Uudenlaista indeksoinnin tehokkuutta geotieteissä

Multilingual Thesaurus of Geosciences. Second edition. • Deutsch • English • Español • Français • Italiano • Russkij • Sponsored by ICSTI and IUGS; edited by J. Gravesteijn, C. Kortman, R. Potenza, and G.N. Rassam. Medford, N.J.; Information Today, Inc. 1995. 645 s.

Eipä uskoisi jonkin tesauruksen lueskelun ja selailun kuuluvan elämyksellisen lukemisen piiriin, mutta niin vain kävi, että yllä mainittu geotieteiden monikielinen tesaurus oli minulle monen ahaa-elämyksen lähde. Kun ikää karttuu, alkaa katsella asioita yhä useammin myös historiallisesta perspektiivistä, muistaa ihmisiä vuosikymmenten takaa ja oppii asettamaan omallekin kohdalle sattuneita yksittäisiä asioita yleisempään, tässä tapauksessa tieteellisen tiedon välityksen kehityksen kehikkoon.

Tulin informaatioalalle työhön 7 vuotta ennen Sputnikia ja siirryin eläkkeelle intensiivisen tietoverkottumisen alun aikaan. Tuollaiseen rupeamaan mahtuu monta harppausta tieto- ja tietoliikennetekniikan kehityksessä. Se taas on vaikuttanut monin tavoin tieteellisen tiedon välittymistä ja välittämistä koskevaan tutkimus- ja kehittämistyöhön. Perusongelmat tosin ääriviivattiin jo pari vuosisataa sitten, mutta sekä ongelmat että ratkaisut ovat aina olleet sidoksissa kunakin aikana käytettävissä olleeseen teknologiseen varustukseen.

Kun toisen maailmansodan jälkeen tieteellinen julkaisutoiminta alkoi lisääntyä rajusti, kansainvälinen tiedeyhteisö heräsi 1940-luvun viimeisinä vuosina pohtimaan tieteellisen tiedon tavoitettavuuden ongelmia. Ratkaisua haettiin tieteellisten julkaisujen laadun kohottamisesta. Yritettiin vaikuttaa sekä tutkijoihin kirjoittajina että eri alojen yhteisöihin julkaisijoina. Ideana oli jäntevöittää alkutuotantoa saamalla yksittäiset tutkijat kirjoittamaan entistä jämäkämpää tekstiä ja neuvomalla päätoimittajia asettamaan julkaisemiskynnyksen entistä korkeammaksi.

Tämä julkaisemistaitojen parantamisen virtaus oli vallalla pitkästi 1960-luvulle ja seuraavallekin vuosikymmenelle, mutta rinnalle alkoivat jo nousta eritoten suurten bibliografisten kustantajien ansiosta myös ideat eri tiedeyhteisöjen velvollisuudesta parantaa tiedon tavoitettavuutta dokumentaation keinoin ja osallistua omien alojensa tiedon tallennus- ja hakujärjestelmien kehittämistyöhön. Nyt esiteltävän monikielisen tesauruksen juuret ulottuvat tuohon saumakohtaan.

Tesaurus on tuotettu Milanossa ylläpidettävästä Multilingual Thesarus- eli MT-tietokannasta, joka sisältää otsikossa mainittujen kuuden kielen lisäksi myös suomen- ja tsekinkieliset hakusanastot. Pienten kielialueiden sanastot eivät ole kiinnostaneet painetun tesauruksen kustantajaa, vaikka tämä toimiikin sponsoroituna ns. non-profit -kustantajana. Valinta on ymmärrettävä eikä asialla ole edes väliä, koska kaikkiin kahdeksaan hakusanastoon pääsee käsiksi suoraan tietokannasta. Sillä taas on väliä, minkälaiseen indeksointifilosofiaan on päädytty ja miten tesaurus ajatuksellisesti ja tietoteknisenä ratkaisuna toimii käytännössä.

Minun on mahdotonta esitellä tämän tesauruksen rakennetta kertomatta ensin sen toimituskunnan jäsenen ja hakusanatyön koordinoijan J. Gravesteijnin indeksointiajattelusta. Tapasin tämän hollantilaissyntyisen, "monikielisen" geologin ensimmäisen kerran vuonna 1969 Ranskassa, mistä hän oli löytänyt leipäpuun sikäläisen Geode -tiedonhakujärjestelmän kehittämisestä. Kun itse olin tuolloin tehnyt vain UDK-luokitusta ja yrittänyt kehittää sitä FID:n puitteissa, kävi elämyksestä perehtyä Gravesteijnin, minulle uudenlaiseen luokitusajatteluun ja erityisesti hänen kehittelemiinsä hakukäsitteiden nuolikaavoihin. Hänen kuningasajatuksensa oli muotoilla geotieteille yhdistetty luokitus- ja indeksointijärjestelmä niin, että se rakentuisi pääluokkien hakukäsitteistä, jotka ilmaistaisiin hakusanojen avulla. Tämä erottelu olisi tietysti pitkälti sopimuksenvaraista, mutta ajan mittaan eri käsitteitä kuvaavat hakusanat vakiintuisivat indeksoinnin käytännössä. Tämä perusidea elää monikielisessä tesauruksessa.

Monikielisen tesauruksen tarkoituksena on mahdollistaa indeksoitujen viitetietojen vaih-

to eri kielien ja bibliografisten tallennusjärjestelmien kesken siten, että hakusanojen vastaavuus käsitetasolla säilyy.

Valtaosan kirjan sivuista vie aakkosellinen, englannin kielellä esitettyjen käsitteellisten avainsanojen (key terms) luettelo. Avainsanoja on 5823. Kullakin on oma numero, joka sitoo yhteen kaikki kuusi erikielistä, samalla rivitasolla ilmoitettua termiä. Avainsanat eivät itse ole hakuelementtejä. Ne ovat kirjaimellisesti avaimia tesaurushankkeessa mukana olevien järjestelmien hakusanoihin.

Järjestelmät taas perustuvat erilaisiin indeksointifilosofioihin. Siksi monikielisessä tesauruksessa osoitettu käsitevastaavuus ei merkitse, että ao. käsite olisi joka järjestelmässä myös vakioitu hakusana. Näissä tapauksissa taulukoissa ohjataan tuttuun tapaan käytössä olevaan hakusanaan ja sen kohdalla mainitaan ne termit, joiden asemasta vakioitua hakusanaa käytetään.

Kunkin avainsanan kohdalla ilmaistaan nelikirjaimisella koodilla myös se, mihin aihepiiriin käsite kuuluu. Aihepiirejä eli pääluokkia on 36. Lähes kolmannes pääluokista, esim. Mineraalit, perustuu ammattipiireissä kansainvälisesti hyväksyttyyn systematiikkaan. Nämä systematiikat on esitetty kirjan loppupuolella erillisinä hierarkkisina luetteloina, mitä pidän hyvänä hakutermien käyttäjän apuneuvona. Niiden avulla on helppoa siirtyä laajempiin, suppeampiin ja rinnakkaisiin hakusanoihin. Osa pääluokista taas käsittää joukon nimeltä mainittuja kohteita kuten esim. luokka Tutkimusmenetelmät ja -laitteet sekä kaikkien järjestelmien voittamaton paha, Miscellanea-luokka. Muut pääluokat koostuvat käsiteryppäistä ja niiden sisäisistä pienoishierarkioista eli ne asettuvat lähestymistavan puolesta em. äärityyppien väliin.

Pääluettelon jälkeen on apuluetteloita, ensin kunkin kielen ja järjestelmän hakusanat aakkosjärjestyksessä. Niinpä englanninkielinen hakusanaluettelo ei suinkaan toista avainsanaluetteloa, vaan se käsittää amerikkalaisessa GeoRef-järjestelmässä käytetyn hakusanaston. Seuraava luettelo esittää kuhunkin pääluokkaan kuuluvat hakutermit, samaten aakkosjärjestyksessä, ja niitä seuraa jo mainitut hierarkkiset luettelot.

Viimeisenä on indeksoinnissa käytettyjen, eri lähtökohdista määritettyjen alueellisten käsitteiden luettelo. Geotieteelliset suuralueet eivät noudata valtiollisia rajoja, mistä hyvä esimerkki on se peruskalliokilpi, joka ulottuu Ruotsista sen koko pituudelta koko Suomen poikki Itä-Karjalaan Äänisjärven seudulle ja Vienanlahteen asti. Käytännön syistä tutkimukset keskittyvät enimmäkseen yhden valtion alueelle. Mittakaava vaihtelee universaalisesta globaaliseen, globaalisesta suuralueelliseen, siitä yhteen, maantieteellisin koordinaatein ilmaistavaan havaintopisteeseen.

Monikielisen tesauruksen kiehtovin piirre on se, että siitä paljastuu miten tietyt asiat käsitteellistetään eri tavalla eri kielialueilla. Itse asiassa käsitteellistämisestä kuultaa läpi asianomaista kieltä käyttäen harjoitetun geologisen tutkimuksen historia ja samalla myös kielialueen geologiset erityispiirteet, lisäksi myös terminologinen vuorovaikutus kielten välillä.

Vaikka mikään tesaurus ei itsessään ole termisanakirja, tämä monikielinen hakusanasto tarjoaa kuitenkin myös terminologiselle työlle hyvän pohjan. MT-tietokannasta voidaan näet tulostaa kootusti tietyn aihepiirin keskeiset käsitteet ja niitä vastaavat erikieliset ilmaisut ja sitten tutkia termien käsitteellisiä ulottuvuuksia eri kielissä. Valaisen asiaa esimerkillä.

Pääluokassa Hyödykkeet ja mineraaliesiintymät on bentoniittiesiintymän käsite. Sitä vastaavien hakusanojen tutkistelu osoittaa, miten saksalainen geologi mieltää tällaisen esiintymän raaka-aineen lähteeksi, ranskalainen ainekseksi ja aineskertymäksi, amerikkalainen kerrostumaksi muiden joukossa, mutta venäläinen geologi sekä kerrostumaksi että "malmiksi", ainakin ei-ammatillisessa puheessa myös syntymäpaikaksi. Millä tavalla itse itsensä selittävä on suomenkielinen esiintymä -sana terminä ja hakusanana?

Suomenkielisinä hakusanat ja luettelot ovat MT-tietokannassa rinnan seitsemän muun kielen kanssa. Tesaurusta kannattavan käsitteellisen yhteensopivuuden ansiosta geologian tutkimuskeskuksen (GTK) informaatiotoimisto voi tämän tietokannan kautta hakea

viitetietoja monen kielen avulla useasta viitetietokannasta, siis käyttää hyväksi yht'aikaa eri järjestelmien toisistaan poikkeavia rakenteita ja erilaisia viitekatteita.

Varsin merkittävä MT-tietokannan käyttösovellus on se, että em. toimisto indeksoi toimialansa julkaisut englanniksi GeoRef-järjestelmään, muuntaa tulokset suomenkielisiksi hakusanoiksi muutamalla näppäimen painalluksella ja asettaa viitetiedot – vips! – käytettäviksi kansallisissa viitetietokannoissa (ARTO ja FinGeo). Tätä on pidettävä taloudellisuudessaan verrattomana innovaationa. Se saanee ajan mittaan laajalti käyttöä, kunhan tarvittavat muiden alojen monikieliset tesaurustietokannat ensin saadaan kootuksi. Se tosin vie reippaanlaisesti aikaa.

Tämän tesauruksen ensimmäinen painos ilmestyi 1988. Sen valmistamiseen kului vuonna 1976 työnsä aloittaneelta työryhmältä 12 vuotta. Tätä versiota on tutkittu monin tavoin, ennen kaikkea seuraamalla sen hakusanojen esiintymistiheyttä indeksoinnissa sekä indeksoijien tuottamien uusien hakusanojen ja niitä vastaavien käsitteiden esiintulon vakiintumista. Vähän käytetyt käsitteet on poistettu yhtä perustellusti kuin uusia on otettu mukaan. Perusrunko on kuitenkin 19 vuoden uurastuksen jälkeen valmis.

Nyt valmistuneen toisen painoksen käsitteistö muuttuu toki jatkossa, sillä uusille ideoille, innovaatioille ja entisestä poikkeaville löydöksille on aina raivattava tilaa. Niin teoreettisesti hyvin perusteltu kuin monikielisen tesauruksen perusrakenne onkin, se on sittenkin siihen vihkiytyneiden aivoitusten tuote. Sen monet käyttömahdollisuudet eivät välttämättä avaudu kaikille hakijoille, esim. ns. loppukäyttäjille. Missään tesaurusta koskevassa esitteessä tms. en ole löytänyt mainintaa aikomuksesta analysoida käyttäjien suorittamia hakuja. Ne ovat usein melkoisen hakuammunnan, erehdysten ja niiden korjaamisen tilkkutäkkiä. Siitäkin olisi hyvä tietää erityisesti tesaurukseen valittujen käsitteiden käyttäjärelevanssin arvioimiseksi.

Indeksointi on varsin kallista työtä. Siksi taannehtivaan indeksointiin ei useimmiten voida osoittaa varoja. Uusin MT-tietokannan käyttösovellus on aihehakujen tekeminen monen kielisiä otsikoita sisältävistä julkaisuluetteloista, joita ei kuunaan ole indeksoitu. GTK on teettänyt tietokannan Aarne Laitakarin julkaisemasta Suomen geologisesta bibliografiasta vuosilta 1934–1970. Tällä tavalla FinGeo-tietokanta, joka ulottuu vuoteen 1971 ja käsittää indeksoituja viitteitä, on saanut "ikälisän", sitä edeltävien 36 vuoden aikana kertyneiden julkaisujen viitetiedot ja otsikkohaut niistä. Hakutulokset ovat jo nyt varsin hyviä, mutta selvää on, että monikielisen tesauruksen toimivuutta otsikkohaussa on parannettava kehittämällä synonyymisanastoja.

Kauan tieteellisiä julkaisuja toimittaneena tiedän, että on kirjoittajia, joita mikään mahti maailmassa ei saa luopumaan lennokkaista otsikoista, vaikka niistä ei käy riittävän selvästi ilmi, mitä tuli tutkituksi. On myös (ollut?) toimituksia, jotka päästävät läpi harhaanjohtavia otsikkoja. Minun esimerkkikokoelmani helmi on otsikko, joka toisti paikallislehdessä olleen uutisankan otsikon, mutta artikkeli kertoi mistä ilmiöstä itse asiassa oli kysymys. Tällaisiin otsikkoihin ei synonyymisanastokaan pure. Entä sitten? Syyttäköötä tekijät itseään!

Vielä on jäljellä yksi näkökohta. Geotieteiden monikielinen tesaurus on perusrakenteen puolesta valmis. Se on valmis työkalu kaikille niille maille (kielialueille), joissa omat tämän alan informaatiojärjestelmät ovat jääneet kehittämättä tai joiden järjestelmät halutaan saattaa sekä kansainvälisesti että kansallisesti yhteensopivaan muotoon. Marssijärjestyksessä Suomi kuuluu edelläkävijöihin.

Toimituskunnan suomalainen jäsen, GTK:n informaatiotoimiston toimistopäällikkö Caj Kortman on osallistunut tesaurustyöhön nuo jo mainitut 19 vuotta, vetänyt hanketta kansainvälisen geologiunionin puitteissa ja kehitellyt GTK:ssa työtoveriensa kanssa MT-tietokannalle edellä kuvaamiani uusia käyttötapoja. Toivottavasti en aiheuta painajaismaiseksi muuttuvaa ryntäystä, kun kehotan asiasta kiinnostuneita tutustumaan GTK:n informaatiotoimistossa monikielisen tesaurustietokannan tarjoamiin toimintamahdollisuuksiin.

Marjatta Okko