Tähtitieteen data ja sen linkitykset julkaisuihin

Eva Isaksson

Suomi lähti kunnianhimoisesti mukaan suureen kansainväliseen tähtitieteen dataprojektiin. Kuulostaako tämä tuoreelta uutiselta? Kovin tuoreesta tapahtumasta ei ollut kyse, sillä vuosiluku oli 1890. Historioitsijat pohtivat vieläkin, söikö tämä vuosikymmeniä kestänyt suurhanke suomalaisen tähtitieteen voimavarat. Kaikki tutkimusdataan liittyvät elementit olivat jo tuolloin olemassa. Tämä prosessoitu data päätyi julkaisuihin, pinoon kookkaita katalogeja.

Jo 1800-luvun lopussa tarjolla oli raakadataa, eli suuren valokuvakartoituksen Helsingin taivaanviipale, jota kuvattiin kylminä talviöinä lasilevyille. Tutkijoiden lisäksi tuli laskijoita, joukko ahkeria naisia jotka mittasivat dataa levyiltä ja redusoivat sitä taulukoiksi. Painetut katalogit löytyvät kirjastoista.

Entä jos joku haluaisi päästä käsiksi alkuperäiseen raakadataan? Aivan ensiksi pitäisi selvittää, missä se on. Mikään luettelotieto ei asiaa paljasta, ja kaikkein vähiten nämä astrofotograafisen kartoituksen painetut luettelot.

Helsingin Observatorion entiset työntekijät ehkä sentään tietävät, että lasilevyt siirrettiin joulukuussa 2009 Kumpulaan, Fysiikan laitoksen kellariin. Seuraava haaste olisi löytää vanhoista puulaatikosta jokin ”datasetti” eli yksittäistä taivaankohtaa esittävä valokuvauslevy, joka toivottavasti saattaa vielä olla käyttökelpoinen.

Taivaan data on avointa kaikille

Yleensä kun puhutaan datasta, ajattelemme digitaalisia tietomassoja. Tähtitieteessä syntyy paljon havaintodataa, joten sitä alettiin tallentaa biteiksi heti tilaisuuden tullen. NASAn kuulennoilla syntynyttä kuvamateriaalia alettiin 1970-luvulla tallentaa uudenlaiseen, avoimen lähdekoodin formaattiin, joka julkistettiin 1981 FITS-formaattina (Flexible Image Transport System).

FITS vakiintui nopeasti tähtitieteellisen datan perusformaatiksi, ja on niin vakaa, että nykytähtitieteilijä pystyy huoletta käsittelemään kolmen vuosikymmenen ikäisiä FITS-tiedostoja. Sen etuihin voidaan lukea mm. se, että tiedostot voivat olla todella suuria kooltaan, ja havaintolaitteet voivat kirjoittaa metadataa suoraan näihin tiedostoihin. FITS Liberator –ohjelmalla jopa pelkkä näppärä harrastelija voi muokata itselleen kuvia kaukoputken tuottamasta raakadatasta.

Tähtitieteen datan avoimuus juontaa juurensa siitä syvään juurtuneesta näkemyksestä, että taivas on avoin kaikille, eivätkä tähdet ole kenenkään omaisuutta. Tutkijat on ollut helppo saada jakamaan dataa keskenään. Toisaalta ne laitteet joilla tämä avoin data on saatu kuuluvat kaikkein kalleimpiin tutkimuslaitteisiin, joita ei lähetetä avaruuteen pikkurahalla.

Tähtitiede tuottaa suuria määriä dataa

Yleensä laitehankkeiden takana ovat isot yhteisiä varoja kanavoivat organisaatiot kuten NASA, ESA (Euroopan avaruusjärjestö) tai ESO (European Southern Observatory). Satojen tuhansien tai miljardien eurojen hintaiset laitteet ovat kukin ainutkertaisia lajissaan. Niillä tehtävät taivaankartoitukset kattavat usein tietyn aallonpituuden ja koko havaittavan taivaan ja tuottavat datamääriä, joita tyypillisesti mitataan petatavuissa. Mikäli havaintolaitteen suuntaamiseen haluaa vaikuttaa, tutkija joutuu anomaan ankarasti kilpailtua havaintoaikaa. Alan käytännön mukaan tällaiselle havaintodatalle saa vuoden mittaisen

Kuva: Morguefile.com / FlyingPete

suoja-ajan, sitten data siirtyy kaikkien saataville.

Lähitulevaisuudessa toteutetaan yhä valtavampia tähtitieteen havaintoprojekteja. Etelä-Afrikkaan ja Australiaan on rakenteilla mittavia radioteleskooppipeltoja. Valmistuttuaan tämä SKA-hanke tuottaa enemmän dataa kuin mitä koko nykyisessä internet-tietovuossa on liikenteessä.

Vielä 2000-luvun alussa tähtitieteen tutkimuksessa uskottiin ns. virtuaaliobservatorioihin. Kaikki taivaalta tallennettu data olisi tietoverkon kautta jokaisen tähtitieteilijän hyppysissä. Virtuaaliobservatorioprojekteja käynnistettiin eri puolilla maailmaa. Sitten Yhdysvaltain kansallinen tiedesäätiö NSF leikkasi 2012 kansallisen virtuaaliobservatorion rahoituksen viidesosaan. Vaikka tutkimusdata on päivän kuuma sana, sen saatavuus voi olla milloin hyvänsä vaarassa, kun jokin keskeinen rahoittaja sulkee rahahanat. Alalle on yllättäen ilmaantunut yksityisiä toimijoita. Microsoft käynnisti 2008 ”World Wide Telescope” –hankkeen, jossa hyödynnetään monia virtuaaliobservatoriota varten kehitettyjä työkaluja.

Tähtitieteen datakeskukset syntyvät

Kirjastonäkökulmasta meitä kiinnostaa, miten kaikki tämä data on järjestetty, ja miten se linkittyy julkaisuihin. Euroopassa ongelmaa on pohtinut vuodesta 1971 lähtien Strasburgissa toimiva tähtitieteen datakeskus CDS (Centre des Données Astronomiques). Sen piirissä toimii tähtitieteilijöitä, IT-asiantuntijoita ja informaatikkoja, jotka kaikki tuovat peliin oman erityisosaamisensa.

CDS:n tietoranteiden suunnittelun lähtökohtana on ollut, että sekä julkaisuissa että havaintodatassa puhutaan kohteista. Taivaalta löytyvät valopisteet tai muunlaiset, himmeämmät läiskät saavat yleensä jonkin tunnisteen tai useampia. Strasburgissa alettiin pitää kirjaa kohteiden nimistä ja kerätä niitä tietokannaksi. Näitä tunnisteita haravoidaan tähtitieteen julkaisujen kokoteksteistä DJIN-nimisellä ohjelmalla, ja tarvittaessa voidaan lennosta luoda kokonaan uusia tunnisteita. Kaukaiset kohteet eivät meidän aikaskaalamme mukaan juurikaan liiku paikoiltaan, joten tietyissä tähtitaivaan koordinaateissa sijaitseva kohde erilaisine havaintodatoineen voidaan yhdistää tiettyyn tunnisteeseen.

CDS on tuottanut tähtitieteilijöiden käyttöön monia datatyökaluja: kohteet kattava SIMBAD, havaintotaulukoita keräävä VizieR ja vuorovai-

kutteisesti käytettävä Aladin-taivaankartasto. Nämä kuuluvat alan tutkijoiden perustyökaluihin. Uuden artikkelin ilmestyessä sen tekstissä esiintyvien kohteiden nimet tunnistetaan ja artikkelit linkittyvät jatkossa suoraan näihin kohteisiin. Myös suurimmat tähtitieteen kustantajat osallistuvat CDS:n työhön mm. huolehtimalla siitä että kohteiden ja taulukoiden tiedot on mahdollisimman helppo poimia teksteistä.

ADS kokoaa kaikki maailman tähtitieteen julkaisut

Atlantin toisella puolen NASAn rahoittama Astrophysics Data System (ADS) puolestaan on rakentanut viitetietokannan, joka sisältäisi kaikki maailmassa ilmestyneet tähtitieteen julkaisut. Näitä oli 2013 noin 10 miljoonaa viitettä. ADS käynnistettiin 1991. Se on pyrkinyt alusta asti rakentamaan linkkejä muihin tietokantoihin. Muut toimijat (CDS, kustantajat, preprint-tietokanta arXiv, tähtitieteen suuret havainto-ohjelmat) ovat sen yhteistyökumppaneita. Niinpä SIMBADiin haravoitujen kohdetunnisteiden avulla voidaan suoraan hakea kirjallisuusviitteitä ADSista. Tämä hienosti hiottu kokonaisuus on sitonut tähtitieteen tutkijat niin tiukasti verkkoonsa, ettei tähtitieteilijä yleensä tietoa hakiessaan harhaannu muualle ADSin ääreltä.

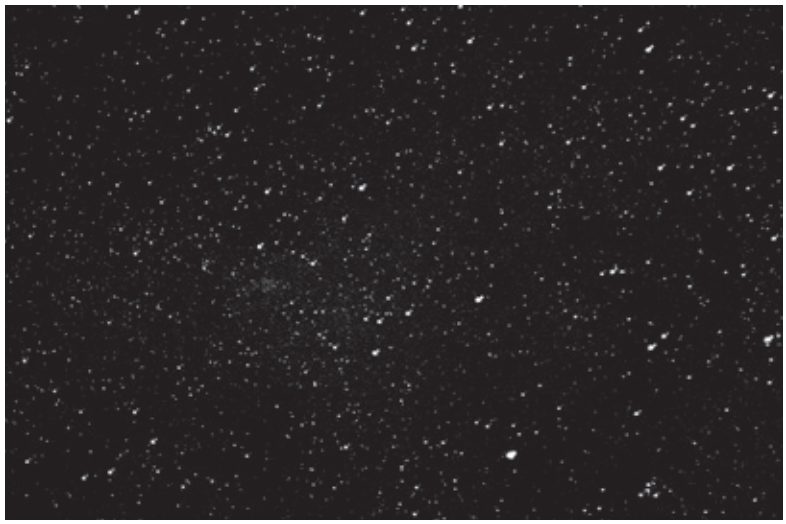
ADSiin alettiin lisätä datalinkkejä vuodesta 1997 lähtien. Tämä tarkoittaa sitä, että kyseisissä artikkeleissa on ollut linkki verkossa löytyvään dataan. Aluksi kyseessä oli vain muutamia kymmeniä linkkejä, mutta muutamassa vuodessa datalinkkien määrä alkoi lisääntyä. Nykyään niiden määräksi on tasaantunut n. 1500 linkiksi vuosittain (Pepe et al. 2013).

Käytännössä kyse on kahdenlaisista linkeistä. Datalinkki voi sijaita keskitetysti ylläpidetyllä palvelimella, josta sen voi olettaa löytyvän jatkossakin. Näin on varsinkin silloin, kun kyse on käsittelemättömästä datasta ja koko taivaan kattavista kartoituksista. Data voi myöskin olla jo kertaalleen käsiteltyä. Tällöin se päätyy herkästi yksittäisten tutkijoiden kotisivuille tai usein lyhytikäisiin ftp-arkistoihin. ADSIn vanhimmista datalinkeistä jopa yli puolet oli jossain vaiheessa ehtinyt muuttua kuolleiksi linkeiksi. Suurin osa näistä oli nimenomaan yksittäisten tutkijoiden ylläpitämiä linkkejä. Sen sijaan keskitetyissä palveluissa säilytetyn datan jatkosaatavuus oli melko vakaata vuosienkin jälkeen.

Datalinkkeja metsästämässä

Alkuperäisen, käsittelemättömän datan käyttäminen ei tähtitieteessä aina ole mahdollista tai käytännöllistä. Sitä voi olla yksinkertaisesti niin paljon, ettei sitä kannata säilyttää sellaisenaan saatavilla. Monelle riittää jo kertaalleen käsitelty data. Erään tähtitieteilijän sanoin: ”Moniko on koskaan käsitellyt SDSS-kuvatiedostoja uudestaan? Antakaa kun arvaan: maa kantaa pinnallaan enintään kymmenen ihmistä jotka ovat ikinä käsitelleet Sloan [Digital Sky Survey] –kuvia uudelleen.”

Eräs keskeinen syy dataviittauksiin on niiden artikkelille tuoma lisäarvo. Henneken ja Accomazzi (2011) tarkastelivat 3814 datalinkattua artik-



Kuva: Morguefile.com / seriousfun

kelia, jotka oli julkaistu 1995-2000 ja vertasivat niiden saamia viittauksia mahdollisimman hyvin näitä vastaaviin linkkaamattomiin artikkeleihin. Kymmenessä vuodessa datalinkatut artikkelit saivat noin 20% enemmän viittauksia kuin verrokit.

Dataviittausten seuranta kiinnostaa erityisesti suuria laiterahoittajia. Lähes kaikilla suurilla tähtitieteen laitteistoilla on nykyään oma kuratoitu viitetietokanta, johon listataan ne artikkelit, joissa on käytetty näillä laitteilla kerättyä dataa. Artikkelit linkitetään muihin viitetietokantoihin (ADS tai Web of Science) joista saadaan viittausmäärät. Näin yksittäisten havaintolaitteistojen tuottavuutta voidaan seurata.

Myös laitteita käyttäneistä tutkijoista voidaan hakuprosesseissa saada kiinnostavia tietoja. Jos vaikkapa suomalaiselta hakijalta toivotaan näyttöä ESO:n havaintolaitteiden käyttökokemuksesta, ESO Telbib-tietokanta kertoo, paljonko ESO:n havaintolaitteilla saatua dataa tutkija on käyttänyt, ja onko se raakadataa vai ehkä muiden käsittelemää.

Seuraava askel tähtitieteen datapalvelun palapelissä on semanttisten teknologioiden tehostettu soveltaminen yhä kasvavaan määrään kokotekstejä. Tietomassojen kasvaessa on viisasta sijoittaa datan hallintaan mahdollisimman aikaisessa vaiheessa, jotta kalliiden havaintolaitteiden tuottama tieto voidaan hyödyntää mahdollisimman hyvin. Kallista dataa ei kannata unohtaa kellareihin! 🕮

Lähteet

E. Henneken & A. Accomazzi (2011), Linking to Data - Effect on Citation Rates in Astronomy. arXiv:1111.3618 [cs.DL]

A. Pepe et al. (2013), Sharing, archiving, and citing data in astronomy. http://authorea.com/288

Linkkejä

ADS: http://labs.adsabs.harvard.edu/adsabs/
CDS: http://cdsweb.u-strasbg.fr/
World Wide Telescope: http://www.worldwidetelescope.org
ESO Telbib: http://telbib.eso.org/

Tietoa kirjoittajasta

Eva Isaksson, kirjastonhoitaja
Helsingin yliopiston kirjasto
Email. Eva.isaksson@helsinki.fi