järjestettiin myös paneelikeskustelu eettisistä näkökulmista tekoälyn hyödyntämiseen arkistoissa. Tekoälyn etiikasta on olemassa monenlaisia malleja, mm. OECD:n (*https://oecd.ai/en/ai-principles*) ja EU:n. Nämä ovat kuitenkin kovin ylätasolla, ja käytännön soveltaminen arkistoissa on haastavaa. Paneelin osallistujat jakoivat omia kokemuksiaan siitä,

Konferenssipaikkana toimi Salamancan yliopiston Maantieteen ja historian tiedekunnan rakennus.

Ensimmäisen päivän aloituksessa oli sali pullollaan.

miten esimerkiksi työpajatyöskentelyllä voidaan asiaa työstää osaksi kunkin omaan työtä tiedonhallinnan parissa. Tekoälyn nähtiin vaikuttavan myös siihen, mitä arkistoidaan: paneelissa keskusteltiin esimerkiksi siitä, tulisiko itse tekoälymalli pyrkiä arkistoimaan.

Käytännön työkaluja arkistoinnin tueksi

Alankomaiden kansallisarkisto esitteli neljän eri sosiaalisen median arkistoinnin työkalun (Archive-It, Archiveweb.page, Browsertrix, Webpreserver) vertailua. Kävi ilmi, että vastaavia vertailuja on tehty ja tehdään monessa paikassa. Yleisöstä ehdotettiinkin yhteiseurooppalaista suositusta työkaluvalintaan. Alankomaissa oli päädytty käyttämään ArchiveWeb.page-työkalua Webrecorder suiten kautta. Keskeisenä nähtiin se, että arkistoon saadaan sosiaalisen median viestien ulkoasu, pelkkä rajapintojen kautta saatu tekstisisältö ei siis riitä. Erityisen kiinnostava oli Elisabeth Klindworthin (Archives of the Max Planck Society) pitämä esitys EMILIA-työkalusta sähköpostien arkistointiin, koska kuten pitkäaikaisimmat *Failin* lukijat muistanevat myös Digitalialla on oma työnkulkunsa sähköpostien arkistointiin. Elisabeth pyysi myös kaikkia yhteistyöstä kiinnostuneita olemaan heihin yhteydessä.

Maiju Pohjola Kansallisarkistostamme esitteli DALAI-hankkeen tuloksia hankkeen päätyttyä. Hankkeessa kehitetyt työkalut tunnistavat digitoiduista aineistoista skannausvirheitä ja sisältöjä. Niitä voi kokeilla Arkkiivi-demossa (*https://arkkiivi.fi/*). Yleisö ehdotti seuraavaksi tutkimuskohteeksi asiakirjojen ja asioiden suhteiden automaattista tunnistusta. Keskustelua herätti myös tyhjien sivujen käsittely digitoinnin ja käyttöön tarjoamisen yhteydessä: ovatko ne turhaan tallennustilaa vieviä vai olennainen osa alkuperäistä aineistoa?

Seuraavat tapaamiset

Kaikki Triennalen esitykset ovat jaossa DLM Forumin jäsenille yhteisön verkkosivuilla. Seuraavat DLM-jäsentapaamiset pidetään Belgiassa ja Unkarissa, järjestelyvuoron kiertäessä EU-puheenjohtajuuskausien mukana. Belgian tapaamisen ajankohta on todennäköisesti keväällä 27–28.5.2024 ja siihen yhdistetään EARK- ja EAG-päivät. Kannattaa siis merkitä ajankohta jo alustavasti kalenteriin.