Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Computer, Communication and Information Sciences

Täsmäytysjärjestelmien kehittäminen linkitetyn datan sanastoille kulttuuriperintöorganisaatiossa

Unni Kohonen

Diplomityö
2024

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijä Unni Kohonen

Työn nimi Täsmäytysjärjestelmien kehittäminen linkitetyn datan sanastoille kulttuuriperintöorganisaatiossa

Koulutusohjelma Master's Programme in Computer, Communication and Information Sciences

Pääaine Computer Science

Vastuuopettaja/valvoja Professori Eero Hyvönen

Työn ohjaaja(t) TkT Osma Suominen

Yhteistyötaho Kansalliskirjasto

Päivämäärä 24.4.2024 **Sivumäärä** 66+6 **Kieli** Suomi

Tiivistelmä

Linkitetyn datan asiasanastot ovat laajalti käytössä esimerkiksi kirjastoissa, museoissa ja kulttuuriperintöorganisaatioissa, tarjoten tiedolle johdonmukaisen esitystavan ja helpottaen tiedon löytämistä ja tulkintaa eri järjestelmissä. Asiasanastojen kehittämisessä ja käyttämisessä tiedon integroiminen eri lähteistä on keskeisessä osassa. Sanastojen yhdistämistä ja niiden tunnisteiden hyödyntämistä muissa konteksteissa kutsutaan täsmäytykseksi. Täsmäyksen automatisaatiolla voidaan nopeuttaa ja helpottaa näitä työnkulkuja.

Yksi kulttuuriperintöorganisaatio Kansalliskirjaston tehtävistä on kehittää ja ylläpitää linkitetyn datan asiasanastoja ja työkaluja niiden käyttämiseen. Tämän diplomityön tavoitteena on kehittää ja arvioida täsmäytysjärjestelmä, joka vastaa Kansalliskirjaston tarpeisiin täsmäytyksen tekemisessä. Tutkimuksessa hyödynnetään suunnittelutieteen menetelmiä. Se koostuu kahdesta kehitysiteraatiosta, joissa arvioidaan rakennettujen järjestelmien toimintaa ja jatkokehitetään niitä saadun palautteen perusteella. Järjestelmiä arvioidaan haastatteluiden sekä laadullisen dataanalyysin keinoin.

Työssä toteutetaan kaksi täsmäytysjärjestelmäprototyyppiä Kansalliskirjaston Skosmos- ja Annif-ohjelmistojen avulla. Ne molemmat sisältävät toimintoja, jotka mahdollistavat täsmäytyksen tekemisen. Kehittämällä kaksi prototyyppiä saadaan laajemmin tietoa täsmäytysjärjestelmien rakentamisesta kulttuuriperintöorganisaation tarpeisiin. Työssä kehitetyt järjestelmäprototyypit ovat REST-rajapintoja, jotka toteuttavat World Wide Web Consortiumin (W3C) täsmäytysrajapintamääritelmän tärkeimmät toiminnallisuudet.

Tutkimus demonstroi, että on mahdollista kehittää täsmäytysjärjestelmä, joka vastaa kulttuuriperintöorganisaation tarpeisiin täsmäytyksen automatisoimiseksi. Kehitettyjen järjestelmien lopullisessa arvioinnissa ilmeni, että Skosmokseen perustuvan prototyyppi olisi valmis käyttöönottoon Kansalliskirjastolla, sillä sen tuottamat tulokset ovat luotettavia ja se sisältää tarpeelliset toiminnallisuudet. Annifiin perustuva prototyyppi puolestaan vaatisi vielä kehittämistä, vaikka sen tuottamat tulokset olivat monipuolisempia.

Avainsanat linkitetty data, kontrolloidut sanastot, täsmäytys, tietuelinkitys, suunnittelutiede

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Unni Kohonen

Title of thesis Developing reconciliation systems for linked data vocabularies in a cultural heritage organization

Programme Master's Programme in Computer, Communication and Information Sciences

Major Computer Science

Thesis supervisor Prof. Eero Hyvönen

Thesis advisor(s) Osma Suominen, D.Sc. (Tech.)

Collaborative partner National Library of Finland

Date 24.4.2024 **Number of pages** 66+6 **Language** Finnish

Abstract

Linked data vocabularies are widely used in institutions such as libraries, museums, and cultural heritage organizations, providing a consistent representation for knowledge and facilitating information discovery and interpretation across various systems. The integration of information from different sources is crucial in the development and use of linked data vocabularies. The process of combining vocabularies and utilizing their identifiers in other contexts is referred to as reconciliation. The automation of reconciliation can speed up and streamline these workflows.

One of the responsibilities of the National Library of Finland is to develop and maintain linked data vocabularies and tools for their use. The objective of this thesis is to develop and evaluate a reconciliation system that meets the reconciliation needs of the National Library of Finland. Research in this thesis utilizes the methods of design science and consists of two development iterations, during which the reconciliation systems are evaluated and further developed based on feedback that was received. The systems are evaluated through interviews and qualitative data analysis.

Two reconciliation system prototypes are developed using the Skosmos and Annif software maintained by the National Library. Both prototypes include the functionality required for reconciliation to be performed. The development of two prototypes provides a broader understanding of the construction of reconciliation systems for the needs of cultural heritage organizations. The prototypes developed in this thesis are REST APIs that implement the primary functionalities of the World Wide Web Consortium's (W3C) reconciliation service API.

This thesis demonstrates that it is feasible to develop a reconciliation system that fulfills the requirements of a cultural heritage organization for automating the reconciliation process. The final evaluation of the systems revealed that the prototype based on Skosmos would be ready for deployment at the National Library, as the results it produces are reliable and it includes the necessary functionalities for reconciliation. The prototype based on Annif, however, requires further development, although it produced more varied results.

Keywords linked data, controlled vocabularies, reconciliation, record linkage, design science

Sisällys

Esipuhe

Tämä työ on tehty toimeksiantona Kansalliskirjastolle yhteistyössä sen Finto- ja Annif-työryhmien kanssa.

Haluan kiittää työni valvojaa professori Eero Hyvöstä sekä ohjaajaani Osma Suomista, joka on ollut suurena apuna koko diplomityöprosessin ajan. Lisäksi haluan kiittää kaikkia muita kollegoitani Kansalliskirjastolla heidän avustaan työn eri vaiheissa.

Kiitokset myös perheelleni heidän jatkuvasta tuestaan.

Helsingissä 24.4.2024
Unni Kohonen

Lyhenteet

CSV	Comma Separated Values
DSRP	Design science research process
HTML	Hyper Text Markup Language
HTTP	Hypertext Transfer Protocol
JSON-LD	JavaScript Object Notation for Linked Data
RDF	Resource Description Framework
RDFS	RDF Schema
REST	Representational state transfer
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
TF-IDF	Term frequency–inverse document frequency
Turtle	Terse RDF Triple Language
URI	Unique Resource Identifier
URL	Unique Resource Locator
W3C	The World Wide Web Consortium
XML	Extensible Markup Language

1 Johdanto

Linkitetyn datan kontrolloidut asiasanastot ovat laajassa käytössä sisällönkuvailussa muun muassa kirjasto- ja museoalalla sekä erilaisissa kulttuuriperintöorganisaatioissa. Nämä sanastot tarjoavat tiedolle tarkan rakenteen ja yhdenmukaisen esitystavan, samalla hyödyntäen linkitetyn datan periaatteita käsitteiden uudelleenmäärittelyn ja päällekkäisyyden välttämiseksi. Näin ne helpottavat tiedon löytämistä ja tulkintaa eri järjestelmissä, edistäen siten tiedon yhteentoimivuutta ja tehokasta tiedonhallintaa. (Harper & Tillett, 2007; Smith, 2021.)

Linkitetyn datan asiasanastojen käytössä ja kehittämisessä tiedon integroiminen useista lähteistä on merkittävässä roolissa. Tämän vuoksi on olennaista, että sanastojen kehittäjät ja käyttäjät pystyvät yhdistämään tietoaineistoja, jotka eivät jaa yhteisiä linkitetyn datan tunnisteita. Tästä tunnisteiden yhdistämisestä on englanninkielisessä kirjallisuudessa käytetty monia nimiä, esimerkiksi ”reconciliation” (Delpeuch, 2019), ”entity resolution” (Christophides, 2012) ja ”data matching” (Christen, 2012), mutta tässä opinnäytetyössä siitä käytetään nimitystä täsmäytys. Täsmäytyksellä on monia käyttökohteita. Sitä voidaan esimerkiksi hyödyntää tilanteessa, jossa halutaan löytää kahdesta eri linkitetyn datan sanastosta toisiaan vastaavat käsitteet ja luoda niiden välille vastaavuussuhteita. Toinen esimerkki on erilaisten museokokoelmien linkittämättömän metatiedon rikastaminen linkitetyn datan asiasanoilla.

Toisiaan vastaavien tunnisteiden tai asiasanojen löytäminen vaikeutuu, kun niiden määrä kasvaa. Sanastojen kehittäjien ja käyttäjien työstä tulee hidasta ja vaivalloista, jos täsmäytys joudutaan tekemään manuaalisesti jokaiselle asiasanalle erikseen. Tämä luo tarpeen järjestelmälle, joka mahdollistaa täsmäytyksen osittaisen automatisaation ja helpottaa sen suorittamista suuremmille asiasanamäärille. Tällainen järjestelmä säästää aikaa ja resursseja sekä parantaa tiedon integroitavuutta. Yksi tapa toteuttaa täsmäytysjärjestelmä on käyttää World Wide Web Consortiumin (W3C) määritelmän mukaista täsmäytysrajapintaa (Delpeuch ym., 2023), joka ehdottaa käyttäjälle sanaston tietueita linkitettäväksi. Tällaista rajapintaa käytetään usein taulukkomuotoisen datan käsittelyyn kehitetyllä OpenRefine-työkalulla (OpenRefine, 2023).

Kulttuuriperintöorganisaatio Kansalliskirjaston yhtenä tavoitteena on tarjota ja kehittää kontrolloituja asiasanastoja sekä edistää sisällönkuvailutyötä. Se tarjoaa useita työkaluja näiden tehtävien helpottamiseksi, esimerkiksi linkitetyn datan sanastoselaimen Finton (Suominen ym. 2014) sekä automaattisen sisällönkuvailun palvelun Finto AI:n (Suominen ym. 2022). Tietoaineistojen yhdistämisellä voidaan tukea sanastokehityksen ja kuvailun prosesseja entisestään. Täsmäytystä on toteutettu Kansalliskirjastossa ulkoisia palveluita ja OpenRefine-työkalua käyttäen, mutta sen omat järjestelmät eivät vielä mahdollista täsmäytystä. Onkin syntynyt tarve laajentaa näitä

täsmäytyksen työnkulkuja myös Kansalliskirjaston tarjoamiin sanastoihin, erityisesti esimerkiksi yleiseen suomalaiseen ontologiaan YSO:on.

Finto-palvelu perustuu Kansalliskirjaston kehittämään ja ylläpitämään Skosmos-ohjelmistoon (Suominen ym. 2015), joka tarjoaa alustan SKOS-muotoisten linkitetyn datan asiasanastojen, ontologioiden ja luokitusten julkaisemiseen ja selaamiseen. Sen avulla voidaan päästä käsiksi sanastoihin selainkäyttöliittymän ja REST-rajapinnan kautta. Sekä käyttöliittymä että REST-rajapinta mahdollistavat asiasanojen hakemisen sanastoista, mutta näin voidaan tehdä vain yksi asiasana kerrallaan ja tunnisteen valinta tulee tehdä manuaalisesti. Finto AI -palvelu perustuu niin ikään Kansalliskirjastossa kehitettyyn ja ylläpidettyyn Annif-ohjelmistoon (Suominen ym. 2022), jonka avulla voidaan tekoälyn keinoin automaattisesti hakea ehdotuksia sopiviksi asiasanoiksi tekstiaineiston perusteella. Myös se tarjoaa selainkäyttöliittymän ja REST-rajapinnan. Annif kykenee käsittelemään useita tekstiaineistoja samanaikaisesti, mutta ei sellaisenaan tue täsmäytyksen apuna käytössä olevia palveluita. Skosmos tai Annif eivät siis kumpikaan tällä hetkellä mahdollista täsmäytyksen automatisaatiota, mutta niiden olemassa olevien toimintojen avulla voidaan toteuttaa täsmäytysjärjestelmät, jotka helpottavat täsmäytyksen tekemistä.

Tämän opinnäytetyön tavoitteena on toteuttaa ja arvioida järjestelmä, joka mahdollistaa täsmäytyksen osittaisen automatisaation erityisesti Kansalliskirjaston sanastokehityksen näkökulmasta. Työ vastaa seuraaviin tutkimuskysymyksiin:

1) Minkälaisia tarpeita kulttuuriperintöorganisaatioissa on täsmäytykselle ja mitkä vaatimukset nämä tarpeet asettavat toteutettavalle järjestelmälle?
2) Miten järjestelmä toteutetaan vastaamaan asetettuja vaatimuksia?
3) Miten järjestelmää arvioidaan ja miten sen toiminta vastaa arviointiperusteita?

Opinnäytetyö ei painotu täsmäytysalgoritmien analyysiin tai kehitykseen, vaan tarkoituksena on keskittyä olemassa olevien ratkaisujen soveltamiseen ja laajentamiseen.

Tämä diplomityö jakautuu teoriaosuuteen ja käytännön toteutuksen kuvaukseen. Teoriaosuus alkaa luvusta 2, jossa esitellään linkitetyn datan periaatteet ja teknologiat sekä linkitetyn datan kontrolloidut sanastot. Luvussa 3 esitellään tarkemmin Skosmos- ja Annif-ohjelmistot. Teoriaosuuden lopussa luvussa 4 käsitellään täsmäytyksen mahdollistavaa tietuelinkitystä, W3C:n täsmäytysrajapintaa ja OpenRefine-työkalua sekä aiemmin toteutettuja täsmäytysjärjestelmiä. Käytännön toteutuksen osuus alkaa luvusta 5, jossa esitellään diplomityössä käytetyt tutkimusmenetelmät. Luvussa 6 puolestaan käydään läpi toteutetut täsmäytysjärjestelmät sekä arvioidaan niiden suoriutumista. Lopuksi luvussa 7 tehdään yhteenveto diplomityöstä, pohditaan sen tuloksia sekä vastataan tutkimuskysymyksiin.

2 Linkitetty data ja linkitetyn datan sanastot

Semanttinen web tarjoaa vision verkosta, jossa tieto on koneellisesti käsiteltävässä muodossa, mikä mahdollistaa älykkäiden verkkopalveluiden ja -sovellusten kehittämisen (Berners-Lee ym., 2001; Hyvönen, 2002). Semanttinen web on rakennettu W3C:n kehittämien teknologioiden ja ylläpitämien standardien varaan (Semantic Web Standards, 2019). Suomessa semanttisen webiä ja sen palveluita on alettu kehittämään vuonna 2002 (Hyvönen, 2021). Vuonna 2003 käynnistyi Suomalaiset semanttisen webin ontologiat (FinnONTO) -hanke, jonka alla kehitettiin muun muassa ONKI-ontologiakirjastopalvelu (Hyvönen ym., 2008). Yksi tapa toteuttaa semanttisen webin periaatteita on linkitetty data. Tässä luvussa käsitellään tarkemmin linkitetyn datan periaatteita, sen teknologioita sekä kontrolloituja sanastoja.

2.1 Linkitetty data

Linkitetyllä datalla (engl. linked data) tarkoitetaan yhteyksien luomista verkon eri lähteistä peräisin olevien aineistojen välille. Sen tavoitteena on luoda globaali tietoverkko (engl. web of data), joka koostuu kaikista julkaistuista tiedoista. Toisin kuin perinteisessä hypertekstiverkossa, jossa dokumentit linkittyvät tyypittämättömillä hyperlinkeillä, linkitetyn datan verkossa luodaan tyypitettyjä linkkejä käyttämällä koneluettavaa standardisoitua dataa. (Bizer ym., 2009.)

Berners-Lee (2006) määrittelee 4 ohjesääntöä linkitetylle datalle:

1) Käytä URI-tunnisteita aineistojen nimeämiseen
2) Käytä HTTP URI-tunnisteita, jotta nimiä voidaan etsiä
3) Kun URI-tunnisteen sisältö haetaan, tarjoa hyödyllistä tietoa standardoidussa muodossa
4) Sisällytä linkkejä muihin URI-tunnisteihin, jotta on mahdollista löytää lisää tietoa

Berners-Lee (2006) ehdottaa myös 5 tähden mallin avoimen linkitetyn datan (engl. linked open data) julkaisemiseen:

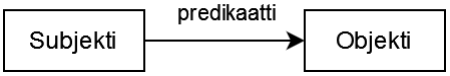
- 1 tähden malli: Data on saatavilla avoimen lisenssin alla missä tahansa muodossa
- 2 tähden malli: Data on saatavilla koneluettavassa muodossa
- 3 tähden malli: Data on saatavilla koneluettavassa muodossa, avoimessa formaatissa
- 4 tähden malli: Data on saatavilla W3C:n standardien mukaisessa muodossa
- 5 tähden malli: Data on linkitetty muuhun verkossa saatavilla olevaan dataan

Nämä Berners-Leen (2006) esittämät periaatteet muodostavat yhteisen kehyksen (avoimen) linkitetyn datan julkaisemiseen ja linkittämiseen verkossa.

Seuraavassa alaluvussa esitellään linkitetyn datan keskeiset teknologiat, jotka ovat olennaisia tässä diplomityössä.

2.2 Linkitetyn datan teknologiat

Kuten edellä esitettiin, linkitetyn datan resurssit tunnistetaan käyttämällä yksikäsitteisiä URI-tunnisteita (engl. Unique Resource Identifier), joista on saatavilla tietoa HTTP-protokollan (engl. Hypertext Transfer Protocol) avulla verkossa. Resurssit ja niiden suhteet muihin resursseihin esitetään RDF-mallin (engl. Resource Description Framework) (Cyganiak ym., 2014) mukaisessa, koneluettavassa muodossa. RDF-data rakentuu kolmikoista (engl. triples), jotka koostuvat subjektista, predikaatista ja objektista. Kolmikoista koostuvaa joukkoa kutsutaan RDF-graafiksi, jossa subjektit ja objektit ovat solmuja ja predikaatit solmujen välisiä kaaria. Kuvassa 2.1 on esitetty esimerkki yksinkertaisesta RDF-graafista. Subjektit ja objektit voivat olla URI-tunnisteisia resursseja, literaaleja tai tyhjiä solmuja. Predikaatit ovat aina URI-tunnisteisia resursseja. RDF-graafeja voidaan kuvata käyttäen montaa eri syntaksia, esimerkiksi Turtle, RDF/XML tai JSON-LD. (Cyganiak ym., 2014.)



Kuva 2.1. Yksinkertainen RDF-graafi.

RDF-mallin laajennus RDF Schema (RDFS) (Brickley & Guha, 2014) tarjoaa sanaston RDF-datan semanttiselle kuvaamiselle. Sen avulla voidaan kuvata resursseista koostuvia ryhmiä eli luokkia (rdfs:Class), ja niiden suhteita toisiinsa eli ominaisuuksia (rdf:Property). RDFS:n ominaisuudet ovat RDF-kolmikoiden predikaatteja. RDFS määrittelee useita ominaisuuksia, kuten rdf:type, jonka avulla RDF-resurssit määritellään jonkin luokan instansseiksi. Luokille voidaan määrittää alaluokkia rdfs:subClassOf-ominaisuudella, jolloin kaikki alaluokan instanssit ovat myös yläluokan instansseja. rdfs:subPropertyOf-ominaisuudella voidaan ilmaista, että resurssit, joilla on jokin ominaisuus, on myös jokin toinen ylätason ominaisuus. Ominaisuuksien arvo- ja määrittelyjoukot voidaan ilmaista rdfs:range- ja rdfs:domain-ominaisuuksilla. Lisäksi RDFS sisältää rdfs:comment- ja rdfs:label-ominaisuudet ihmisen luettavissa olevan tiedon määrittelyyn. (Brickley & Guha, 2014.)

RDF(S)-muotoista dataa voidaan hakea SPARQL-kyselykielellä SPARQL-rajapinnasta (Garlik & Seaborne, 2013). Suurin osa mahdollisista kyselyistä sisältää kyselymallin (engl. query pattern) eli joukon RDF-kolmikoita muistuttavia kolmikoita, joissa subjektit, objektit ja predikaatit voidaan korvata muuttujilla. Kyselymalli vastaa RDF-datan jotain aligraafia, joka palautetaan vastauksena. SPARQL-kyselyt voivat olla muodoltaan SELECT-, DESCRIBE-

, ASK- tai CONSTRUCT-mallisia. SELECT-kysely palauttaa kyselymallissa olevien muuttujien arvot sitä vastaavasta RDF-datan aligraafista. DESCRIBE-kysely puolestaan palauttaa koko aligraafin, joka vastaa kyselymallia. ASK-kysely palauttaa totuusarvon, joka ilmaisee vastaako kyselymalli jotain aligraafia. CONSTRUCT-kyselyn avulla voidaan rakentaa kyselymallin perusteella uusi graafi. SPARQL:n avulla on mahdollista osoittaa kyselyt useaan RDF-datalähteeseen yhdistettyjen kyselyiden avulla (engl. federated query). (Garlik & Seaborne, 2013.)

RDF- ja RDFS-standardien päälle on rakennettu useita muita linkitetyn datan standardeja. Niistä yksi on kontrolloitujen strukturoitujen sanastojen kuvaukseen tarkoitettu SKOS-tietomalli (engl. Simple Knowledge Organization System) (Miles & Bechhofer, 2009). Kuten RDFS, se perustuu luokkiin ja ominaisuuksiin. SKOS-pohjaiset sanastot koostuvat käsitteistä (skos:Concept), jotka mallintavat joitain abstrakteja tai konkreettisia asioita. SKOS-käsitteistä voidaan muodostaa käsiteskeemoja (skos:ConceptScheme), joilla on mahdollista mallintaa sanastoja. Samankaltaisista käsitteistä voidaan myös muodostaa kokoelmia (skos:Collection). SKOS tarjoaa useita ominaisuuksia, esimerkiksi skos:inScheme ja skos:topConceptOf, joiden avulla käsitteitä voidaan liittää käsiteskeemoihin. Käsiteskeeman sisällä käsitteiden välille voidaan määrittää hierarkkisia ja semanttisia suhteita esimerkiksi skos:broader-, skos:narrower- ja skos:related-ominaisuuksien avulla. Eri käsiteskeemojen käsitteiden välille puolestaan voidaan määrittää vastaavuussuhteita esimerkiksi skos:exactMatch- ja skos:closeMatch-ominaisuuksilla. Käsitteille voidaan määritellä termejä skos:prefLabel-, skos:altLabel- ja skos:hiddenLabel-ominasuuksilla. Skos:definition-ominaisuudella niille voidaan lisäksi antaa määritelmä. (Miles & Bechhofer, 2009.)

2.3 Linkitetyn datan kontrolloidut sanastot

Smith (2021) määrittelee kontrolloidut sanastot luetteloksi termejä, joita käytetään viestinnän standardoimiseen ja ymmärtämisen helpottamiseen. Ne koostuvat tarkasti valikoiduista termeistä, jotka muodostavat tietyn aihealueen kuvailemiseen käytetyn sanavaraston (Hedden, 2008). Smithin (2021) mukaan ne ovat tärkeitä sekä tieteellisen keskustelun helpottamisessa ja standardoinnissa että tiedon kuvailun ja järjestämisen parantamisessa. Kontrolloitujen sanastojen avulla voidaan tehostaa esimerkiksi museo- ja kirjastoaineistojen luettelointia ja hakua.

Harpring (2010, luku 2.3) esittää useita tyyppejä kulttuurialan kontrolloiduille sanastoille, esimerkiksi asiasanalistat, kontrolloidut listat, auktoriteettitiedostot, tesaurukset ja ontologiat. Nämä eroavat toisistaan lähinnä laajuutensa ja rakenteensa kompleksisuuden suhteen. Harpring erottelee lisäksi toisistaan strukturoidut ja strukturoimattomat kontrolloidut sanastot. Strukturoidut sanastot hänen mukaansa korostavat käsitteiden välisiä suhteita, esimerkiksi termien vastaavuutta ja niiden välisiä hierarkiasuhteita.

Harper ja Tillett (2007) ovat esittäneet, että kontrolloidut sanastot voivat toimia osana linkitetyn datan tietoverkon rakentamista. Heidän mukaansa olemassa olevat, esimerkiksi kirjastoalan, sanastot voidaan muuntaa linkitetyn datan teknologioita hyödyntävään muotoon, jolloin ne parantavat tiedon hakemista ja lisäävät tietolähteiden saavutettavuutta ja yhteentoimivuutta. Janowicz ym. (2014) puolestaan esittävät, että Berners-Leen (2006) linkitetyn datan viiden tähden malli on ollut vain perusedellytys toimivan linkitetyn datan tuottamiseen ja he painottavat sanastojen merkitystä linkitetyn datan käytettävyyden parantamisessa. He ovatkin esittäneen Berners-Leen viiden tähden mallia vastaavan mallin linkitetylle datalle ja sen sanastoille:

- 0 tähden malli: linkitetty data ilman sanastoa
- 1 tähden malli: linkitetty data, joka sisältää ihmisen luettavan kuvauksen käytetystä sanastosta
- 2 tähden malli: tieto sanastosta on tarjolla koneluettavassa muodossa
- 3 tähden malli: sanasto linkittyy muihin sanastoihin
- 4 tähden malli: sanastosta on saatavilla koneluettavaa metatietoa
- 5 tähden malli: muut sanastot linkittyvät sanastoon

Esimerkki sanastosta, joka on saatavilla SKOS-muodossa linkitettynä datana, on Yhdysvaltain Kongressin kirjaston ylläpitämä Library of Congress Subject Headings (LCSH) -asiasanasto[1]. Muita linkitetyn datan sanastoja on esimerkiksi Kansalliskirjaston ylläpitämä SKOS-muotoinen yleinen suomalainen ontologia[2] (YSO), joka pohjautuu yleiseen suomalaiseen asiasanastoon[3] (YSA) (Hyvönen ym., 2008) sekä Yhdysvaltain kansallisen lääketieteellisen kirjaston Medical Subject Headings (MeSH) -sanasto[4].

[1] https://www.loc.gov/cds/products/product.php?productID=209
[2] https://finto.fi/yso/
[3] https://finto.fi/ysa/
[4] https://www.ncbi.nlm.nih.gov/mesh/

3 Kansalliskirjaston sanastotyön ohjelmistot

Tässä luvussa esitellään lyhyesti Skosmos ja Annif, kaksi Kansalliskirjaston ylläpitämää sanastotyössä käytettyä ohjelmistoa. Luvussa käsitellään niiden toiminnallisuutta, arkkitehtuuria ja käyttötapauksia. Ensimmäisessä alaluvussa keskitytään Skosmokseen ja toisessa Annifiin.

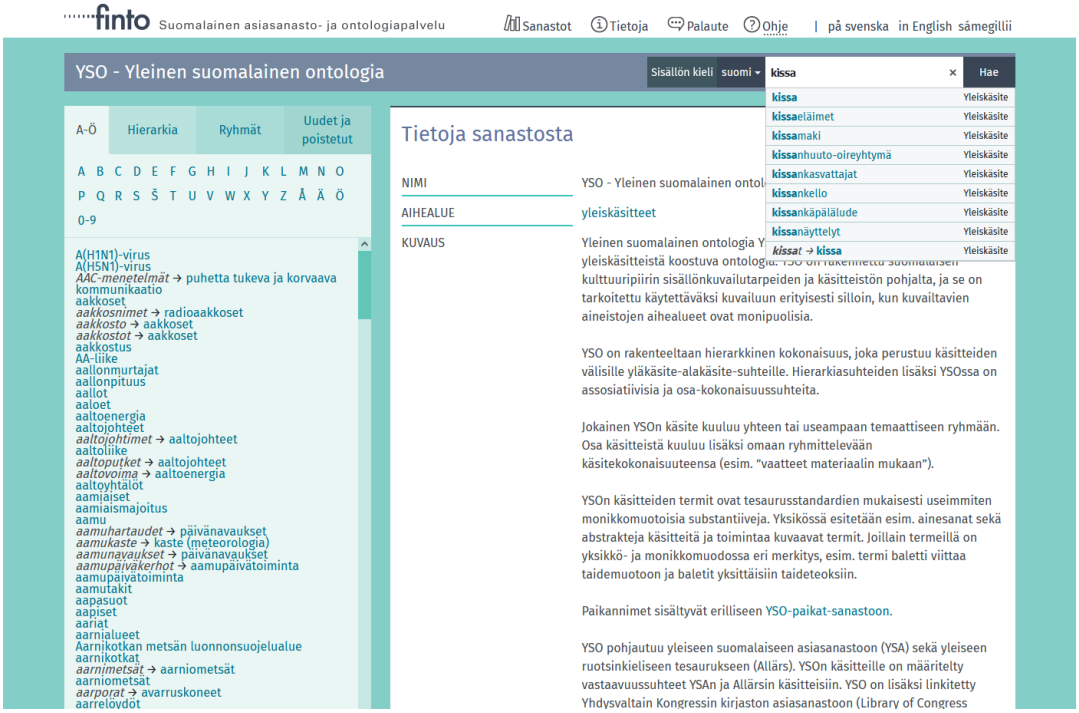
3.1 Skosmos

Skosmos (Suominen ym., 2015) on Kansalliskirjaston kehittämä ja ylläpitämä web-pohjainen ohjelmisto SKOS-muotoisten linkitetyn datan sanastojen julkaisuun ja selaamiseen. Se perustuu FinnONTO-hankkeessa kehitettyyn ONKI-ontologiapalveluun (Suominen ym., 2014).

Skosmoksen avulla voidaan selata sanastojen käsitteitä, niiden muodostamia hierarkioita, ryhmiä sekä sanastoihin tehtyjä muutoksia. Se tukee monikielisiä SKOS-sanastoja ja myös mahdollistaa niiden selaamisen eri kielillä. Skosmoksen tarjoamat sanastot on tallennettu RDF-kolmikkotietokantaan (esimerkiksi Apache Jena Fuseki) ja sen toiminnot perustuvat tietokannan tarjoamaan SPARQL-rajapintaan. Se tarjoaa selainkäyttöliittymän lisäksi REST-rajapinnan sanastojen käsittelyyn. Rajapinta mahdollistaa esimerkiksi käsitteen tai kokonaisen sanaston RDF-datan noutamisen sekä käsitteiden etsimisen sanaston sisällä tai yhtäaikaisesti kaikista sanastoista. (Suominen ym., 2015.)

Kansalliskirjasto tarjoaa oman Skosmos-instanssinsa, Finton[5], joka sisältää YSO:n ja muiden yleisten sanastojen lisäksi useita erityisalojen ontologioita, sanastoja ja luokituksia. Kuvassa 3.1 on esitetty Finton selainkäyttöliittymässä YSO:n etusivu sekä käsitteiden hakutoiminnallisuus. Myös muita Skosmos-pohjaisia palveluita on tarjolla verkossa, esimerkiksi UNESCO:n ylläpitämä UNESCO-tesaurus[6] ja YK:n elintarvike- ja maatalousjärjestön AGROVOC-tesaurus[7].

[5] https://finto.fi/
[6] https://vocabularies.unesco.org/
[7] https://agrovoc.fao.org/



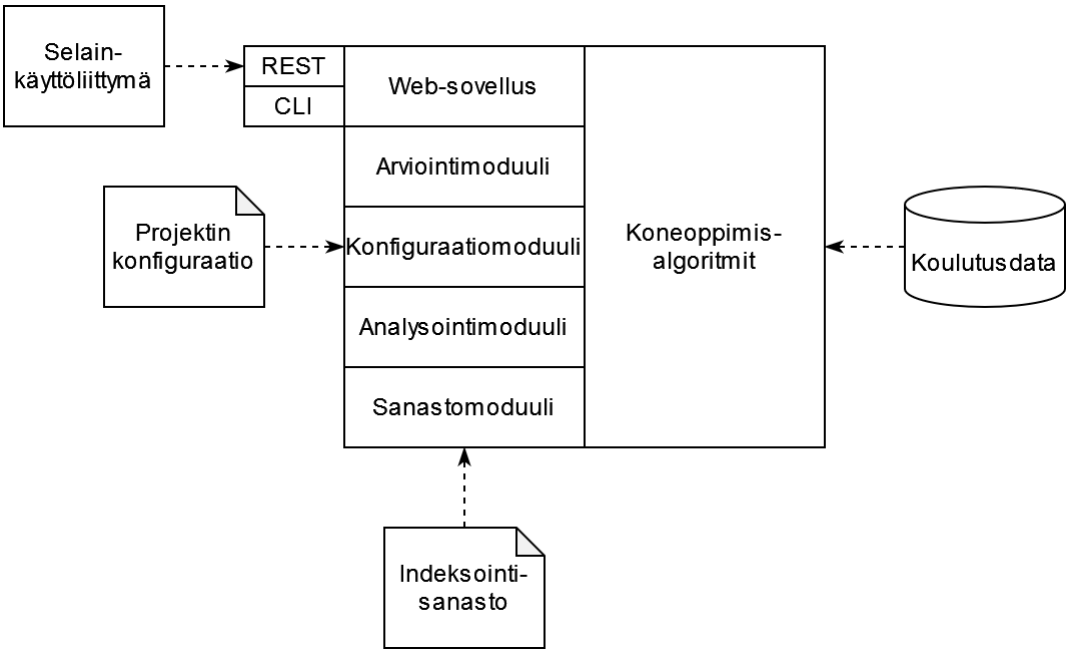
Kuva 3.1. YSO-ontologian etusivu Finto-palvelussa. Ylhäällä oikealla näkyy Finton hakutoiminnallisuus.

3.2 Annif

Annif (Suominen ym., 2022) on Kansalliskirjastossa kehitetty ja ylläpidetty ohjelmisto tekstiaineistojen automaattiseen kuvailuun asiasanoilla. Se käyttää koneoppimisalgoritmeja analysoimaan syötettyjä tekstiaineistoja ja ehdottaa asiasanoja niiden indeksoimiseksi.

Annifin ehdottamat asiasanat haetaan indeksointisanastosta, joka voidaan määritellä joko yksinkertaisena tekstitiedostona tai SKOS-muodossa. Koulutusaineistona toimivat jo olemassa olevat indeksoidut tekstiaineistot ja metadatatietueet. Annif yhdistää sanaston ja käytetyt algoritmit sekä arviointi- ja analysointimoduulin yhdeksi projektiksi, jonka kautta kuvailuehdotuksia tehdään. Se tarjoaa komentorivikäyttöliittymän sekä REST-rajapinnan toiminnallisuuden integroimiseen muihin järjestelmiin. Myös yksinkertainen selainkäyttöliittymä tarjotaan pääasiassa testausta varten. REST-rajapinta sisältää metodeja esimerkiksi projektin tietojen tarkasteluun sekä kuvailuehdotusten tekemiseen. (Suominen ym., 2022.)

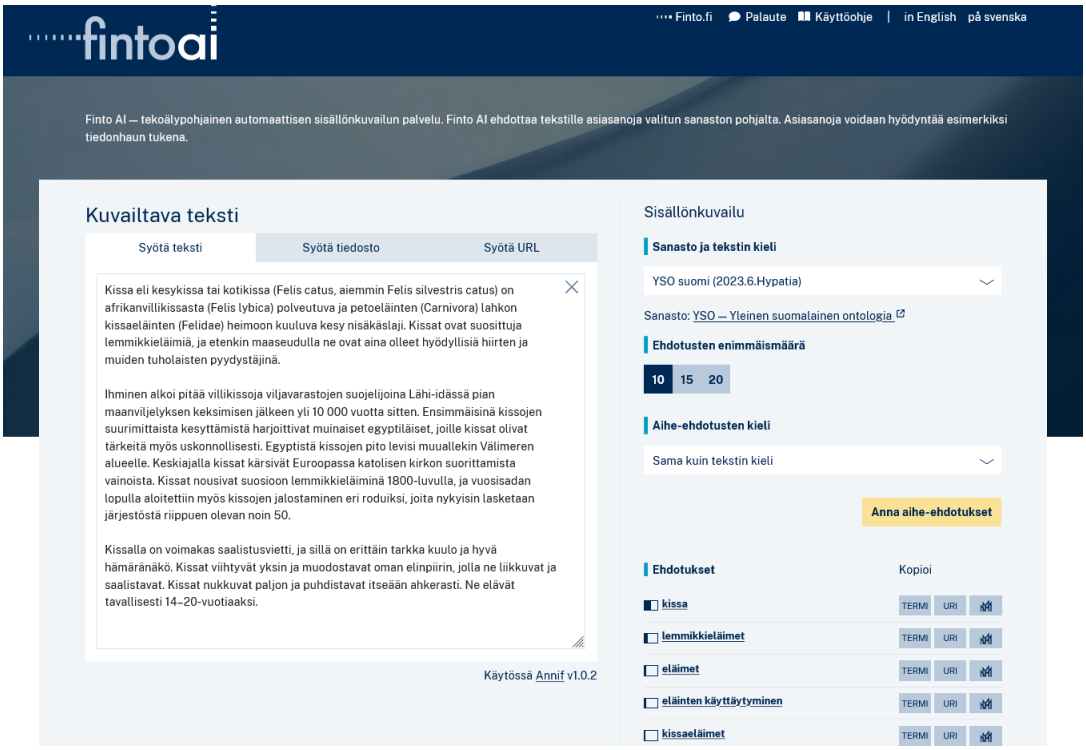
Kuvassa 3.2 on havainnollistettu Annifin perusarkkitehtuuri. Siinä on esitetty projektien arviointi-, konfiguraatio-, analysointi- ja sanastomoduulit sekä niiden käyttämät koneoppimisalgoritmit. Myös kaikkien projektien käsittelyyn tarkoitettu web-sovellus on esitetty kuvassa.



Kuva 3.2. Annif-ohjelmiston perusarkkitehtuuri. Mukailtu Suomisen ym. (2022) kuvasta 1.

Annif on käytössä Kansalliskirjaston Finto AI -palvelussa[8], joka tarjoaa Annifin REST-rajapintaa hyödyntävän selainkäyttöliittymän tekstiaineistojen indeksointiin. Sen avulla voidaan käyttää YSO-, KAUNO[9]- tai YKL-sanastoja[10] tekstin kuvailuun. Finto AI:n lisäksi Annif on käytössä myös esimerkiksi Saksan ja Ruotsin kansalliskirjastoissa. Kuvassa 3.3 on esitetty esimerkki Finto AI:n käytöstä tekstin kuvailussa. Kuvassa näkyy kuvailtava teksti vasemmalla ja kuvailun asetukset oikealla sekä asiasanaehdotukset asetusten alla.

[8] https://ai.finto.fi/
[9] https://finto.fi/kauno/
[10] https://finto.fi/ykl/



Kuva 3.3. Esimerkki Finto AI -palvelun käytöstä tekstiaineiston kuvailussa. Teksti haettu Wikipediasta[11].

[11] https://fi.wikipedia.org/wiki/Kissa

4 Täsmäytys

Täsmäytyspalvelun toteuttamisen kannalta on välttämätöntä ymmärtää, miten toisiaan vastaavat tietueet löydetään eri tietolähteistä sekä mitä teknologioita siihen voidaan käyttää. Alaluvussa 4.1. esitellään tietuelinkityksen teoreettinen tausta ja sen eri vaiheet sekä joitakin keskeisimpiä algoritmeja. Alaluvussa 4.2. kuvataan W3C:n täsmäytysrajapintamääritelmä sekä sitä hyödyntävä OpenRefine-työkalu, joiden avulla tietuelinkityksen periaatteita voidaan soveltaa käytännössä täsmäytyspalveluiden rakentamisessa ja käyttämisessä. Lopuksi alaluvussa 4.3. tarkastellaan kirjallisuudessa aiemmin esitettyjä täsmäytyspalveluiden toteutuksia.

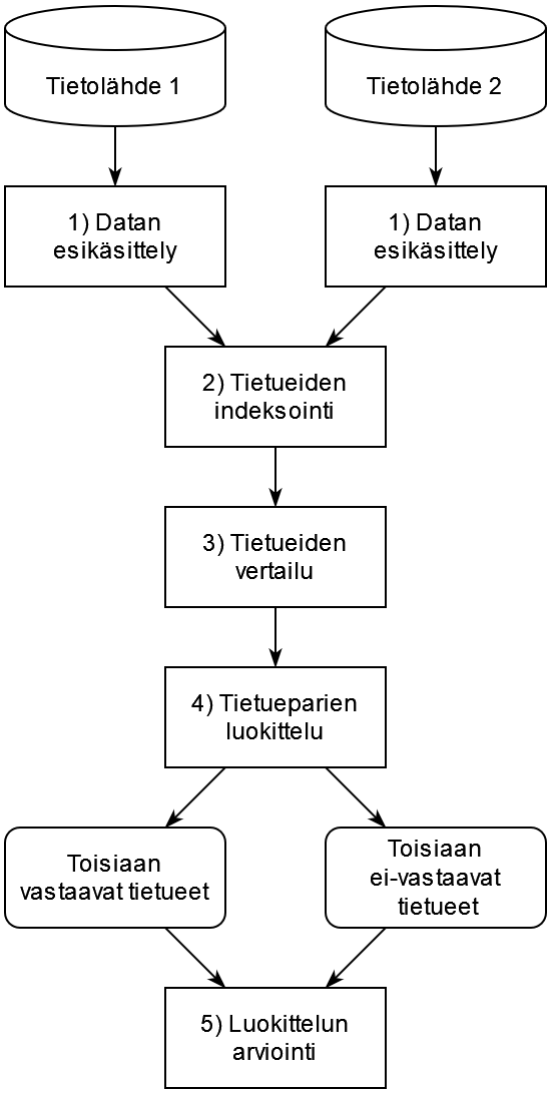
4.1 Tietuelinkitys

Tietuelinkityksellä (englanniksi usein esimerkiksi record linkage tai entity resolution) tarkoitetaan prosessia, jossa tunnistetaan ja yhdistetään useiden eri tietolähteiden tietueita, jotka viittaavat samoihin reaalimaailman objekteihin. Tietolähteinä voivat toimia esimerkiksi relaatiotietokannat tai linkitetyn datan sanastot. Tietueiden linkityksessä käytetään niiden yhteneviä ominaisuuksia, esimerkiksi relaatiotietokannan sarakkeiden arvoja tai linkitetyn datan tietueiden graafirakenteita. (Christophides ym., 2015.)

Christen (2012) jakaa tietuelinkityksen viiteen eri vaiheeseen:

1) Datan esikäsittelyyn
2) Tietueiden indeksointiin
3) Tietueiden vertailuun
4) Tietueparien luokitteluun
5) Luokittelun arviointiin

Tässä jaottelussa datan esikäsittelyllä tarkoitetaan prosessia, jossa tietokantojen sisältämä data muutetaan helpommin vertailtavaan muotoon. Tietuelinkityksen yhteydessä tämä voi tarkoittaa esimerkiksi kirjoitusvirheiden korjaamista tai turhien sanojen ja merkkien poistamista. Tietueiden indeksoinnin avulla yritetään löytää tietueet, joita kannattaa verrata toisiinsa ja siten vähentää linkitysprosessin laskennallista kompleksisuutta. Vertailuvaiheessa tietueita verrataan keskenään käyttämällä erilaisia algoritmeja ja sillä tavoin niistä muodostetaan pareja. Saadut parit luokitellaan seuraavassa vaiheessa joko toisiaan vastaaviksi tai ei-vastaaviksi. Viimeisessä vaiheessa on mahdollista vielä arvioida saatujen luokitusten laatua sekä koko linkitysprosessin kompleksisuutta. Kuva 4.1 havainnollistaa Christenin tietuelinkitysprosessin vaiheet.



Kuva 4.1. Tietuekytkentäprosessi. Mukailtu Christenin (2012) kuvasta 2.1.

Alaluvussa 4.1.1. esitetään täsmällisempi matemaattinen määritelmä tietuelinkitykselle. Alaluvussa 4.1.2. keskitytään tarkemmin kuvailemaan tietueiden indeksointia ja alaluvussa 4.1.3. tietueiden vertailua ja tietueparien luokittelua.

4.1.1 Tietuelinkityksen matemaattinen määritelmä

Tietuelinkityksessä tietueet, joita kutsutaan myös entiteeteiksi, voidaan ajatella joukkoina ominaisuus-arvopareja, jotka koostuvat ominaisuuksien nimistä ja niiden arvoista. Entiteetti e voidaan siis määritellä seuraavasti:

$$e = \{(a_i, v_j) \mid a_i \in A, v_j \in V\},$$

missä A on joukko ominaisuuksia ja V on joukko ominaisuuksien arvoja. Tietuelinkityksen ongelmana on siis löytää entiteettijoukoista E_i ja E_j ne entiteettiparit $(e_i, e_j) \in E_i \times E_j$, joiden jäsenet edustavat samaa objektia

reaalimaailmassa. Tämä tapahtuu löytämällä vastaavuusfunktio $M: E_i \times E_j \rightarrow \{tosi,\ epätosi\}$, joka jakaa entiteettiparit toisiaan vastaaviin tai ei-vastaaviin joukkoihin. (Christophides ym., 2015.)

4.1.2 Tietueiden indeksointi

Jos tietueiden vertailuvaiheessa jokainen mahdollinen tietuepari (eli entiteettijoukkojen E_1 ja E_2 karteesinen tulo) käsiteltäisiin erikseen, olisi tähän vaadittu aika $O(n^2)$. Tämä ei ole riittävän tehokasta suurille tietoaineistoille, joten vertailujen määrää on pystyttävä vähentämään. Ratkaisu tähän ongelmaan saadaan tietueiden indeksoinnista, joka voidaan jakaa kahteen vaiheeseen, lohkomiseen (engl. blocking) ja suodattamiseen (engl. filtering). Lohkomisessa potentiaalisesti toisiaan vastaavat tietueet jaetaan niiden ominaisuuksien (esimerkiksi relaatiotietokannan sarakkeiden arvojen) perusteella lohkoihin, joiden sisäisiksi vertaukset on rajoitettu. Suodattamisella puolestaan hylätään sellaiset tietueparit, jotka ovat toisiaan vastaamattomia. (Köpcke ym., 2010.)

Lohkomisessa käytetään jotain lohkomismenetelmää, joka tuottaa lohkojoukon B jakamalla entiteettijoukon E osiin. Osat voivat olla päällekkäisiä tai erillisiä lohkomismenetelmästä riippuen. Lohkomismenetelmä tuottaa ensin kaikista entiteeteistä tunnusmerkkejä (engl. signature), joiden samankaltaisuus kuvastaa entiteettien samankaltaisuutta, minkä jälkeen entiteetit jaetaan lohkoihin tunnusmerkkien avulla. Lohkomisen tavoitteena on maksimoida löydettyjen aitojen vastaavuuksien määrä samalla kun minimoidaan sellaisten vertailujen määrää, jotka tehdään toisiaan vastaamattomien entiteettien välillä. Lohkomisella saadaan merkittävästi vähennettyä turhia vertailuja ja siten nopeutettua tietuelinkitysprosessia. Tällöin löydettyjen vastaavuuksien määrä kuitenkin vähenee, koska osa vastaavista entiteeteistä asetetaan eri lohkoihin. (Papadakis ym., 2020.)

Gu ja Baxter (2004) kuvaavat suodattamista lohkomisprosessin jälkikäsittelyvaiheena, jossa suurimpien lohkojen entiteettejä vertaillaan likimääräisesti toisiinsa ei-vastaavien parien löytämiseksi. Tässä menetelmässä valitaan jokin suodatusmuuttuja, jonka avulla turhat entiteettiparit voidaan poistaa lohkoista. Papadakis ym. (2023) esittävät tämän lisäksi myös toisen suodatustekniikan, lohkojen puhdistamisen, jossa vähennetään lohkojen päällekkäisyyttä ja siten tehostetaan todellisten parien löytymistä. He ehdottavat suurimpien lohkojen poistamista kokonaan, sillä ne todennäköisesti koostuvat entiteettipareista, jotka esiintyvät myös muissa lohkoissa.

4.1.3 Tietueiden vertailu ja tietueparien luokittelu

Kuten alaluvussa 4.1.1. mainittiin, tietuelinkityksen tavoitteena on löytää ne entiteettiparit $(e_i, e_j) \in E_i \times E_j$, jotka kuvaavat samaa reaalimaaliman

objektia käyttämällä vastaavuusfunktiota $M(e_i, e_j)$. Käytännössä vastaavuusfunktio määritellään käyttämällä jotain samankaltaisuusfunktiota $sim: E_i \times E_j \rightarrow \mathbb{R}$ ja kynnysarvoa $\theta \in \mathbb{R}$:

$$\mathrm{M}(e_i, e_j) = \begin{cases} tosi, \text{jos } sim(e_i, e_j) \geq \theta \\ epätosi \text{ muutoin} \end{cases}.$$

Ideaalinen samankaltaisuusfunktio löytäisi aina toisiaan vastaavat entiteettiparit, mutta tällaista funktiota ei ole olemassa kaikelle datalle. Funktiot pyrkivätkin siis maksimoimaan löytyneiden todellisten parien määrää samalla minimoiden löytyneiden toisiaan vastaamattomien parien määrää. (Christophides ym., 2015.)

Christophides ym. (2015) esittää kaksi kategoriaa samankaltaisuusfunktioille linkitetyn datan vertailuun: sisältöön perustuvat ja entiteettien suhteisiin perustuvat menetelmät. Sisältöön perustuvilla menetelmillä heidän mukaansa mitataan käytännössä entiteettien ominaisuus-arvoparien samankaltaisuutta eli niiden ominaisuuksien arvoja verrataan toisiinsa. Näistä menetelmistä ehkä tunnetuin on Levenštеinin etäisyys (Levenštein, 1966), joka kertoo pienimmän määrän operaatioita (merkin lisääminen, poistaminen tai korvaaminen), joilla kaksi merkkijonoa voidaan muuntaa samaksi merkkijonoksi. Entiteettien suhteisiin perustuvia menetelmiä Christophides ym. puolestaan kuvaavat sellaisiksi, joissa entiteettejä verrataan toisiinsa perustuen niiden suhteisiin naapureidensa kanssa. Nämä menetelmät voidaan heidän mukaansa jakaa puupohjaisiin ja graafipohjaisiin menetelmiin taustalla olevan tietorakenteen perusteella.

Samankaltaisuusfunktioita käytetään siis tietueiden luokitteluun eli niiden avulla tehdään lopullinen päätös entiteettien jakamisesta toisiaan vastaaviin tai ei-vastaaviin pareihin. Aiemmin annetun vastaavuusfunktion M määrittelyn mukaan tämä tarkoittaa käytännössä oikean kynnysarvon θ asettamista, mikä voidaan tehdä esimerkiksi koneoppimisen menetelmillä (Christophides ym., 2015). Christen (2012) esittää myös muita tapoja luokitella entiteettiparit, esimerkiksi käyttäen todennäköisyyteen tai sääntöihin perustuvia menetelmiä.

4.2 W3C:n täsmäytysrajapintamääritelmä ja OpenRefine

Aiemmassa alaluvussa 4.1 käsiteltiin tietuelinkityksen teoriaa ja keskityttiin menetelmiin, joissa kaksi tietoaineistoa yhdistetään toisiinsa yhdellä kertaa ilman merkittävää panosta käyttäjältä. Täsmäytystä on mahdollista tehdä myös käyttäjäohjattuna tietuelinkityksen periaatteita soveltaen. Yksi merkittävimmistä käyttäjäohjattuja täsmäytysjärjestelmiä tukevista teknologioista on W3C:n entiteettien täsmäytys -yhteisöryhmän[12] määrittelemä

[12] https://www.w3.org/community/reconciliation/

täsmäytysrajapinta (Delpeuch ym., 2023). Rajapintamääritelmä antaa puitteet täsmäytysjärjestelmän tarjoamiselle HTTP(S)-palveluna. Tässä alaluvussa esitellään täsmäytysrajapintamääritelmä ja sitä hyödyntävä, taulukkomuotoisen datan käsittelyyn tarkoitettu OpenRefine-työkalu.

4.2.1 Täsmäytysrajapintamääritelmä

W3C:n rajapintamääritelmä määrittää tietomallin, johon täsmäytys perustuu. Tietolähteen, johon täsmäytyskyselyt kohdistuvat, oletetaan koostuvan entiteeteistä, joilla on tunniste (engl. id), nimi (engl. name), lista tyyppejä (engl. type) ja mahdollinen kuvaus (engl. description). Tyypit kuvaavat kategorioita, joihin entiteetit voidaan jakaa ja ne koostuvat tunnisteesta, nimestä sekä mahdollisesti listasta tyyppejä, jotka ovat tyypin ylätyyppejä (engl. broader). Entiteeteillä voi olla myös ominaisuuksia (engl. property), jotka koostuvat tunnisteesta ja nimestä. (Delpeuch ym., 2023.)

Rajapintamääritelmä määrittää kuusi toimintoa täsmäytysjärjestelmälle:

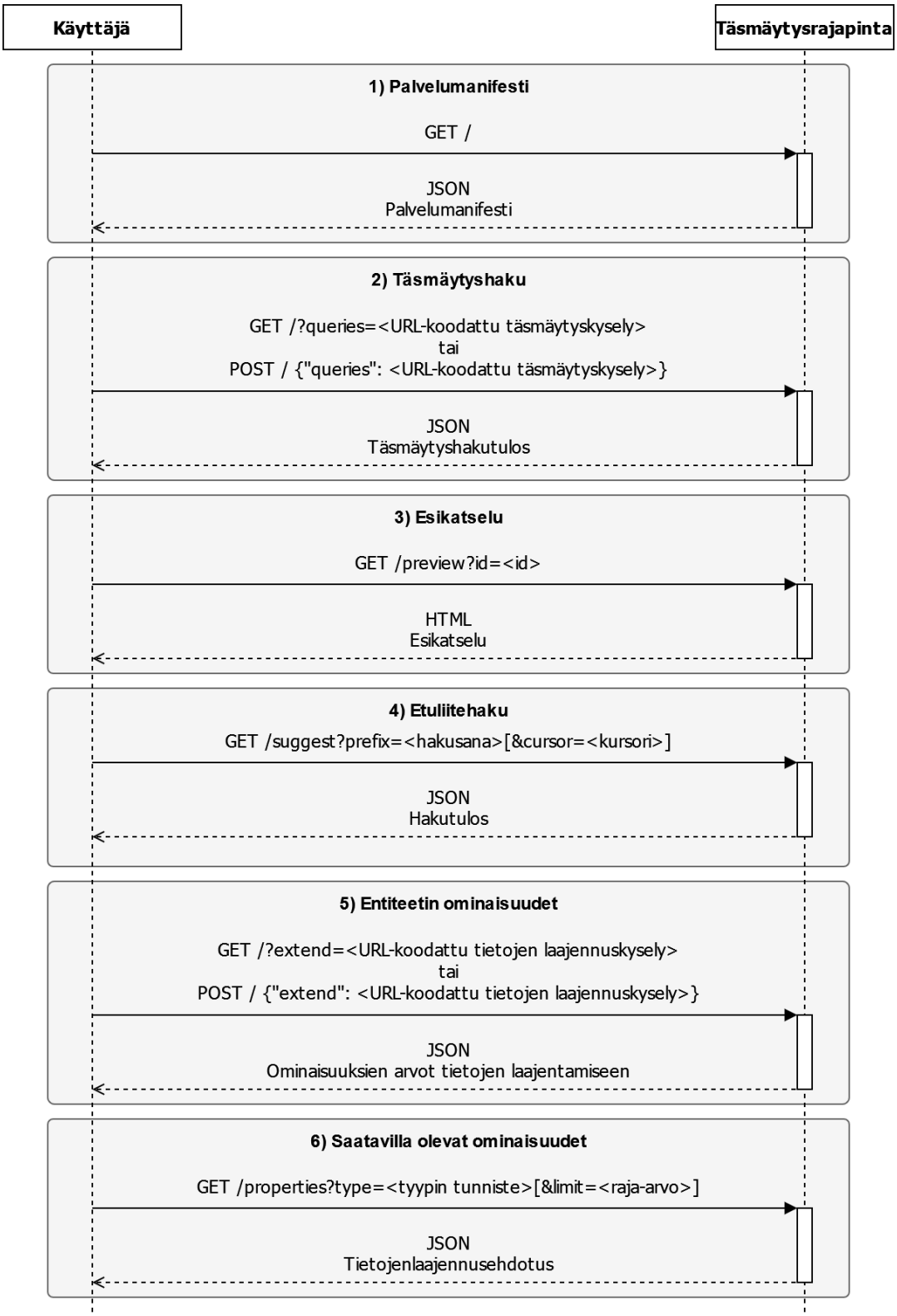
1) järjestelmää kuvaavan palvelumanifestin palauttaminen
2) täsmäytyshaku
3) entiteettien esikatselu
4) etuliitehaku
5) entiteetin ominaisuuksien arvojen palauttaminen
6) järjestelmän kautta saatavilla olevien ominaisuuksien palauttaminen

Nämä on esitetty kuvassa 4.2 sekvenssikaavion muodossa. Palvelumanifesti antaa kuvauksen täsmäytysjärjestelmästä ja sisältää esimerkiksi järjestelmän nimen, sen entiteettien oletustyypit sekä toiminnallisuuksien palvelupisteiden verkko-osoitteet. Täsmäytyshaku on palvelun tärkein toiminto: sen avulla käyttäjä linkittää entiteetit toisiinsa. Hakuun liitetään linkitettävien entiteettien nimet sekä mahdollisesti niiden tyypit ja ominaisuuksien arvot. Järjestelmä vastaa listalla hakutuloksia, jotka sisältävät muun muassa kandidaattientiteettien tunnisteet, nimet, tyypit ja pisteytyksen sekä tiedon siitä pitääkö palvelu kyseistä kandidaattia riittävän hyvänä linkitettäväksi. Pisteytys kertoo kuinka hyvin kandidaatti vastaa haussa annettua entiteettiä. Perushakutoiminnallisuuden lisäksi rajapinta tarjoaa kolme muuta toimintoa täsmäytyksen helpottamiseen. Entiteettien esikatselun (engl. preview service) avulla käyttäjä voi tarkastella entiteettien metatietoja, esimerkiksi niiden kuvauksia ja suhteita muihin entiteetteihin, HTML-tiedoston muodossa. Etuliitehaulla (engl. suggest service) käyttäjä voi hakea entiteettejä, tyyppejä tai ominaisuuksia yksi kerrallaan erillisten palvelupisteiden välityksellä. Sen tarkoituksena on mahdollistaa automaattinen täydennys eli sen odotetaan palauttavan vain tuloksia, jotka vastaavat annettua etuliitettä. Etuliitehaun tulokset sisältävät haetun entiteetin, tyypin tai ominaisuuden tunnisteen, nimen sekä mahdollisesti kuvauksen ja entiteettien tapauksessa myös tyypin. Tietojen laajentamistoiminnolla (engl. extend service) voidaan hakea

entiteeteille kutsussa määriteltyjen ominaisuuksien arvoja. Rajapinnan avulla on lisäksi mahdollista hakea ominaisuudet, jotka ovat oleellisia jonkin entiteettityypin kannalta. Rajapinta palauttaa käyttäjälle siis listan ominaisuuksista kysytylle tyypille tietojenlaajennusehdotuksena (engl. data extension property proposal). (Delpeuch ym., 2023.)

Rajapintamääritelmä jättää päätöksen linkitysten tekemisestä viime kädessä siis käyttäjälle. Se automatisoi prosessia antamalla ehdotuksia entiteeteistä ja pisteyttämällä ehdotukset, mutta käyttäjälle annetaan etuliitehaulla vapaus valita mikä tahansa entiteetti linkitettäväksi. Määritelmä ei yleisesti muutenkaan ota kantaa taustalla tehtävän tietuelinkitysprosessin vaiheisiin. Kandidaattientiteettien hakemisen nopeuttamiseksi voidaan täsmäytysjärjestelmän tietolähteen tietueet indeksoida niiden nimien tai ominaisuuksien perusteella. Rajapintamääritelmän toteuttavat täsmäytysjärjestelmät käyttävät useimmiten jotain olemassa olevaa hakukonetta (esimerkiksi ElasticSearch[13]) entiteettien löytämiseen ja käyttävät sen tarjoamia indeksointimenetelmiä (Delpeuch, 2019). Indeksointia voidaan jatkaa hakutuloksia suodattamalla käyttäen täsmäytyshakukyselyssä annettuja tyyppejä. Kandidaattientiteetit haetaan useimmiten tietolähteestä ennen niiden vertailua haettuun entiteettiin (Delpeuch, 2019). Vertailun tuloksena saadaan kaikille kandidaateille pisteytys, joka perustuu entiteettien nimiin ja mahdollisesti ominaisuuksiin. Täsmäytysjärjestelmät hyödyntävät myös pisteytyksessä usein käyttämänsä hakukoneen tuottamaa pisteytystä (Delpeuch 2019), mutta pisteet voidaan laskea myös käyttämällä esimerkiksi alaluvussa 4.1.3 esiteltyjä menetelmiä. Järjestelmä luokittelee kandidaatit lopuksi haettua entiteettiä vastaaviksi tai ei-vastaaviksi esimerkiksi asettamalla jonkin kynnysarvon pisteytykselle.

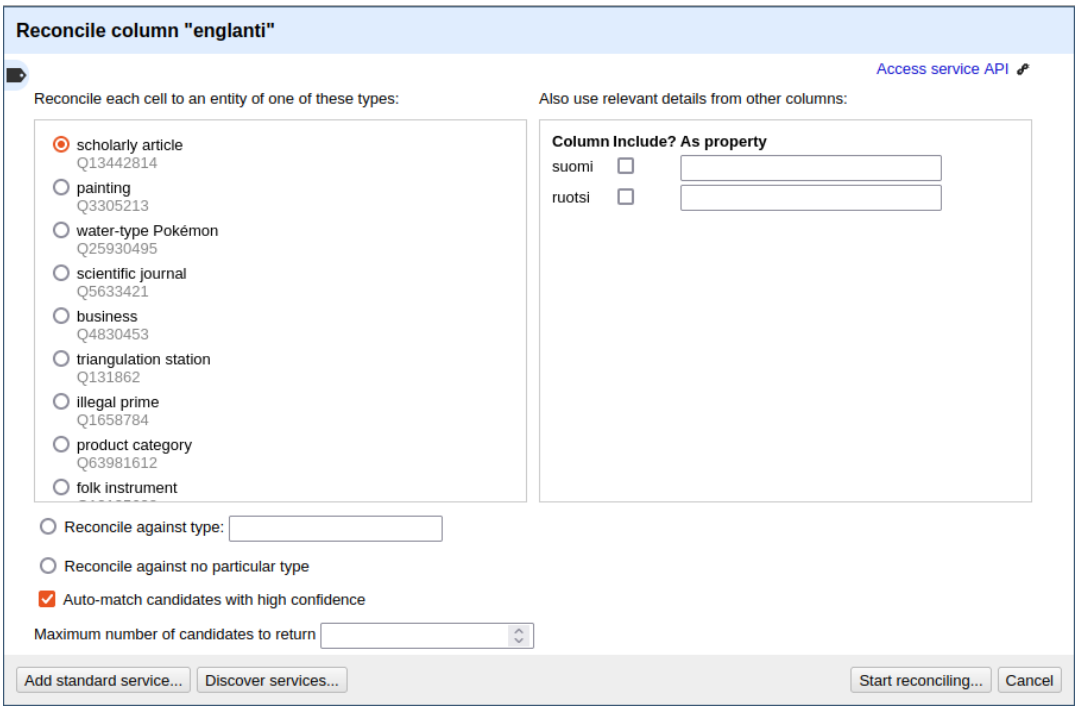
[13] https://www.elastic.co/elasticsearch



Kuva 4.2. Sekvenssikaavio W3C:n täsmäytysrajapintamääritelmän toiminnallisuuksista.

4.2.2 OpenRefine-ohjelmisto

W3C:n määritelmän mukaisten täsmäytysrajapintojen hyödyntämiseen yleisimmin käytetty ohjelmisto on OpenRefine (OpenRefine, 2023). Se tukee kaikkia edellisessä alaluvussa 4.2.1 esiteltyjä rajapintamääritelmän pääasiallisia ominaisuuksia. OpenRefine on integroitu kiinteästi Wikipediaa tukevan linkitetyn datan tietokannan Wikidatan täsmäytysrajapintaan ja se mahdollistaa muutosten lataamisen suoraan Wikidataan (Delpeuch, 2022). Sen avulla on mahdollista kuitenkin käyttää myös muita W3C:n määritelmän mukaisia rajapintoja. Myös muita työkaluja täsmäytysrajapintojen käyttöön on olemassa, esimerkiksi Cocoda[14], mutta tässä työssä keskitytään OpenRefine-ohjelmistoon, sillä se oli ollut käytössä Kansalliskirjastossa jo aiemmin.

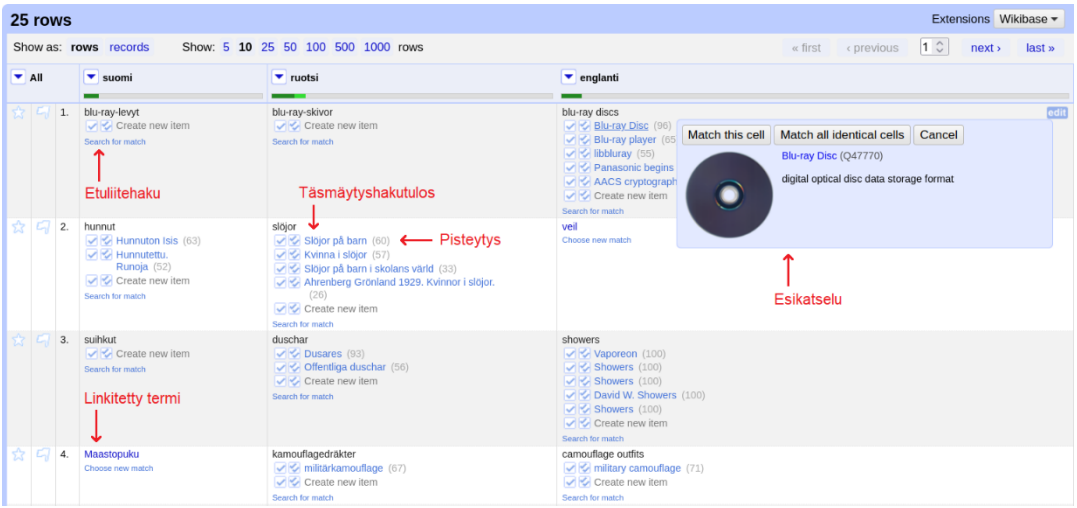


Kuva 4.3. OpenRefine-ohjelmiston näkymä täsmäytyksen parametrien asettamiseen.

OpenRefinen avulla on siis mahdollista täsmäyttää taulukkomuotoisia aineistoja käyttämällä mitä tahansa W3C:n rajapintamääritelmän toteuttavaa palvelua. Ennen täsmäytyshaun aloittamista täytyy käyttäjän määrittää haun parametrit, esimerkiksi entiteettien tyyppi, rajapinnan tarjoamilla asetuksilla (Delpeuch, 2023.). Kuvassa 4.3 on esimerkki Wikidatan englanninkielisen täsmäytysrajapinnan asetuksien määrittämisestä. Siinä näkyy entiteettien tyypin valinta vasemmalla sekä entiteettien ominaisuuksien arvojen asettaminen oikealla. Täsmäytyshaku voidaan tehdä ”Start reconciling...”-painikkeen avulla. Kuvassa 4.4 on puolestaan annettu esimerkki sanalistasta,

[14] https://coli-conc.gbv.de/cocoda/

jonka sarakkeet on täsmäytetty Wikidatan suomen-, ruotsin- ja englanninkielisillä rajapinnoilla. Käyttäjälle näytetään rajapinnan palauttamat täsmäytyshakutulokset ja niiden pisteytykset jokaiselle termille, mistä hän valitsee linkitettävät Wikidatan käsitteet. Osalle sanalistan termeistä on kuvassa tehty linkitys (esimerkiksi termille ”Maastopuku”). Käyttöliittymässä näytetään lisäksi käsitteiden esikatselu, kun niitä osoitetaan hiirellä. Jokaiselle solulle voidaan myös tehdä linkitys etuliitehaun avulla käyttämällä ”Search for match”-painiketta. Kuvaan on merkitty punaisella tekstillä ja nuolilla täsmäytyshaun tulokset ja niiden pisteytys, etuliitehaku, esikatselu sekä esimerkki linkitetystä termistä.



Kuva 4.4. OpenRefine-työkalun näkymä osittain täsmäytetystä sanalistasta. Oikeassa yläkulmassa näkyy ”Blue-ray Disc”-käsitteen[15] esikatselu.

4.3 Aiemmat täsmäytyksen toteutukset linkitetylle datalle

Tähän alalukuun on kerätty kirjallisuudesta esimerkkejä toteutetuista ratkaisuista linkitetyn datan täsmäytykseen. Ratkaisuja tarkastellaan niiden valitsemien tietueiden linkityksen menetelmien sekä yleisen arkkitehtuurin kannalta. Taulukkoon 4.1 on koottu käsiteltyjen täsmäytysjärjestelmien ominaisuuksia. Tässä alaluvussa tarkemmin tarkasteltujen järjestelmien lisäksi on toteutettu monia muita täsmäytysrajapintamääritelmän toteuttavia täsmäytyspalveluja. Useita niistä on listattu taulukkoon Wikidatan ylläpitämällä verkkosivulla, joka ilmaisee palveluiden palvelupisteet sekä niiden tukemat toiminnot (Reconciliation service test bench, n.d.).

[15] https://www.wikidata.org/wiki/Q47770

Taulukko 4.1. Käsiteltyjen täsmäytysjärjestelmien ominaisuuksien vertailu.

	Wikidata[16]	**ARIADNE**[17]	**lobid-gnd**[18]	**lobid-organisations**[19]	**Getty Vocabulary Program**[20]
Sanasto, jota varten kehitetty	Wikidata	AAT	GND	Saksan ISIL-rekisteri ja Deutsche Bibliotheksstatistiks -sanasto	ATT, ULAN, TGN
Mahdollista yleistää mille tahansa sanastolle helposti	Ei	Ei	Ei	Ei	Ei
Hakumenetelmä	Sisäiset hakurajapinnat ja SPARQL	SPARQL	ElasticSearch	ElasticSearch	Menetelmä ei tiedossa
Pisteytys	ElasticSearch	Ei pisteytystä	TF-IDF ja ElasticSearch	ElasticSearch	Menetelmä ei tiedossa
Käyttää W3C:n täsmäytysrajapintamääritelmää	Kyllä	Ei	Kyllä	Kyllä	Kyllä
W3C:n täsmäytysrajapintamääritelmästä toteutetut toiminnot	Kaikki päätoiminnot	-	Kaikki päätoiminnot	Ainoastaan täsmäytyshaku	Täsmäytyshaku, esikatselu, ominaisuuksien ja niiden arvojen haku sekä ominaisuuksien etuliitehaku

4.3.1 Wikidatan täsmäytyspalvelu

Wikidata tarjoaa W3C:n täsmäytysrajapintamääritelmän mukaisen palvelun täsmäytyksen tekemiseen. Rajapinta sallii täsmäytyshakujen tekemisen entiteettien nimien, ominaisuuksien ja tyyppien avulla. Se sisältää myös entiteettien esikatselun, etuliitehaun sekä ominaisuuksien ja niiden arvojen haun. Etuliitehaulla voidaan hakea entiteettejä, tyyppejä tai ominaisuuksia. Esikatselu näyttää entiteetin nimen, kuvan, tunnisteen ja kuvauksen, kuten kuvassa 4.4 on esitetty. (Delpeuch, 2020.)

Delpeuchin (2020) esittämä täsmäytysrajapinta on olemassa olevien rajapintojen päälle rakennettu kääre, joka kääntää siihen tehdyt kutsut Wikidatan sisäisten rajapintojen kutsuiksi. Hänen mukaansa täsmäytysrajapinta käyttää Wikidatan tarjoamia haku- ja täydennyspalveluja sekä SPARQL-kyselyitä kandidaattitietueiden löytämiseen. Näin saaduista kandidaateista suodatetaan pois ne, jotka eivät vastaa haetun entiteetin tyyppiä. Kandidaatit pisteytetään sen mukaan, kuinka samankaltaisia ne ovat

[16] https://wikidata.reconci.link/en/api
[17] https://vmt.ariadne.d4science.org/vmt/vmt-app.html
[18] https://lobid.org/gnd/reconcile
[19] https://lobid.org/organisations/reconcile
[20] https://services.getty.edu/vocab/reconcile/

verrattuna haettuun entiteettiin ominaisuuksien arvojen ja nimen perusteella. Pisteytys pohjautuu Levenšteinin etäisyyttä käyttävään mittaan (Delpeuch, 2019). Wikidatan täsmäytysrajapintaa ei siis voida suoraan ottaa käyttöön muille sanastoille tai alustoille, sillä se on rakennettu Wikidatan sisäisten toiminnallisuuksien varaan ja perustuu sen sisäisiin tietorakenteisiin.

Rajapinta pyrkii nopeuttamaan täsmäytysprosessia monin tavoin, esimerkiksi rinnakkaistamalla rajapintakutsuja. Se myös pyrkii välttämään hakurajapintojen käyttöä esimerkiksi tapauksissa, joissa hakuterminä käytetään Wikidatan tunnistetta, jolloin vastaava tietue palautetaan suoraan. Lisäksi, jos täsmäytyshaussa annetaan ominaisuutena tunniste, yritetään vastaava tietue löytää ilman hakurajapintoja SPARQL-kyselyllä. Wikidatan rajapinta käyttää ElasticSearch-hakukonetta ja sen hakuindeksointia hakutoiminnoissa, mikä nopeuttaa myös osaltaan kaikkia tehtyjä hakuja. (Delpeuch, 2020.)

4.3.2 ARIADNE-projekti

Arkeologisen tutkimuksen avuksi kehitetty ARIADNE-projekti integroi dataa useasta kontrolloidusta sanastosta ja tarjoaa sen verkkoportaalin kautta käytettäväksi (Meghini ym., 2017). Tietojen integroimista varten kehitettiin selainpohjainen työkalu, jonka avulla asiantuntijat pystyvät yhdistämään tietoja useista linkitetyn datan sanastoista (Binding & Tudhope, 2016).

Kontrolloidut sanastot linkitetään työkalun avulla yhteen keskussanastoon, Art & Architecture Thesaurus -tesaurukseen (AAT), mikä mahdollistaa niiden välisen yhdistetyn haun. Työkalu käyttää SPARQL-kyselyitä ja merkkijonojen sumeaa vertailua tietueiden löytämiseen AAT:stä ja linkitettävistä sanastoista. Työkalu esittää myös muuta semanttista informaatiota tietueista, kuten hierarkkisen kontekstin sekä lyhyen kuvauksen. Näiden tietojen avulla työkalua käyttävä asiantuntija luo haluamiansa linkitetyn datan suhteita tietueiden välille. (Binding & Tudhope, 2016.)

Linkitys siis tapahtuu käsin ja jokainen linkitettävä tietue tulee etsiä sanastoista erikseen. Työkalu ei myöskään pisteytä kandidaattitietueita lainkaan eikä osoita, mitkä haun tuloksista todennäköisimmin vastaisivat haluttua tietuetta. Tämä voi hidastaa ja vaikeuttaa linkitystä tekevän asiantuntijan työtä. Kyseinen lähestymistapa ei toisin sanoen automatisoi täsmäytysprosessia merkittävästi. Työkalu on lisäksi rakennettu tukemaan linkitystä ainoastaan AAT-sanastoon, eikä sen käyttöönottaminen muussa kontekstissa olisi helppoa.

4.3.3 Lobid-palvelu

Lobid-palvelu (engl. Linking Open Bibliographic Data) on saksalaisen korkeakoulukirjastopalveluiden keskuksen hbz:n (saksaksi hochschulbibliothekszentrum) tarjoama linkitetyn datan palvelu kirjastoille, joka sisältää

kolme sanastoa: lobid-gnd, lobid-organisations ja lobid-resources sekä niiden selainkäyttöliittymät ja hakurajapinnat (Pohl, 2018). Lobid-gnd- ja lobid-organisations-palveluille on tarjolla myös W3C:n täsmäytysrajapintamääritelmän mukaiset rajapinnat. Lobid-gnd pohjautuu saksalaisten kirjastoalan laitosten hallinnoiman Die Gemeinsame Normdatei (GND) -sanastoon (Steeg ym., 2019) ja lobid-organisations puolestaan Saksan ISIL-rekisteriin sekä kirjastoja kuvaavaan Deutsche Bibliotheksstatistik -sanastoon (hbz, n.d.). Rajapinnat on kehitetty toimimaan yllä mainittujen sanastojen kanssa osana lobid-palvelua, eikä niitä voi ottaa helposti muiden sanastojen tai organisaatioiden käyttöön.

Kuten Wikidatan rajapinta, myös lobid-gnd mahdollistaa entiteettien hakemisen niiden nimien, tyyppien ja ominaisuuksien perusteella (Steeg & Pohl, 2019). Se sisältää myös entiteettien esikatselun, etuliitehaun sekä ominaisuuksien ja niiden arvojen haun. Etuliitehaku on toteutettu entiteeteille, tyypeille ja ominaisuuksille. Esikatselu sisältää nimen ja tunnisteen lisäksi muuta tietoa entiteetistä, esimerkiksi henkilöä kuvaavaan tietueeseen liitetyn kuvan. Lobid-organisations-palvelun rajapinta puolestaan mahdollistaa täsmäytyshaun ainoastaan entiteeteillä, eikä se toteuta muita toiminnallisuuksia lainkaan (Reconciliation service test bench, n.d.).

Lobid-gnd-palvelun täsmäytysrajapinta käyttää tietueiden hakemiseen ElasticSearch-hakukonetta (Steeg ym., 2019). Järjestelmän lähdekoodista (hbz, 2024a) nähdään, että se tekee hakuja useisiin sanaston kenttiin, esimerkiksi tietueiden ensisijaisiin nimiin ja niiden tunnisteisiin. Täsmäytyskyselyssä annetut ominaisuudet ketjutetaan lähdekoodissa osaksi ElasticSearch-kyselyä ja ne vaikuttavat siten haun tuloksiin. Tuloksista myös suodatetaan koodissa pois ne, jotka eivät vastaa täsmäytyskyselyssä annettuja tyyppejä. Täsmäytyshaun tulosten pisteytyksessä rajapinta käyttää ElasticSearchin antamaa pisteytystä sekä TF-IDF-menetelmää (engl. term frequency–inverse document frequency) pisteiden antamiseen (Delpeuch, 2019). Lobid-organisations-palvelun lähdekoodista (hbz, 2024b) nähdään, että se toimii samaan tapaan ja käyttää ElasticSearch-kyselyjä tietueiden hakemiseen ja pisteytykseen.

4.3.4 Getty Vocabulary Program -täsmäytyspalvelu

Getty Research Institute (GRI) -tutkimuslaitoksen linkitetyn datan Getty Vocabularies -ohjelma tarjoaa usean linkitetyn datan sanaston taiteen alalle, muun muassa AAT:n, Getty Thesaurus of Geographic Names -tesauruksen (TGN) ja Union List of Artist Names -sanaston (ULAN) (Getty Research Institute, n.d.). Näille kolmelle sanastolle tarjotaan yhteinen W3C:n määritelmän mukainen täsmäytysrajapinta. Rajapinnassa kohdesanasto valitaan asettamalla kyselyn tyypiksi sanaston nimi. (Garcia, 2023.)

GRI:n rajapinta toteuttaa täsmäytyshaun lisäksi entiteettien esikatselun sekä ominaisuuksien ja niiden arvojen haun. Etuliitehaku tarjotaan

ainoastaan ominaisuuksille. Täsmäytyshaussa voidaan käyttää ominaisuuksia käsitteiden rajaamiseen, mutta tyypillä valitaan sanasto, johon täsmäytys kohdistuu, joten muita käsitteiden tyyppejä ei voida hyödyntää. Esikatselu sisältää tietoja entiteeteistä, esimerkiksi henkilön syntymä- ja kuolinajan, esineen kuvauksen tai käsitteen sijoittumisen sanaston hierarkiaan. (Garcia, 2023.)

GRI:n sanastojen rajapintaa ei ole toteutettu avoimen lähdekoodin ohjelmistona, joten sen sisäisestä toiminnasta, esimerkiksi käsitteiden hakumenetelmästä tai pisteytyksen määräytymisestä, ei ole saatavilla tietoa. Se ei siten myöskään sovellu muiden sanastojen tai organisaatioiden käyttöön.

4.3.5 Soveltuvuus Kansalliskirjaston tarpeisiin

Mitään tässä alaluvussa tarkastelluista järjestelmistä ei voida helposti ottaa käyttöön Skosmoksen tai Annifin kanssa, joten on käytettävä jotain muuta lähestymistapaa. Tietuelinkitykseen on olemassa monia työkaluja, esimerkiksi Python-kirjasto dedupe[21], mutta ne keskittyvät pääosin taulukkomuotoisen datan käsittelyyn ja eivät siten sovellu linkitetyn datan sanastojen kanssa työskentelyyn. W3C:n määritelmän mukaisten täsmäytysrajapintojen perustamiseen on olemassa joitakin työkaluja, esimerkiksi Reconcile-csv[22] ja csv-reconcile[23], jotka luovat rajapinnan CSV-muotoiselle datalle, eli ne eivät myöskään sovellu suoraan linkitetyn datan sanastojen käyttöön. RDF Extension[24] on työkalu, jonka avulla voidaan käyttää SPARQL-rajapintoja täsmäytyksessä, mutta se toimii OpenRefine-työkalun laajennuksena eikä sen avulla siis voida tarjota täsmäytyspalvelua verkon yli. Toisin sanoen ei ole olemassa järjestelmää, jonka avulla voitaisiin helposti ottaa käyttöön täsmäytyspalvelu mielivaltaisia RDF-muotoisia sanastoja varten. Tämän vuoksi Skosmokselle ja Annifille on mielekkäintä toteuttaa omat järjestelmänsä alusta alkaen.

Kuten tässä alaluvussa on näytetty, linkitetyn datan sanastojen täsmäytykseen hyvin yleisesti käytetty menetelmä on W3C:n määritelmään perustuva täsmäytysrajapinta. Wikidatan sivustolla rajapintoja on listattu tämän työn kirjoittamisen aikaan 80 (Reconciliation service test bench, n.d.) ja tässä alaluvussa käsitellyistä palveluista neljä viidestä perustuivat rajapintamääritelmään. Se on siis laajasti käytössä oleva teknologia, jota myös monet työkalut tukevat. Tämä on osasyy sille, miksi se otettiin käyttöön tässä työssä kehitettävissä täsmäytysjärjestelmissä. Alaluvussa 6.1. käsitellään teknologiavalinnan perusteluita tarkemmin.

Vaikka tässä alaluvussa käsiteltyjä järjestelmiä ei voidakaan suoraan ottaa käyttöön, voivat ne auttaa toteutettavien järjestelmien kehittämisessä. Esimerkiksi tapaa, jolla Wikidatan täsmäytysrajapinta on toteutettu kääreenä

[21] https://github.com/dedupeio/dedupe
[22] https://github.com/okfn/reconcile-csv
[23] https://github.com/gitonthescene/csv-reconcile
[24] https://github.com/stkenny/grefine-rdf-extension

sen rajapintojen päälle, voidaan hyödyntää Skosmoksen täsmäytysjärjestelmässä käyttämällä sen REST-rajapintaa.

5 Tutkimusmenetelmät

Tässä luvussa esitellään opinnäytetyössä käytetyt tutkimusmenetelmät. Ensimmäisessä alaluvussa kuvataan käytetty tutkimusprosessi tarkemmin ja esitellään pääasiallinen tutkimusmenetelmä, suunnittelutiede. Toisessa alaluvussa puolestaan esitellään tutkimuksessa käytetyt datalähteet ja datan keräysmenetelmät.

5.1 Suunnittelutiede

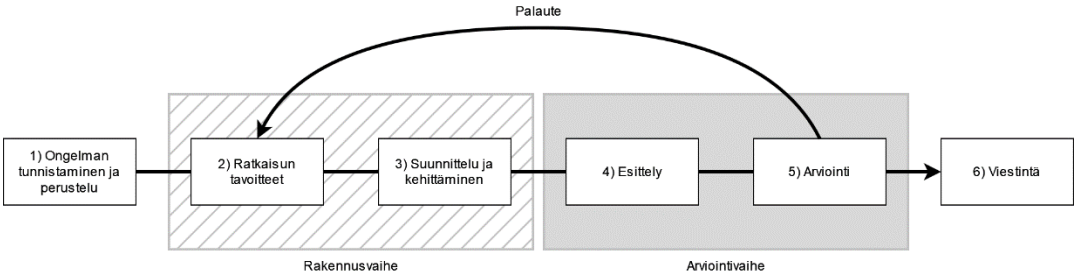
Johannesson ja Perjons (2014) määrittelevät suunnittelutieteen tutkimusparadigmana, jossa kehitetään ja tutkitaan artefakteja, joilla ratkaistaan käytännön ongelmia. Kehitettävän artefaktin tulee Hevnerin ym. (2004) mukaan ratkaista jokin aiemmin ratkaisematon ongelma. Tämä voi tarkoittaa joko olemassa olevan tietopohjan laajentamista tai vanhan tiedon soveltamista uusilla ja innovatiivisilla tavoilla. Informaatiotieteessä artefaktit voidaan jakaa neljään tyyppiin: konstruktiot, mallit, menetelmät ja instanssit (March & Smith, 1995). Konstruktiot esittävät sanaston ja symbolit, joilla kuvataan ongelmia ja niiden ratkaisuja. Niitä voidaan käyttää hyväksi rakentamaan malleja, joilla puolestaan kuvataan erilaisia tilanteita. Konstruktioihin ja malleihin pohjautuen voidaan luoda metodeja, jotka määrittävät toimet jonkin tehtävän suorittamiseen. Instanssi puolestaan toteuttaa konkreettisen työkalun tai tietojärjestelmän käyttäen hyväkseen konstruktioita, malleja ja metodeja. (March & Smith, 1995). Johdannossa esitetyt tutkimuskysymykset selvästi ohjaavat konkreettisen tietojärjestelmän kehittämiseen, niinpä tässä opinnäytetyössä toteutettavat artefaktit ovat tyypiltään instansseja. Artefakteina toimii kaksi täsmäytysjärjestelmäprototyyppiä, jotka on kehitetty Annif- ja Skosmos-ohjelmien avulla. Rajapinnat noudattavat alaluvussa 4.2.1 esiteltyä täsmäytysrajapintamääritelmää ja ne on kuvattu tarkemmin luvussa 6.

Peffers ym. (2007) esittävät mallin suunnittelutieteen tutkimusprosessille (engl. Design science research process, DSRP), joka koostuu kuudesta osasta:

1) ongelman tunnistaminen ja perustelu
2) ratkaisun tavoitteet
3) suunnittelu ja kehittäminen
4) esittely
5) arviointi
6) viestintä

Toisaalta Hevner ym. (2004) luonnehtivat artefaktin kehittämistä eräänlaiseksi hakuprosessiksi ja painottavat suunnittelutieteen luontaista iteratiivisuutta. He esittävät, että artefakti tulisi kehittää sykleissä, jotka koostuvat arviointivaiheesta ja rakennusvaiheesta. Arviointivaiheessa saadaan palautetta suunnitteluprosessin ja artefaktin laadusta, jota sitten käytetään

rakennusvaiheessa niiden edelleen kehittämiseen. DSRP ja Hevnerin ym. (2004) esittämä prosessi on yhdistetty tässä työssä käytettäväksi tutkimusprosessiksi. Rakennusvaihe siis koostuu DSRP:n osista 2 ja 3 ja arviointivaihe osista 3 ja 4. Rakennusvaiheessa valmistunutta artefaktia käytetään tarkoituksenmukaisessa kontekstissa ja sen toiminnallisuutta arvioidaan tarkoituksenmukaisella tavalla. Palautteen perusteella artefaktin vaatimukset arvioidaan uudelleen ja sitä kehitetään edelleen vastaamaan uusiin vaatimuksiin. Prosessi on esitetty kuvassa 5.1.



Kuva 5.1. Diplomityössä käytetty tutkimusprosessi. Rakennusvaiheen osat on esitetty viivoitetulla taustalla ja arviointivaiheen osat yhtenäisellä taustalla. Mukailtu Peffersin ym. (2007) kuvasta 1.

Venable ym. (2016) esittävät viitekehyksen suunnittelutieteen arvioinnille, joka pohjautuu kahteen ulottuvuuteen. Ensimmäisessä ulottuvuudessa erotetaan toisistaan formatiiviset ja summatiiviset (engl. formative ja summative) arvioinnin menetelmät. Formatiivisella arvioinnilla pyritään parantamaan arvioitavan prosessin tuloksia, kun taas summatiivisella arvioinnilla pyritään arvioimaan, missä määrin tulokset vastaavat odotuksia. Toinen ulottuvuus puolestaan erottaa toisistaan keinotekoisen ja naturalistisen (engl. artificial ja naturalistic) arvioinnin. Keinotekoinen arviointi testaa hypoteeseja laboratoriokokeiden ja simulaatioiden kaltaisten menetelmien avulla. Naturalistinen arviointi puolestaan tutkii ratkaisun suorituskykyä sen todellisessa ympäristössä esimerkiksi tapaustutkimuksen avulla. Venable ym. kuvaavat arviointiprosessia tutkimuksen aikana etenemiseksi formatiivisista summatiivisiin arvioihin ja keinotekoisista naturalistisiin arvioihin. He kutsuvat tapaa, jolla eteneminen tapahtuu, arviointistrategiaksi. He esittävät neljä strategiaa: nopea ja yksinkertainen, inhimillisten riskien arviointi, teknisten riskien arviointi ja puhtaasti tekninen arviointi. Nopeassa ja yksinkertaisessa arvioinnissa arviointi-iteraatioita toteutetaan suhteellisen vähän ja siinä edetään nopeasti summatiiviseen ja naturalistiseen arviointiin. Se onkin valittu strategiaksi tähän työhön, sillä rakennettavat artefaktit ovat suhteellisen yksinkertaisia ja arviointi-iteraatioita ei ole mahdollista toteuttaa suurta määrää johtuen rajallisesta ajasta ja koehenkilöiden rajallisesta saatavuudesta.

Peffers ym. (2012) löysivät kahdeksan suunnittelutieteellisessä tutkimuksessa käytettyä arviointimenetelmää. Heidän mukaansa käytetyimmät

menetelmät olivat tekninen testaus, jossa arvioidaan artefaktin teknistä suorituskykyä (eikä vaikutusta todelliseen maailmaan) sekä havainnollistava skenaario, jossa artefaktia sovelletaan todellisen maailman tilanteeseen. Tähän työhön on pääasialliseksi arviointimenetelmäksi valittu näistä jälkimmäinen kontekstuaalisten haastattelujen muodossa. Arviointimenetelmiä käsitellään tarkemmin alaluvussa 5.2.

Hevner ym. (2004) esittävät 7 ohjesääntöä suunnittelutieteen tekemiseen. Ne on listattu taulukossa 5.1. Tässä opinnäytetyössä on pyritty rakentamaan tutkimusprosessi näiden ohjesääntöjen mukaiseksi ja noudattamaan niitä koko tutkimuksen ajan. Ohjesääntöjen noudattamisen onnistumista on arvioitu alaluvussa 7.2.

Taulukko 5.1. Suunnittelutieteen ohjesäännöt (Hevner ym., 2004).

Ohjesääntö	**Kuvaus**
1. Artefaktin tuottaminen	Suunnitelutieteen tutkimuksen on tuotettava käyttökelpoinen artefakti konstruktion, mallin, menetelmän tai instanssin muodossa.
2. Ongelman merkitys	Suunnittelutieteen tutkimuksen tavoitteena on kehittää teknologisia ratkaisuja tärkeisiin ja merkityksellisiin ongelmiin.
3. Artefaktin arviointi	Artefaktin hyödyllisyys, laatu ja tehokkuus on osoitettava täsmällisesti hyvin laadittujen arviointimenetelmien avulla.
4. Tutkimuksen kontribuutio	Tehokkaan suunnittelutieteen tutkimuksen on tarjottava selkeä ja todennettavissa oleva kontribuutio toteutetun artefaktin avulla.
5. Tutkimuksen tarkkuus	Suunnittelutieteellinen tutkimus perustuu tarkkojen menetelmien soveltamiseen sekä artefaktin rakentamisessa että arvioinnissa.
6. Suunnittelu hakuprosessina	Toimivan artefaktin etsiminen edellyttää käytettävissä olevien keinojen hyödyntämistä haluttujen päämäärien saavuttamiseksi ja samalla ongelmaympäristön lainalaisuuksien täyttämistä.
7. Tutkimuksesta tiedottaminen	Suunnittelutieteen tutkimus on esiteltävä tehokkaasti sekä teknologiaan että johtamiseen suuntautuneelle yleisölle.

5.2 Datan keräys ja arviointimenetelmät

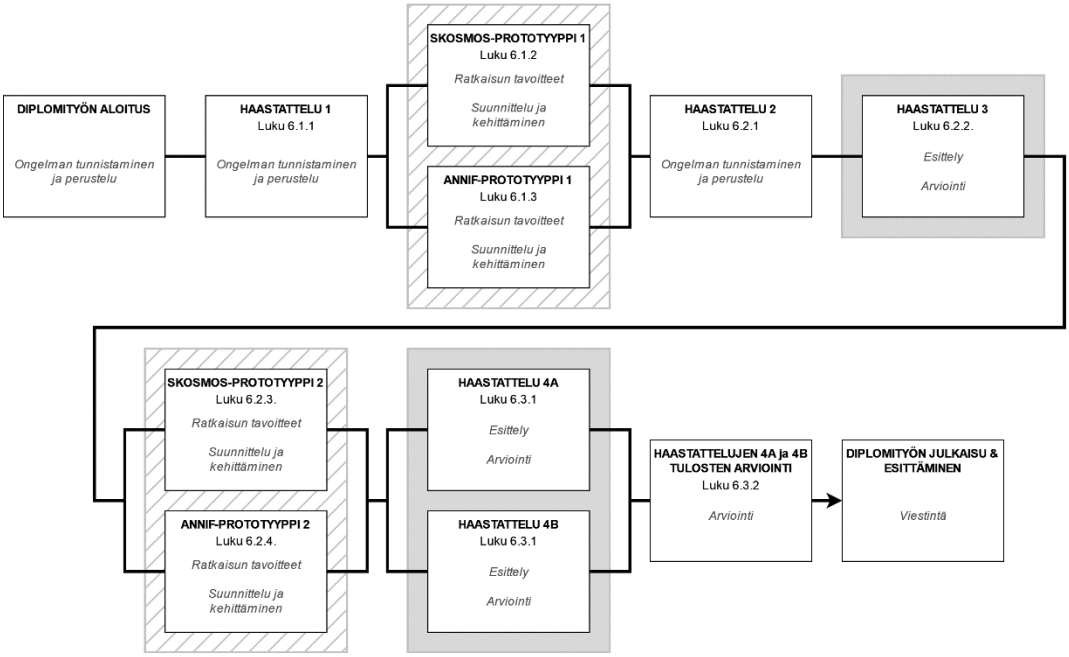
Yhtenä diplomityön pääasiallisena datalähteenä ja arviointimenetelmänä käytettiin Kansalliskirjaston sanastotyön tekijöiden kanssa suoritettuja haastatteluja. Haastattelukierroksia toteutettiin yhteensä 4 diplomityöprosessin eri vaiheissa. Haastatteluista kahdessa käytettiin datan keräysmenetelmänä kontekstuaalista haastattelua (engl. contextual interview), jossa haastateltavaa tarkkaillaan tosielämän tilanteessa ja hänelle esitetään kysymyksiä hänen toiminnastaan (Holtzblatt, 2005, s. 80). Kontekstuaalisen haastattelun etuna on, että siinä saadaan paljon tietoa haastateltavan toiminnasta hänen todellisessa ympäristössään. Muut haastattelut toteutettiin avoimena ja puolistrukturoituna haastatteluna.

Aivan diplomityöprosessin alussa pidettiin avoin haastattelu (Haastattelu 1) Kansalliskirjaston sanastokehittäjille sekä teknisille kehittäjille. Sen tavoitteena oli saada käsitys siitä, mitä tarpeita Kansalliskirjastossa on täsmäytykselle sekä kehitettäville täsmäytysjärjestelmille ja siten vastata tutkimuskysymykseen 1. Tämän jälkeen pidettiin ensimmäinen kontekstuaalinen haastattelu (Haastattelu 2) eräälle Kansalliskirjaston sanastokehittäjälle. Tämän haastattelun tavoitteena oli kartuttaa tietoa täsmäytyksen työnkuluista Kansalliskirjastossa sekä vastata tutkimuskysymyksiin 1 ja 2. Seuraavaksi pidettiin puolistrukturoitu haastattelu (Haastattelu 3) kahdelle Kansalliskirjaston sanastokehittäjälle, jonka tarkoituksena oli kerätä palautetta toteutettujen täsmäytysjärjestelmien toiminnasta ja käytettävyydestä sekä vertailla niitä toisiinsa. Tämä haastattelu vastasi tutkimuskysymyksiin 2 ja 3. Lopuksi pidettiin kaksi viimeistä kontekstuaalista haastattelua (Haastattelut 4A ja 4B), jotka koostuivat tehtäväosiosta ja kysymysosiosta. Niiden tavoitteena oli kerätä tietoa järjestelmien käytettävyydestä ja vastata tutkimuskysymykseen 3.

Haastattelun 4 tuloksena saatiin lisäksi yhteensä neljä haastateltavien tekemää, rajapintojen avulla tuotettua linkitystä YSO-sanastoon. Näitä linkityksiä verrattiin toisiinsa ja analysoitiin laadullisesti artefaktien toimivuuden arvioimiseksi. Haastattelujen tuloksista sekä tuotetuista rajapinnoista on kerrottu tarkemmin luvussa 6.

6 Tulokset

Tässä luvussa käsitellään haastattelujen tulokset ja kuvataan kehitettyjen rajapintojen arkkitehtuuri ja niiden toiminnallisuus. Luvussa edetään kronologisesti tutkimusprosessin läpi. Ensimmäisessä alaluvussa käsitellään Haastattelu 1 ja ensimmäiset rajapintaprototyypit. Seuraavassa alaluvussa käsitellään Haastattelut 2 ja 3 sekä lopulliset prototyypit. Viimeisessä alaluvussa käsitellään Haastattelut 4A ja 4B ja niiden tuloksia analysoidaan. Kuva 6.1 havainnollistaa tämän luvun sekä tutkimuksen etenemisen.



Kuva 6.1. Diplomityön toteutunut tutkimusprosessi. Rakennusvaiheen osat on esitetty viivoitetulla taustalla ja arviointivaiheen osat yhtenäisellä harmaalla taustalla.

6.1 Ensimmäinen iteraatio

Tutkimusprosessin ensimmäisessä vaiheessa määriteltiin ratkaistava ongelma ja selvitettiin ratkaisun alustavat tavoitteet. Tämä tehtiin avoimen haastattelun avulla. Asetettujen tavoitteiden saavuttamiseksi toteutettiin ensimmäiset rajapintaprototyypit.

6.1.1 Haastattelu 1

Haastattelussa 1 selvitettiin Kansalliskirjaston aiempia käytäntöjä täsmäytyksen tekemisessä, siihen käytettyjä menetelmiä sekä tarpeita siihen tulevaisuudessa. Haastattelu toteutettiin avoimena ryhmähaastatteluna, eikä siihen laadittu tarkkoja kysymyksiä etukäteen. Haastatteluun osallistui

Kansalliskirjaston sanastokehittäjiä sekä teknisiä kehittäjiä. Haastattelusta luotiin sen aikana muistiinpanot, jotka kirjoitettiin puhtaaksi jälkikäteen.

Haastattelussa selvisi, että sanastokehittäjät olivat tehneet täsmäytystä Kansalliskirjastossa pääasiassa käyttäen OpenRefine-työkalua, jonka he olivat kokeneet hyväksi tähän tarkoitukseen. Tekniset kehittäjät olivat lisäksi kehittäneet pieniä itsenäisiä ohjelmistoja yksittäisten sanastojen ja sanalistojen täsmäytykseen tarpeen mukaan. Täsmäytyksen ensisijainen käyttötapaus kirjastossa oli YSO:n käsitteiden linkittäminen Wikidata-palveluun käyttäen OpenRefine-työkalua. Tärkeimpinä tulevaisuuden tarpeina täsmäytykselle haastateltavat mainitsivat kahden sanaston linkittämisen toisiinsa (käyttäen esimerkiksi skos:exactMatch- tai skos:closeMatch-suhteita) sekä erillisten sanastojen yhdistämisen uudeksi sanastoksi. Myös linkittämättömien sanalistojen linkitys olemassa oleviin sanastoihin tuotiin esiin mahdollisena käyttökohteena. Tärkeimpänä Kansalliskirjaston ylläpitämänä kohdesanastona haastateltavat pitivät YSO:a, mutta myös kansallisbibliografisten toimijatietojen KANTO[25] ja suomalaisten ydinontologioiden kokoelma KOKO[26] mainittiin merkittävinä sanastoina.

Haastattelun 1 sekä alaluvussa 4.3.5. esitettyjen huomioiden perusteella päätettiin toteuttaa W3C:n täsmäytysrajapintamääritelmän mukainen täsmäytyspalvelu vastaamaan Kansalliskirjaston tarpeisiin. Haastattelussa tuli ilmi, että täsmäytystyön tekemistä halutaan jatkaa OpenRefine-ohjelmiston avulla, joten W3C:n määritelmän käyttäminen on luonnollista. Aiemmin esitellyt Skosmos- ja Annif-ohjelmistot tarjoavat toimintoja, jotka mahdollistavat täsmäytysrajapintojen integraation niihin, joten tässä työssä kehitetään täsmäytysrajapintaprototyypit molempien ohjelmistojen avulla. Rajapintojen integroiminen Skosmokseen ja Annifiin mahdollistaa myös linkitysten tekemisen millä tahansa niiden tarjoamilla sanastoilla, minkä haastateltavat mainitsivat yhtenä tarpeenaan täsmäytystyössä. Kehittämällä kaksi erillistä prototyyppiä voidaan tutkia erilaisia lähestymistapoja täsmäytykseen ja saada laajemmin tietoa kulttuuriperintöorganisaation tarpeisiin vastaavien täsmäytysjärjestelmien rakentamisesta. Lisäksi ennen järjestelmien kehityksen aloittamista ei ollut selvää kumpi ohjelmistoista soveltuisi täsmäytyksen tekemiseen paremmin.

Haastattelun perusteella määriteltiin vaatimukset toteutettaville rajapinnoille. Molemmilta prototyypeiltä vaadittiin, että ne tarjoavat verkko-osoitteen, jonka kautta täsmäytyspalvelu on saatavilla sekä täsmäytystoiminnallisuuden yksinkertaisimmassa muodossaan. Nämä ovat myös vähimmäisvaatimukset täsmäytysrajapinnalle, jotta se olisi yhteensopiva OpenRefine-työkalun kanssa. Seuraavissa alaluvuissa 6.1.2. ja 6.1.3. esitetään tarkempi tekninen kuvaus rajapinnoista.

[25] https://finto.fi/finaf/
[26] https://finto.fi/koko/

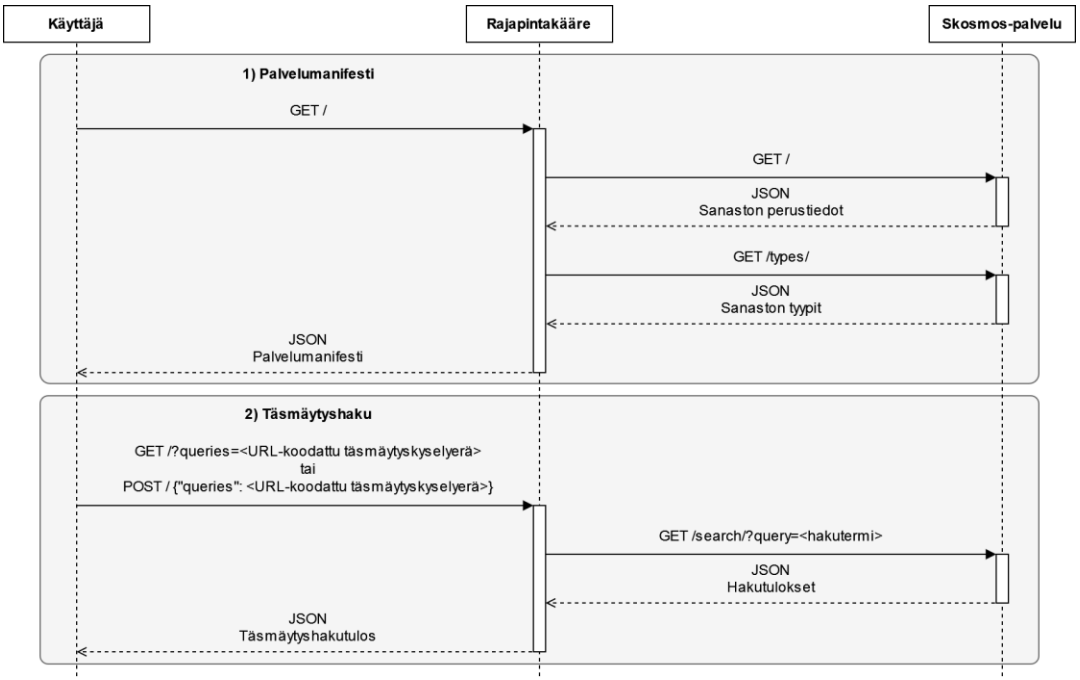
6.1.2 Skosmos-rajapintaprototyyppi 1

Ensimmäinen täsmäytysrajapintaprototyyppi Skosmokselle toteutettiin rajapintakääreenä sen oman rajapinnan päälle eli sen tehtävänä on kääntää täsmäytysrajapintaan tulevat kutsut Skosmos-palvelun, esimerkiksi Finton, rajapintakutsuiksi. Toteutettu prototyyppi käyttää hyväkseen Finto-palvelun rajapintaa täsmäytyshakujen tekemiseen ja kohdesanaston perustietojen hakemiseen, mutta on yhteensopiva myös minkä tahansa muun Skosmos-palvelun kanssa. Lisäksi sillä voidaan käsitellä kaikkia Skosmos-palvelussa olevia sanastoja. Prototyyppi kehitettiin Python-ohjelmointikielellä (versio 3.10) ja Flask-verkkosovelluskehyksellä (versio 2.3). Flask mahdollistaa yksinkertaisten verkkosovellusten toteuttamisen helposti ja nopeasti, minkä takia se valittiin käyttöön rajapintaan.

Kuten Haastattelun 1 perusteella oli päätetty, rajapinnasta tehtiin minimaalinen täsmäytysrajapintamääritelmän versiota 0.2 noudattava toteutus. Se tarjoaa yhden palvelupisteen verkko-osoitteessa ”/<sanastoID>/reconcile”. Palvelupiste palauttaa joko 1) palvelumanifestin tai 2) täsmäytyshaun tulokset JSON-objektina, riippuen lähetetyn HTTP-pyynnön rakenteesta. Rajapinnan toiminnot on esitetty kuvassa 6.2 sekvenssikaavion muodossa. Palvelumanifesti sisältää rajapintamääritelmän mukaiset pakolliset kentät, joiden arvot haetaan Skosmos-palvelun rajapinnasta. Pakollisista kentistä ”SchemaSpace”, joka määrittää URI:n rajapinnan palauttamien entiteettien luokalle, jätettiin kuitenkin tyhjäksi, sillä resurssit Skosmos-sanastoissa ei ole osa mitään tiettyä luokkaa. Täsmäytyshaku tukee pakollista ”queries”-kenttää sekä valinnaisia ”type”- ja ”limit”-kenttiä. Se tekee pyynnön Skosmos-palvelun rajapinnan hakumetodille jokaista haettua entiteettiä kohden ja palauttaa hakutulokset, joissa entiteettien tunnisteina toimivat Skosmos-resurssien URI:t ja niminä resurssien skos:prefLabelit. Hakutuloksia ei pisteytetä, vaan pistekentän arvoksi asetetaan aina 1. Rajapintaprototyyppi ei huomioi kyselyn kieltä, joten hakutulokset haetaan kaikkien kielten perusteella ja tulokset ovat aina suomenkielisiä. JSON-muotoinen esimerkki palvelumanifestista löytyy liitteestä A.

Rajapintakääreelle perustettiin testi-instanssi Kansalliskirjaston hallitsemalle palvelimelle seuraavaa arviointi-iteraatiota varten. Se käyttää hyväkseen Finto-palvelun kehitysversion rajapintaa[27] ja mahdollistaa täsmäytyksen kaikkiin sen tarjoamiin sanastoihin.

[27] https://api.dev.finto.fi/rest/v1/



Kuva 6.2. Sekvenssikaavio ensimmäisen Skosmos-prototyypin toiminnallisuuksista.

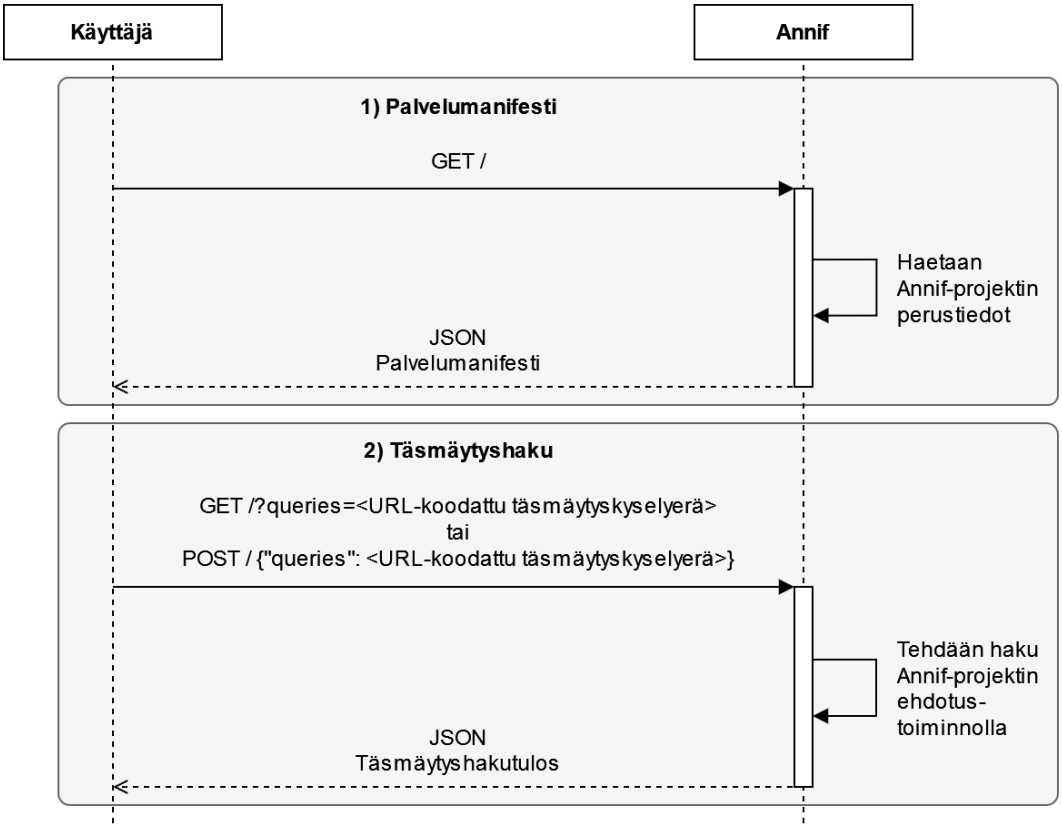
6.1.3 Annif-rajapintaprototyyppi 1

Ensimmäinen rajapintaprototyyppi Annifille toteutettiin sen olemassa olevan rajapinnan laajennuksena. Se käyttää siis hyväkseen Annifin toimintoja omassa toiminnallisuudessaan ja on yhteensopiva kaikkien Annif-instanssiin määriteltyjen projektien (ja siten sanastojen) kanssa. Annif on toteutettu Python-ohjelmointikielellä ja käyttää REST-rajapinnan toteutuksessa Flask- ja Connexion-verkkosovelluskehyksiä (Connexion versio 2.14.2).

Samoin kuin Skosmos-prototyyppistä, myös Annif-prototyypistä toteutettiin Haastattelun 1 perusteella minimaalinen täsmäytysrajapintamääritelmän versiota 0.2 noudattava versio. Myös Annifin rajapinta tarjoaa yhden palvelupisteen palvelumanifestin hakemiseen sekä täsmäytyshakujen tekemiseen verkko-osoitteessa ”/projects/<projektiID>/reconcile”. Nämä on esitetty kuvassa 6.3 sekvenssikaavion muodossa. Palvelumanifesti sisältää rajapintamääritelmän mukaiset pakolliset kentät, mutta ”identifierSpace”-kenttä (URI-nimiavaruus, johon palautettavat entiteetit kuuluvat) jätettiin kuitenkin tyhjäksi, sillä Annif ei edellytä entiteettien kuuluvan mihinkään tiettyyn URI-avaruuteen. Manifestissa palautetaan vain yksi oletustyyppi, koska Annifissa entiteeteille ei ole määritelty tyyppejä. Täsmäytyshaku tukee pakollista ”queries”-kenttää sekä valinnaista ”limit”-kenttää. ”Types”-kentän arvoa ei oteta huomioon täsmäytyshauissa. Haut tehdään käyttämällä Annifin projektin ehdotustoiminnallisuutta, joka palauttaa ehdotetuille asiasanoille URI:n, nimen sekä pisteytyksen. Nämä annetaan täsmäytyshaun

tuloksena. Hakutulosten kieli määräytyy Annifin projektin kielen perusteella. Liitteestä A löytyy esimerkki JSON-muotoisesta palvelumanifestista.

Prototyypistä luotiin Docker-kontti, joka lisättiin Kansalliskirjaston hallitsemaan projektiin OpenShift-konttialustalle seuraavaa arviointi-iteraatiota varten. Testi-instanssiin lisättiin kolme YSO:on perustuvaa projektia suomen-, ruotsin ja englannin kielellä.



Kuva 6.3. Sekvenssikaavio ensimmäisen Annif-prototyypin toiminnallisuuksista.

6.2 Toinen iteraatio

Tutkimusprosessin toisessa vaiheessa laajennettiin ja tarkennettiin vaatimuksia ratkaisulle kontekstuaalisen sekä puolistrukturoidun haastattelun avulla. Uusia tavoitteita vastaamaan toteutettiin lopulliset rajapintaprototyypit.

6.2.1 Haastattelu 2

Haastattelussa 2 selvitettiin kontekstuaalisen haastattelun avulla, miten täsmäytystä tehdään Kansalliskirjastossa OpenRefine-työkalun avulla. Haastateltavana toimi Kansalliskirjaston sanastokehittäjä (kehittäjä B). Haastattelun aikana haastateltavaa pyydettiin näyttämään käyttämänsä työnkulut

OpenRefine-ympäristössä sekä selittämään suullisesti, miten hän käyttää sitä YSO:n käsitteiden täsmäyttämiseen Wikidata-palveluun. Haastattelussa hyödynnettiin lisäksi ennalta laadittuja haastattelukysymyksiä. Haastateltavan kommenteista ja vastauksista laadittiin muistiinpanot haastattelun aikana, jotka kirjoitettiin puhtaaksi haastattelun jälkeen. Haastatteluun 2 laaditut kysymykset löytyvät liitteestä B.

Haastattelussa selvisi, että linkattavista YSO:n käsitteistä tuodaan OpenRefineen erikseen monikko- ja yksikkömuodot sekä erikieliset versiot omiin sarakkeisiinsa. Kehittäjä tekee täsmäytyksen jokaiselle sarakkeelle erikseen ja niiden tulosten perusteella hän päättää mihin Wikidatan käsitteeseen luodaan linkki kyseiselle YSO:n käsitteelle. Kehittäjä tekee päätöksen linkitettävästä käsitteestä ensin tarkastelemalla kandidaattitermiä sekä esikatselun kuvaustekstiä. Jos ne vaikuttavat vastaavan YSO:n käsitettä, kehittäjä avaa kandidaatin Wikidata-sivun sekä mahdollisesti vastaavan Wikipedia-sivun. Näiden sivujen tietojen perusteella hän tekee lopullisen päätöksen linkitettävästä käsitteestä.

Haastateltava piti rajapinnan palauttamia pisteitä epäselvinä, minkä vuoksi hän ei ole käyttänyt niitä hyväksi linkityspäätöksen teossa. Toisena puutteena Wikidatan täsmäytysrajapinnassa hän mainitsi sen, että rajapinta ei mahdollista tiettyjen käsitetyyppien poissulkemista täsmäytystuloksista. Vaikeuksia tuotti myös se, että monikkomuotoisille termeille sekä termeille, jotka sisältävät erikoismerkkejä, ei usein täsmäytyshaulla löydy vastaavuuksia Wikidatasta, vaikka ne olisivatkin olemassa.

6.2.2 Haastattelu 3

Haastattelussa 3 pyrittiin arvioimaan täsmäytysrajapintatoteutusten ensimmäisten prototyyppien toimivuutta sekä niiden puutteita puolistrukturoidun haastattelun avulla. Kahdelle sanastokehittäjälle (kehittäjä A ja kehittäjä B) esiteltiin rajapintojen toiminnallisuutta käyttäen esimerkkidatana museokokoelmien kuvailussa käytettyä Siida-museon sanalistaa, joka ei sisällä linkitetyn datan tunnisteita. Se sisälsi sekä suomen- että pohjoissaamenkielisiä termejä. Sanalista tuotiin OpenRefine-alustalle taulukoksi ja täsmäytys tehtiin YSO-sanastoon molempien rajapintojen avulla. Haastateltavia pyydettiin kommentoimaan rajapintoja haastattelun aikana ja heiltä kysyttiin ennalta laadittuja kysymyksiä niiden toiminnallisuudesta. Haastateltaville demonstroitiin rajapintojen käyttöä OpenRefine-ympäristössä, mutta he eivät itse tehneet täsmäytystä. Kommenteista ja vastauksista otettiin haastattelun aikana muistiinpanot, jotka kirjoitettiin puhtaaksi haastattelun jälkeen. Haastattelun 3 kysymykset löytyvät liitteestä B.

Haastattelussa kävi ilmi, että molemmista rajapinnoista puuttui useita olennaisia ominaisuuksia. Haastateltavat toivoivat molempiin rajapintoihin mahdollisuutta hakea käsitteitä yksitellen sanastosta sekä mahdollisuutta esikatsella käsitteiden ominaisuuksia (esimerkiksi muun kielisiä termejä).

He olivat myös sitä mieltä, että olisi hyödyllistä saada käsitteiden URI-tunnisteet erikseen esille jollain tapaa. Lisäksi Skosmoksen rajapinnalle toivottiin mahdollisuutta valita kieli, jolla täsmäytys tehdään. Molempien rajapintojen palauttamissa tuloksissa ilmeni myös puutteita. Annifin rajapinta ei löytänyt kaikille käsitteille vastaavuuksia, vaikka ne olisivatkin olleet olemassa YSO:ssa, eikä Skosmoksen rajapinta kyennyt löytämään vastaavia käsitteitä erikoismerkkejä sisältäville termeille. Toisaalta, toisin kuin Haastattelussa 2 testattu Wikidatan rajapinta, Skosmoksen rajapinta kykeni löytämään sekä monikko- että yksikkömuodossa esiintyviä termejä.

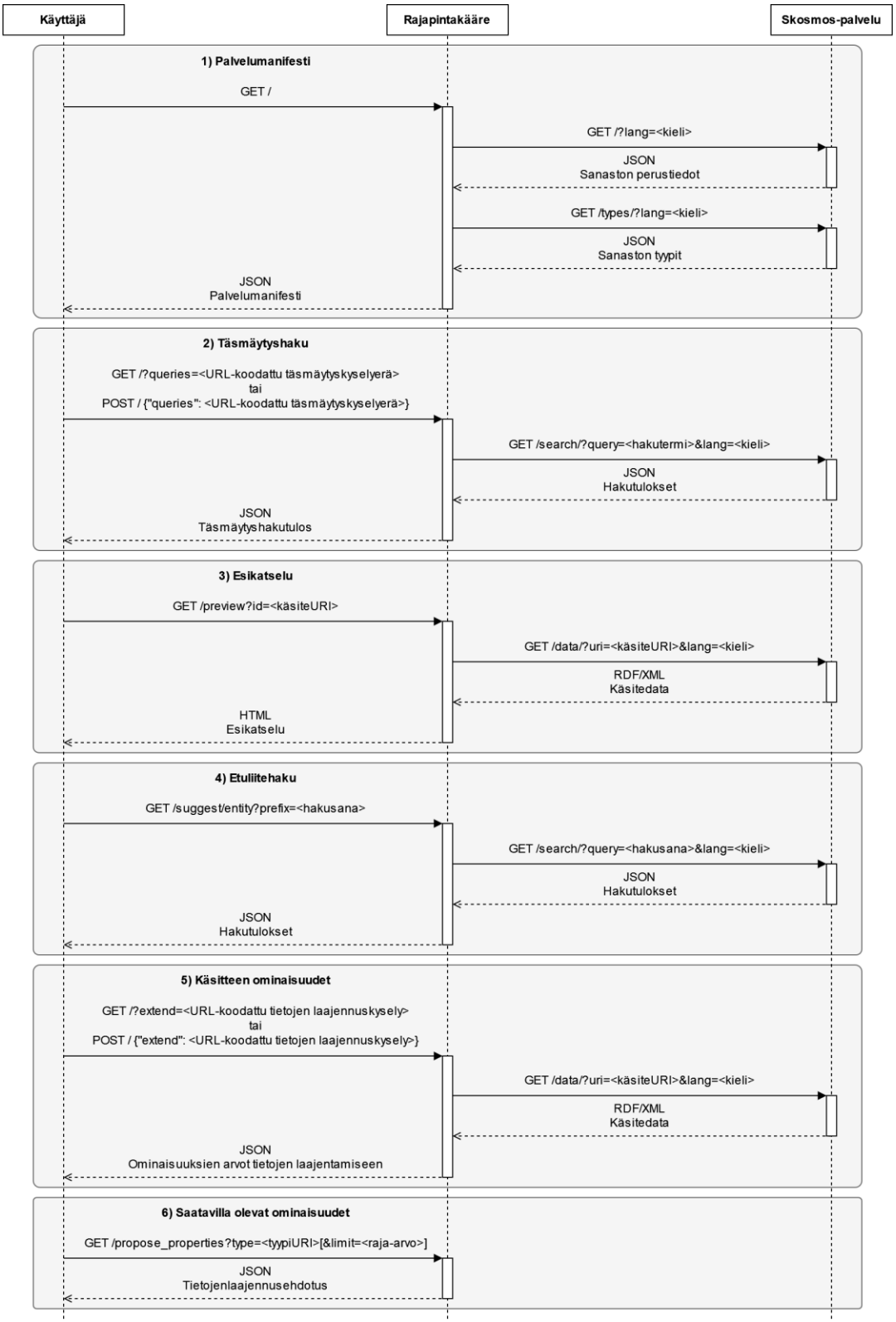
Haastattelujen 2 ja 3 perusteella laajennettiin vaatimuksia rajapintaprototyypeille. Mahdollisuus esikatsella käsitteen tietoja ja hakea käsitteitä suoraan sanastosta siirtymättä OpenRefine-ympäristön ulkopuolelle helpottaa linkitystyön tekemistä, joten niitä tukeva toiminnallisuus asetettiin vaatimukseksi seuraavan iteraation prototyypeille. Myös toiminnallisuus käsitteiden URI-tunnisteiden sekä muiden ominaisuuksien hakemiseen asetettiin vaatimukseksi. Skosmos-prototyypiltä vaadittiin myös mahdollisuutta kielen valintaan sekä parannuksia erikoismerkkejä sisältävien termien haun osalta. Skosmoksen rajapinnalta ei vaadittu erillistä pisteytystä, koska Haastattelussa 2 ilmeni, että haastateltava ei ollut kokenut sitä hyödylliseksi Wikidatan rajapinnassa.

6.2.3 Skosmos-rajapintaprototyyppi 2

Skosmos-prototyypin lopullinen versio laajensi prototyypin 1 rajapintaa mahdollistaen kielen valinnan ja parantaen täsmäytyshaun tuloksia erikoismerkkejä sisältävien hakujen osalta. Rajapintaan lisättiin myös käsitteiden esikatselu, etuliitehaku sekä ominaisuuksien ja niiden arvojen haku. Prototyypin palvelumanifestiin lisättiin tiedot uusille ominaisuuksille. Esimerkki siitä löytyy liitteestä A. Lopullinen rajapintaprototyyppi sisältää siis kuusi toimintoa:

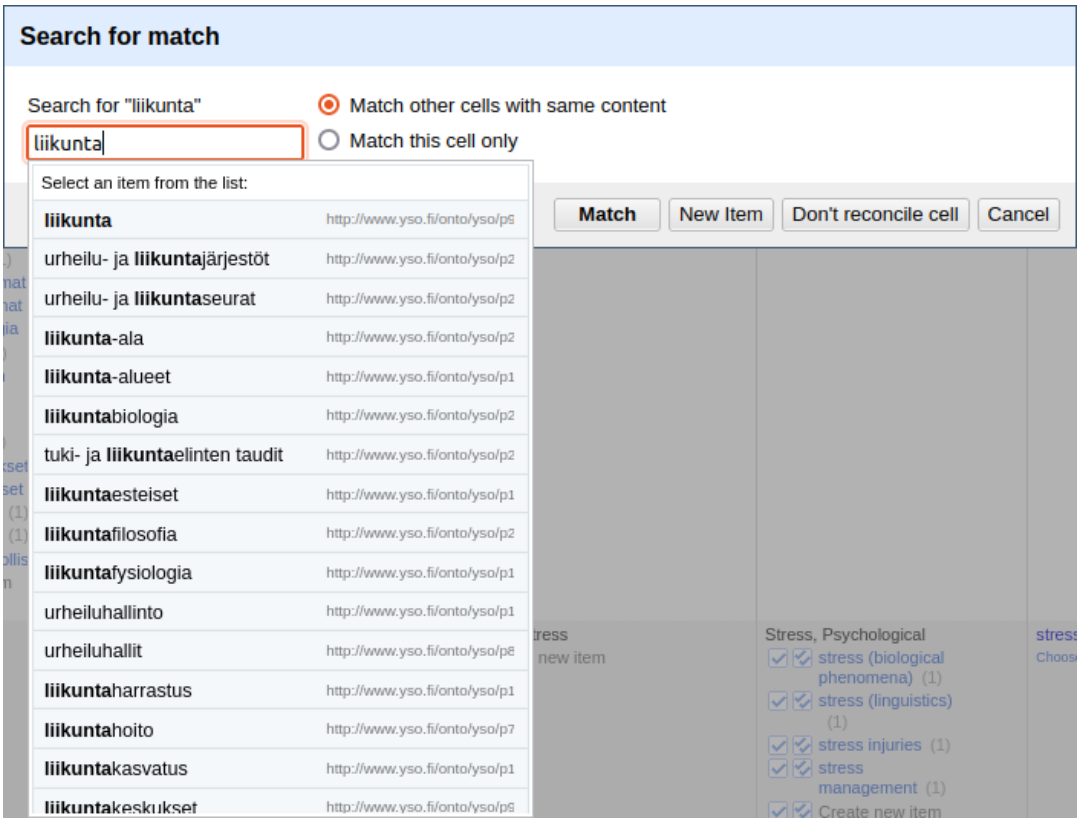
1) palvelumanifestin palauttaminen
2) täsmäytyshaku
3) käsitteiden esikatselu
4) etuliitehaku
5) käsitteen ominaisuuksien arvojen palauttaminen
6) järjestelmän kautta saatavilla olevien ominaisuuksien palauttaminen

Ne on esitetty kuvassa 6.4 sekvenssikaavion muodossa. Rajapintaprototyyppiin tehdyt muutokset päivitettiin Kansalliskirjaston palvelimelle lopullista arviointia varten.



Kuva 6.4. Sekvenssikaavio lopullisen Skosmos-prototyypin toiminnallisuuksista.

Kielirajaus muuttaa rajapinnan täsmäytyspalvelupisteen verkko-osoitteen muotoon ”/<sanastoID>/<kieli>/reconcile”. Kieli siis valitaan linkitettävien termien kielen mukaan ja hakutulokset palautetaan samalla kielellä. Kielivalinnan lisäksi täsmäytyshakua parannettiin sellaisten termien osalta, jotka sisältävät tiettyjä erikoismerkkejä. Erikoismerkilliset haut jaetaan osiin, joille tehdään erilliset pyynnöt Skosmos-palvelun rajapintaan. Yksi haku tehdään aina koko termillä ja yksi haku ensimmäistä erikoismerkkiä edeltävällä osalla. Jos termi sisältää sulkutarkenteen, tehdään lisäksi haku sulkujen sisäisellä osalla. Esimerkiksi termille ”adressit (vetoomukset)” tehtäisiin siis kolme pyyntöä Skosmos-palvelun rajapintaan, merkkijonoilla ”adressit (vetoomukset)”, ”adressit” ja ”vetoomukset”.



Kuva 6.5. Lopullisen Skosmos-prototyypin etuliitehaun tulokset haulle ”liikunta”[28] OpenRefine-ympäristössä.

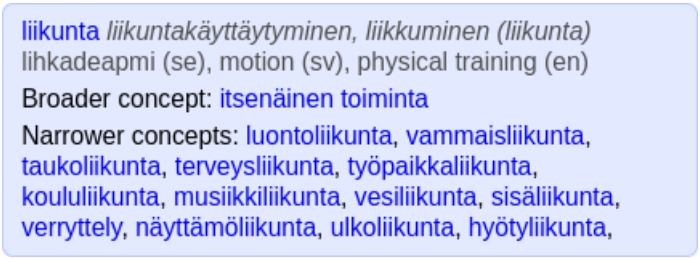
Rajapintaan lisätyn etuliitehaun tarkoituksena on mahdollistaa käsitteiden hakeminen OpenRefine-ympäristön sisällä. Täsmäytysrajapintamäärittelmän mukaan etuliitehaku voidaan tarjota entiteeteille, tyypeille ja ominaisuuksille, mutta Skosmoksen rajapinta toteuttaa näistä vain ensimmäisen. Entiteettejä voidaan hakea verkko-osoitteella ”/<sanastoID>/<kieli>/reconcile/suggest/entity”. Se käyttää Skosmos-palvelun hakumetodia käsitteiden

[28] http://www.yso.fi/onto/yso/p916

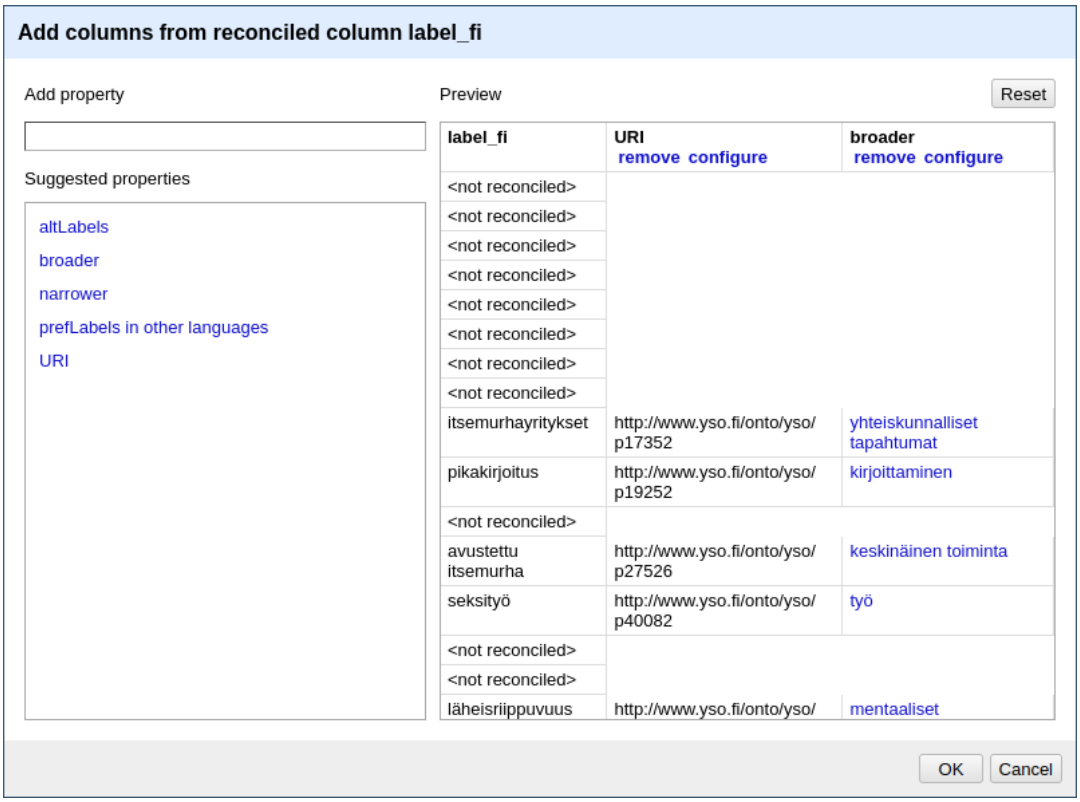
hakemiseen ja sen tulokset ovat identtiset täsmäytyshaun tuloksien kanssa. Kuvassa 6.5 on esitetty esimerkki etuliitehaun tuloksista OpenRefine-ympäristössä.

Esikatselu mahdollistaa täsmennyshaun tuloksena saatujen käsitteiden esikatselun. Rajapinta vastaa verkko-osoitteeseen ”/<sanastoID>/<kieli>/reconcile/preview” saapuneisiin kutsuihin HTML-tiedostolla, joka sisältää käsitteen ensisijaisen termin ja vaihtoehtoiset termit (skos:prefLabel ja skos:altLabel) haetulla kielellä sekä ensisijaiset termit muilla saatavilla olevilla kielillä. Lisäksi näytetään käsitteen määritelmä (skos:definition) sekä sen ylä- ja alakäsitteet (skos:broader ja skos:narrower). Käsitteen data haetaan Skosmos-palvelun rajapinnasta RDF/XML-tiedostona, josta esikatselun tiedot poimitaan SPARQL-kyselyiden avulla. Kuvassa 6.6 on esitetty esimerkki esikatselunäkymästä OpenRefine-ympäristössä.

Ominaisuuksien arvojen hakutoiminnallisuus mahdollistaa täsmäytettyjen käsitteiden rikastamisen kohdesanaston datalla. Rajapintaprototyyppi tarjoaa viisi ominaisuutta, joiden arvot käsitteelle voidaan hakea: URI, ensisijainen termi muilla kielillä, vaihtoehtoiset termit haetulla kielellä sekä ylä- ja alakäsitteet. Saatavilla olevat ominaisuudet voidaan hakea verkko-osoitteesta ”/<sanastoID>/<kieli>/reconcile/propose_properties”. Ominaisuuksien arvot haetaan Skosmos-palvelun rajapinnasta RDF/XML-tiedostona samalla tavalla kuin esikatselutoiminnallisuudessa. Kuvassa 6.7 on esitetty esimerkki ominaisuuksien arvojen haun palauttamista arvoista OpenRefine-ympäristössä.



Kuva 6.6. Lopullisen Skosmos-prototyypin esikatselu käsitteelle ”liikunta” OpenRefine-ympäristössä. Ensimmäisellä rivillä nähdään käsitteen ensisijainen termi sekä vaihtoehtoiset termit. Seuraavalla rivillä sen ensisijaiset termit muilla kielillä. Käsitteestä esitetään myös sen ylä- ja alakäsitteet.



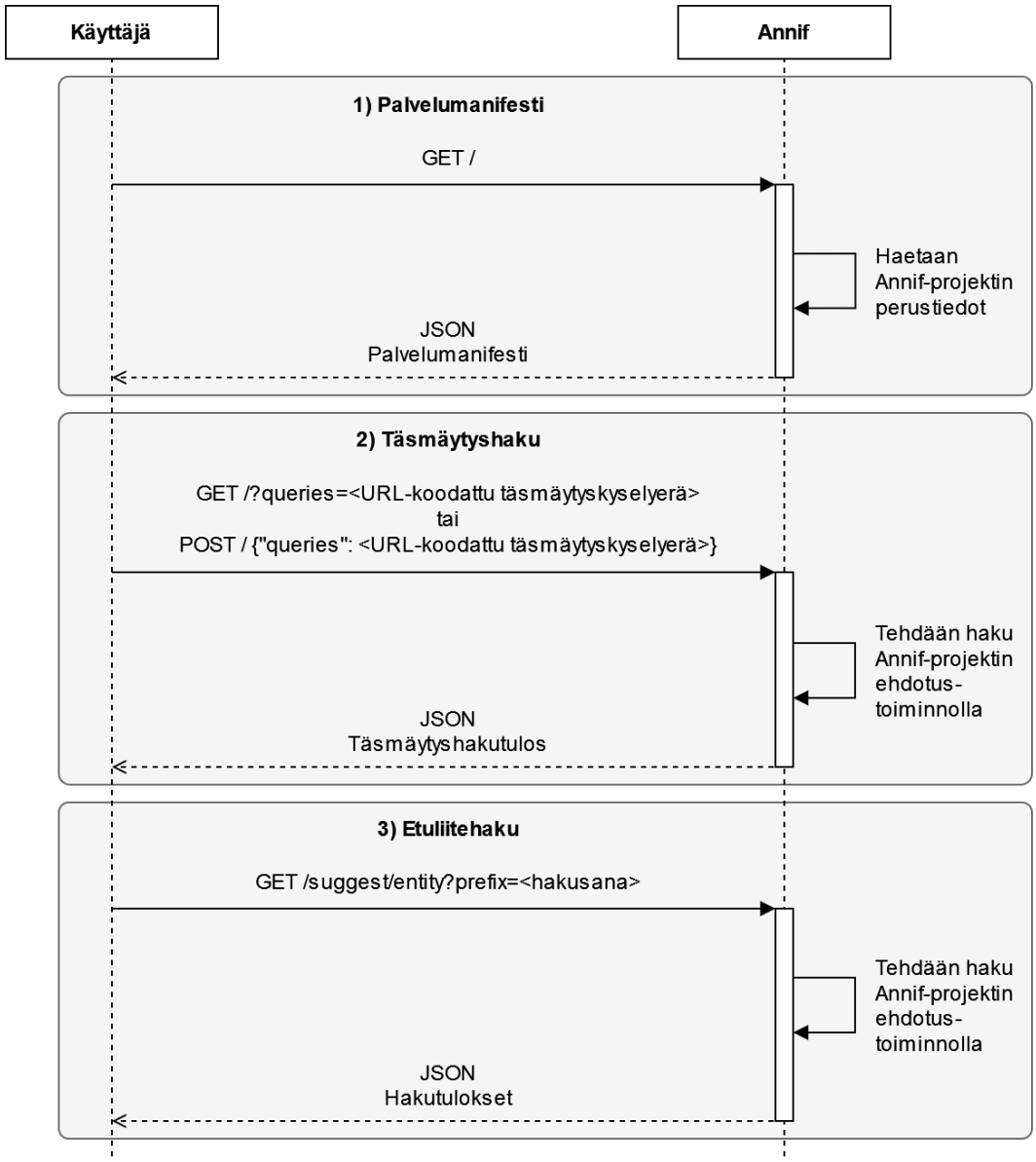
Kuva 6.7. Lopullisen Skosmos-prototyypin ominaisuuksien arvojen haun palauttamat arvot ominaisuuksille ”URI” ja ”broader” OpenRefine-ympäristössä. Kuvan vasemmalla puolella nähdään saatavilla olevat ominaisuudet.

6.2.4 Annif-rajapintaprototyyppi 2

Annifin rajapintaa ei päivitetty juurikaan lopullisessa prototyypissä. Siihen lisättiin etuliitehaku eli se sisältää kolme toimintoa:

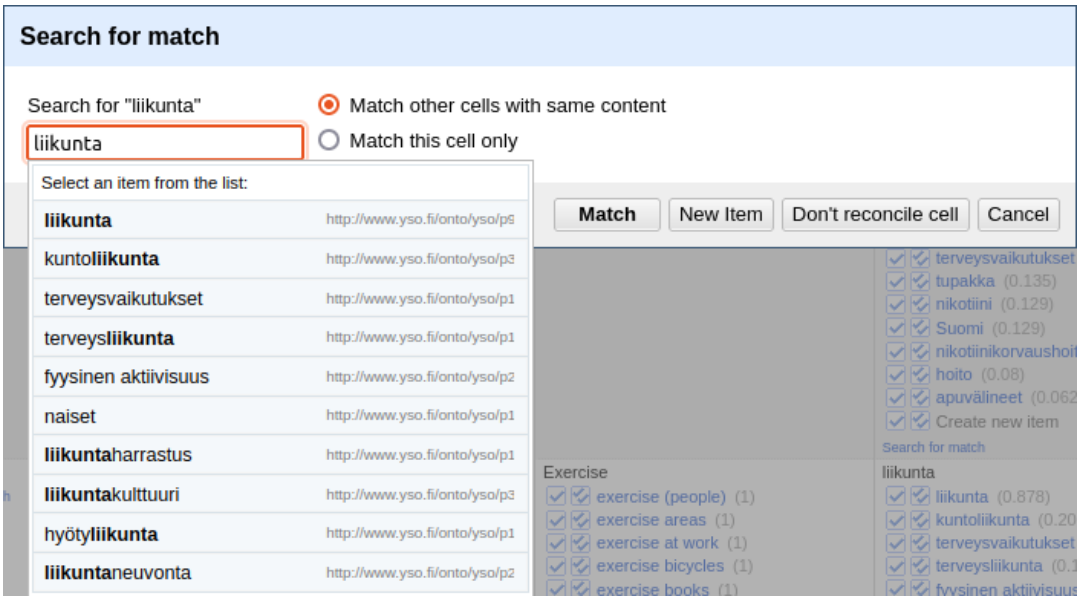
1) palvelumanifestin palauttaminen
2) täsmäytyshaku
3) etuliitehaku

Nämä on esitelty kuvassa 6.8 sekvenssikaaviossa. Prototyypin palvelumanifesti päivitettiin lisäämällä siihen tiedot etuliitehausta. Esimerkki siitä löytyy liitteestä A. Tietueiden esikatselua tai ominaisuuksien arvojen hakua ei toteutettu, vaikka niiden lisääminen olisi periaatteessa ollut mahdollista. Annif tarjoaa mahdollisuuden tallentaa käytetyn sanaston data RDF-muotoisena tiedostona, jota olisi voinut käsitellä suoraan SPARQL-kyselyillä kuten Skosmos-rajapinnan tapauksessa. Käytännössä sanaston lataamiseen tiedostosta kului kuitenkin liikaa aikaa (pahimmillaan yli 2 minuuttia sanastoa kohden), joten nämä toiminnot eivät olleet toteuttamiskelpoisia tässä muodossa. Rajapintaprototyyppiin tehdyt muutokset päivitettiin Docker-kuvaan, joka vietiin Kansalliskirjaston OpenShift-konttialustalle lopullista arviointia varten.



Kuva 6.8. Sekvenssikaavio lopullisen Annif-prototyypin toiminnallisuuksista.

Etuliitehaku toteutettiin samalla tavalla kuin Skosmos-rajapinnassa käyttämällä hyväksi täsmäytyshakumetodia. Se ei siis hae käsitteitä merkkijonojen alun perusteella kuten rajapintamääritelmän mukaan olisi tarkoitus. Tämä johtuu samasta ongelmasta sanaston lataamisessa kuin muissa toiminnallisuuksissa. Myös Annifin etuliitehausta toteutettiin ainoastaan entiteettihaku, joka on saatavilla verkko-osoitteessa ”/projects/<projektiID>/reconcile/suggest/entity”. Muita etuliitehaun toiminnallisuuksia ei toteutettu, koska Annif ei tarjoa käsitteille tyyppejä tai ominaisuuksia, joita hakea. Esimerkki etuliitehaun tuloksesta OpenRefine-ympäristössä on esitetty kuvassa 6.9.



Kuva 6.9. Lopullisen Annif-prototyypin etuliitehaun tulokset haulle ”liikunta” OpenRefine-ympäristössä.

6.3 Lopullinen arviointi

Prototyyppien valmistumisen jälkeen ne arvioitiin vielä viimeisten kontekstuaalisten haastattelujen avulla. Niitä on käsitelty alaluvussa 6.3.1. Lisäksi haastatteluista saatua dataa on analysoitu prototyyppien arvioimiseksi alaluvussa 6.3.2.

6.3.1 Haastattelut 4A ja 4B

Haastattelut 4A ja 4B toteutettiin kontekstuaalisina haastatteluina, jotka koostuivat tehtävä- ja kysymysosuuksista. Haastateltavina toimivat samat Kansalliskirjaston sanastokehittäjät (kehittäjä A ja kehittäjä B) kuin haastattelussa 3, mutta tällä kertaa haastattelut toteutettiin kummallekin kehittäjälle erikseen. Tehtäväosiosta tallennettiin ruudunkaappausohjelmalla video ja koko haastattelusta nauhoitettiin äänitallenne. Haastattelun aikana otettiin myös muistiinpanoja, jotka kirjoitettiin puhtaaksi haastattelun jälkeen. Haastattelun 4 tehtäväosion tehtävät sekä haastattelukysymykset löytyvät liitteestä B.

Tehtäväosiossa sanastokehittäjälle annettiin testiaineisto kahdessa taulukkotiedostossa. Ensimmäisenä aineistona käytettiin MeSH-asiasanaston versiota, joka sisältää englanninkielisten termien lisäksi myös suomen- ja ruotsinkielisiä lääketieteen termejä. MeSH:stä haettiin ”käyttäytyminen”-käsitteen[29] alakäsitteet (379 kappaletta) Finto-palvelun SPARQL-rajapinnan avulla, josta valittiin satunnaisesti 25 käsitettä. Toisena aineistona käytettiin

[29] http://www.yso.fi/onto/mesh/D001519

samaa Siida-museon sanalistaa kuin haastattelussa 3 ja myös siitä valittiin satunnaisesti 25 käsitettä. MeSH-aineisto koostui suomen-, ruotsin- ja englanninkielisistä termeistä, Siida-aineisto puolestaan suomen- ja pohjoissaamenkielisitä termeistä. Aineistot valittiin vastaamaan haastattelussa 1 ilmenneitä potentiaalisia käyttötapauksia ja linkitysympäristö pyrittiin luomaan haastattelua 2 vastaavaksi. Käsitteiden lukumäärä valittiin suhteellisen alhaiseksi, jotta testaustilanteen kesto pysyisi kohtuullisena. Molemmat aineistot tuotiin OpenRefine-työkaluun taulukoiksi ja sanastokehittäjiä pyydettiin tekemään niille linkitys kuten he sen normaalioloissa tekisivät käyttäen sekä Skosmoksen että Annifin rajapintoja. Kaikille kielille tehtiin täsmäytyshaut molemmilla rajapinnoilla ja linkitykset tehtiin molemmille testiaineistoille suomenkielisiin YSO:n käsitteisiin. Kehittäjiä pyydettiin lisäksi testaamaan rajapintojen esikatselutoiminnallisuutta, etuliitehakua sekä ominaisuuksien arvojen hakua. Lopuksi syntyneet taulukot linkityksineen tallennettiin myöhempää laadullista analyysiä varten. Tehtävien tekoon kulunut aika mitattiin ja mahdolliset virhetilanteet kirjattiin ylös. Tehtäväosion tavoitteena oli kerätä tietoa rajapintojen käytettävyydestä ja suorituskyvystä.

Tehtäväosion jälkeen suoritettiin lyhyt puolistrukturoitu haastattelu, jossa kehittäjää pyydettiin arvioimaan rajapintojen käytettävyyttä ja suorituskykyä ja vertaamaan niitä toisiinsa ennalta laadittujen kysymysten avulla. Haastatteluosion tavoitteena oli syventää tehtäväosiossa saatua tietoa rajapintojen käytettävyydestä.

Molemmat haastateltavat olivat sitä mieltä, että Skosmoksen rajapinnan palauttamat ehdotukset olivat hyödyllisiä ja etenkin erikoismerkkejä sisältäville käsitteille vastaavuuksien löytyminen mainittiin hyödyllisenä ominaisuutena. Haastateltavien mukaan Skosmoksen rajapinnan esikatselu- ja etuliitehakutoiminnallisuus toimivat odotetusti ja olivat avuksi linkitysten tekemisessä. Kehittäjän A mukaan molemmat ominaisuudet vähensivät OpenRefine-projektista poistumista ja siten suoraviivaistivat linkitysprosessia. Hän ei lisäisi muita kenttiä esikatselunäkymään, mutta kehittäjä B ottaisi mukaan myös käsitteen temaattisen ryhmän (skos:collection) sekä sen linkityssuhteet muihin sanastoihin (esim. skos:closeMatch Wikidataan). Molempien kehittäjien mukaan Skosmoksen toiminnallisuus ominaisuuksien arvojen hakuun ei ole välttämätön linkitysten tekemisessä, mutta sen tarjoamista käsitteiden tunnisteista voi olla hyötyä, kun linkitykset viedään johonkin muuhun ympäristöön.

Vaikka Annifin rajapinta tarjoaa molempien haastateltavien mukaan semanttisesti monipuolisempia tuloksia kuin Skosmoksen rajapinta, olivat ne usein epäoleellisia ja niiden paljous osittain hankaloitti oikean käsitteen löytämistä. Molemmat kehittäjät käyttivätkin enemmän aikaa Annifin täsmäytyshaun tulosten läpikäymiseen tehtäväosion aikana. Kehittäjän A mukaan Annifin etuliitehaku oli periaatteessa kiinnostava ja tarjosi myös sellaisia käsitteitä, jotka eivät olisi heti tulleet mieleen, mutta molemmat haastateltavat olivat sitä mieltä, että sen avulla oli ajoittain vaikeaa löytää oikeaa käsitettä.

Tehtäväosiossa sattuikin tilanteita, joissa kohdesanastossa olisi ollut haluttu käsite, mutta rajapinta ei tarjonnut sitä, jolloin käsitettä ei voitu linkittää lainkaan. Kehittäjä A toivoi myös, että Annifin rajapinnassa olisi käsitteiden esikatselu, vaikka se ei sisältäisikään tietoja käsitteestä, koska sen puute voi saada käyttäjän epäilemään rajapinnan toiminnallisuutta.

Haastatteluosuuden lopuksi kehittäjiltä kysyttiin vielä, käyttäisivätkö he toteutettuja järjestelmiä työssään. Kumpikin haastateltava oli sitä mieltä, että Skosmos-prototyyppiä voitaisiin käyttää tulevaisuudessa, mutta Annif-prototyyppi vaatisi heidän mielestään vielä kehitystä ennen sen käyttöönottoa. Kehittäjän A mukaan rajapinnat ovat hyödyllisiä etenkin aiemmin linkittämättömien sanalistojen täsmäytyksessä ja erityisesti täsmäytysprosessin alussa parhaiden kandidaattien löytämisessä. Täsmäytystyötä pitäisi kuitenkin jatkaa myös muussa ympäristössä, sillä rajapinnat eivät aina löydä kaikkia vastaavuuksia.

6.3.2 Haastattelujen 4A ja 4B tulosten arviointi

Haastatteluissa 4A ja 4B syntyneet linkitykset tallennettiin haastattelujen jälkeen ja niistä luotiin analysoitavat taulukot. Taulukossa 6.1 on esitetty MeSH-sanaston linkitys ja taulukossa 6.2 Siida-sanalistan linkitys YSO:on. Taulukoissa on esitetty alkuperäiset linkitettävät MeSH:n tai Siidan termit ja haastateltavien tekemät linkitykset Skosmoksen ja Annifin rajapinnoilla. Taulukoihin on merkitty vihreällä ne linkitykset, jotka on tehty oikeaan YSO:n käsitteeseen tai jätetty linkittämättä, kun YSO:ssa ei ole vastaavaa käsitettä. Punaisella sekä X-symbolilla on puolestaan merkitty ne linkitykset, jotka eivät ole kohdistuneet oikeaan YSO:n käsitteeseen. Epäselvät kohdat on merkitty keltaisella ja kysymysmerkillä. Taulukoissa on lisäksi mukana sarake työn tekijän kommenteille.

Taulukoista nähdään, että Skosmoksen rajapinnalla on onnistuttu löytämään lähestulkoon kaikki YSO:n käsitteet, joilla on vastaavuus testiaineistoissa. MeSH:n aineistossa oli muutamia epäselvyyksiä, esimerkiksi Kehittäjä B on linkannut MeSH:n termin ”henkinen rasitus”[30] YSO:n termiin ”stressi”[31], kun taas kehittäjä A on jättänyt tämän linkittämättä. Onkin epäselvää vastaavatko nämä täysin toisiaan ja lopullinen päätös riippuisi sovellettavasta linkitysstrategiasta. Siidan aineistossa kehittäjä A on tehnyt oikeat linkitykset kaikille termeille muutamaa epäselvää tapausta lukuun ottamatta. Kehittäjä B ei ole tehnyt linkityksiä neljälle termille, joilla olisi ollut vastaavat YSO:n käsitteet. Kaksi näistä on luultavasti jäänyt tekemättä epähuomiossa. Kahden muun termin linkittämiseen kehittäjä A oli käyttänyt Skosmos-rajapinnan etuliitehakua, mutta kehittäjä B ei ollut tehnyt näin.

[30] http://www.yso.fi/onto/mesh/D013315
[31] http://www.yso.fi/onto/yso/p133

Virheet selittyvät ainakin osittain sillä, että haastattelussa 4B ei jäänyt aikaa Siidan linkityksen tekemiseen huolellisesti.

Annifin rajapinnalla pystyttiin tekemään miltei kaikki samat linkitykset kuin Skosmoksen rajapinnalla. Sekä MeSH:n että Siidan aineistoissa oli kuitenkin yksi termi, jota Annifin täsmäytyshaku ei kyennyt löytämään. Nämä olivat MeSH:n ”seksityö”[32] sekä Siidan ”oven, astian kahva”, jonka kehittäjä A linkitti Skosmoksen rajapinnalla YSO:n ”kahvat”-käsitteeseen[33]. Kumpaakaan näistä ei kyetty löytämään myöskään etuliitehaun avulla, sillä se käyttää samaa Annifin sisäistä hakumenetelmää kuin tavanomainen täsmäytyshaku. Tämä vaikeuttaa käsitteiden hakemista niissä tilanteissa, joissa täsmäytyshaku ei ole löytänyt oikeaa käsitettä. Toisaalta Annifin rajapinta löytää myös sellaisia käsitteitä, joita Skosmoksen rajapinta ei löydä, esimerkiksi Siidan termi ”marjan poiminta” ei löytynyt Skosmoksen täsmäytyshaulla, mutta Annifin haku tarjosi sen ensimmäisenä ehdotuksena.

Rajapintojen tuottamien tulosten samankaltaisuus selittyy ainakin osittain sillä, että haastateltavat näkivät molempien rajapintojen täsmäytyshakujen tulokset rinnakkain. He pystyivät siis käyttämään toisen rajapinnan tuloksia avuksi linkittämisessä. Voidaan kuitenkin todeta, että Skosmoksen rajapinnan avulla halutut käsitteet löytyvät luotettavammin. Vaikka täsmäytyshaku ei löytäisi oikeaa käsitettä, se voidaan kuitenkin linkittää etuliitehaun avulla, mikä Annifin rajapinnalla ei ole aina mahdollista. Toisaalta Annifin rajapinta tuottaa monipuolisempia tuloksia ja sen avulla on mahdollista löytää käsitteitä semanttisen vastaavuuden eikä vain merkkijonojen samankaltaisuuden perusteella. Esimerkiksi MeSH:n käsitteelle ”hölkkääminen”[34] Annifin haku on löytänyt vastineeksi YSO:n käsitteen ”juoksu”[35], josta kehittäjä B on tehnyt linkityksen.

Taulukko 6.1. Haastattelujen 4A ja 4B MeSH-linkityksen tulokset. Tunnisteiden ”mesh:”- ja ”yso:”-etuliitteet vastaavat verkko-osoitteita ”http://www.yso.fi/onto/mesh/” ja ”http://www.yso.fi/onto/yso/”

MeSH-käsite	Kehittäjän A Skosmos-linkitykset	Kehittäjän B Skosmos-linkitykset	Kehittäjän A Annif-linkitykset	Kehittäjän B Annif-linkitykset	Kommentti
jäähdyttelyliikunta (mesh:D064590)					
liikenneraivo (mesh:D000077315)					
sosiaalinen verkostoituminen (mesh:D060756)	?	?	?	?	Verkostoituminen (yso:p20000) YSO:ssa

[32] http://www.yso.fi/onto/mesh/D011477
[33] https://finto.fi/yso/fi/page/p39278
[34] http://www.yso.fi/onto/mesh/D007590
[35] http://www.yso.fi/onto/yso/p9087

MeSH-käsite	Kehittäjän A Skosmos-linkitykset	Kehittäjän B Skosmos-linkitykset	Kehittäjän A Annif-linkitykset	Kehittäjän B Annif-linkitykset	Kommentti
spatiaalinen prosessointi (mesh:D065855)					
vesipiipun polttaminen (mesh:D000073867)					
pakkotoiminta (mesh:D003192)					
hölkkääminen (mesh:D007590)	?	?	?	juoksu (yso:p9087) ?	Juokseminen (mesh:D012420) on hölkkäämisen (mesh:D007590) yläkäsite
sosiaalinen deprivaatio (mesh:D000091489)					
itsemurhayritys (mesh:D013406)	itsemurhayritykset (yso:p17352)	itsemurhayritykset (yso:p17352)	itsemurhayritykset (yso:p17352)	itsemurhayritykset (yso:p17352)	
pikakirjoitus (mesh:D012781)	pikakirjoitus (yso:p19252)	pikakirjoitus (yso:p19252)	pikakirjoitus (yso:p19252)	pikakirjoitus (yso:p19252)	
kielen liiketavat (mesh:D014061)					
avustettu itsemurha (mesh:D017236)	avustettu itsemurha (yso:p27526)	avustettu itsemurha (yso:p27526)	avustettu itsemurha (yso:p27526)	avustettu itsemurha (yso:p27526)	
seksityö (mesh:D011477)	seksityö (yso:p40082)	seksityö (yso:p40082)	✕	✕	Annifin rajapinta ei löytänyt haettua termiä täsmäytyshaulla eikä etuliitehaulla
itsensäsilpominen (mesh:D012652)					
addiktio ruokaan (mesh:D000073932)					
läheisriippuvuus (mesh:D017004)	läheisriippuvuus (yso:p17875)	läheisriippuvuus (yso:p17875)	läheisriippuvuus (yso:p17875)	läheisriippuvuus (yso:p17875)	
sähkösavukkeiden höyryjen hengittäminen (mesh:D000072137)					
liikunta (mesh:D015444)	liikunta (yso:p916)	liikunta (yso:p916)	liikunta (yso:p916)	liikunta (yso:p916)	
henkinen rasitus (mesh:D013315)	?	stressi (yso:p133) ?	?	stressi (yso:p133) ?	YSO:n stressi (yso:p133) viittaa biologiseen stressiin, mikä mahdollisesti vastaisi MeSH:n fysiologista stressiä (mesh:D013312)
alkoholiabstinenssi (mesh:D064829)					Päihteettömyys (yso:p24215) YSO:ssa,

MeSH-käsite	Kehittäjän A Skosmos-linkitykset	Kehittäjän B Skosmos-linkitykset	Kehittäjän A Annif-linkitykset	Kehittäjän B Annif-linkitykset	Kommentti
					MeSH:ssä ei vastaavaa
informaatiolukutaito (mesh:D058980)	informaatiolukutaito (yso:p15108)	informaatiolukutaito (yso:p15108)	informaatiolukutaito (yso:p15108)	informaatiolukutaito (yso:p15108)	
tutkimusraportti (mesh:D058028)	tutkimusraportit (yso:p236)	tutkimusraportit (yso:p236)	tutkimusraportit (yso:p236)	tutkimusraportit (yso:p236)	
potilastyytyväisyys (mesh:D017060)					
tupakointi (mesh:D000073865)	tupakointi (yso:p10017)	tupakointi (yso:p10017)	tupakointi (yso:p10017)	tupakointi (yso:p10017)	
kestävyysharjoittelu (mesh:D000076663)	kestävyysharjoittelu (yso:p7676)	kestävyysharjoittelu (yso:p7676)	kestävyysharjoittelu (yso:p7676)	kestävyysharjoittelu (yso:p7676)	

Taulukko 6.2. Haastattelujen 4A ja 4B Siida-linkityksen tulokset. Tunnisteiden ”yso:”-etuliite vastaa verkko-osoitetta ”http://www.yso.fi/onto/yso/”

Siidan termi	Kehittäjän A Skosmos-linkitykset	Kehittäjän B Skosmos-linkitykset	Kehittäjän A Annif-linkitykset	Kehittäjän B Annif-linkitykset	Kommentti
marjan poiminta	marjanpoiminta (yso:p2898)	✕	marjanpoiminta (yso:p2898)	marjanpoiminta (yso:p2898)	Skosmoksen täsmäytyshaku ei löytänyt oikeaa termiä, Kehittäjä A käytti etuliitehakua sen löytämiseen
mekko, mekot	mekot (yso:p20766)	mekot (yso:p20766)	mekot (yso:p20766)	mekot (yso:p20766)	
keittokoukku					
metsästys	metsästys (yso:p3547)	metsästys (yso:p3547)	metsästys (yso:p3547)	metsästys (yso:p3547)	
maitosiivilä	siivilät (yso:p16565) ?	?	siivilät (yso:p16565) ?	?	Saamenkielinen termi ”silli” vastaa suomen siivilää
laukunleuka, -leuat					
hylly, hyllyt	hyllyt (yso:p9126)	hyllyt (yso:p9126)	hyllyt (yso:p9126)	hyllyt (yso:p9126)	
tupsu, tupsut	tupsut (yso:p27735)	tupsut (yso:p27735)	tupsut (yso:p27735)	tupsut (yso:p27735)	
messinkikattila					
jiekiö, jiekiöt					

Siidan termi	Kehittäjän A Skosmos-linkitykset	Kehittäjän B Skosmos-linkitykset	Kehittäjän A Annif-linkitykset	Kehittäjän B Annif-linkitykset	Kommentti
juurköysi, köydet	?	?	?	?	YSO:ssa on käsite köydet (yos:p2905), mutta on epäselvää ovatko nämä vastaavia
huivi, huivit	huivit (yso:p10457)	huivit (yso:p10457)	huivit (yso:p10457)	huivit (yso:p10457)	
sanko, sangot	sangot (yso:p21702)	✕	sangot (yso:p21702)	sangot (yso:p21702)	Skosmos-linkki on luultavasti jäänyt tekemättä epähuomiossa
verkatakki					
musiikki		musiikki (yso:p1808)			Siidan sanalista oli virheellinen tämän termin suhteen (suomen- ja saamenkieliset termit erosivat toisistaan), joten sitä ei huomioida
Inari	?	?	Inari (yso:p105976) ?	Inari (yso:p105976) ?	Annifin sanastossa on mukana myös termejä YSO-paikat-sanastosta. Kehittäjän A mukaan tämä on hyvä, mutta Kehittäjän B mukaan ne pitäisi voida poistaa hakutuloksista
ajohihna, -hihnat					
inarinsaamelaiset	inarinsaamelaiset (yso:p25252)	inarinsaamelaiset (yso:p25252)	inarinsaamelaiset (yso:p25252)	inarinsaamelaiset (yso:p25252)	
suutari	suutarit (ammatit) (yso:p1540)	✕	suutarit (ammatit) (yso:p1540)	✕	Kehittäjä B on luultavasti jättänyt linkityksen tekemättä epähuomiossa
kovasin					
etto					
työvaate, työvaatteet	työvaatteet (yso:p9195)	työvaatteet (yso:p9195)	työvaatteet (yso:p9195)	työvaatteet (yso:p9195)	
tinalankatyö, -työt					
oven, astian kahva	kahvat (yso:p39278)				Annifin täsmäytys- ja etuliitehaku eivät kumpikaan löytäneet haettua termiä. Kehittäjä A käytti Skosmoksen etuliitehakua linkittämisessä.

Siidan termi	Kehittäjän A Skosmos-linkitykset	Kehittäjän B Skosmos-linkitykset	Kehittäjän A Annif-linkitykset	Kehittäjän B Annif-linkitykset	Kommentti
		✕	✕	✕	
koriste, koristeet	koristeet (yso:p4463)	koristeet (yso:p4463)	koristeet (yso:p4463)	koristeet (yso:p4463)	

7 Yhteenveto ja pohdinta

Tässä luvussa kootaan yhteen tutkimuksen tulokset, arvioidaan tutkimusprosessia ja pohditaan tutkimuksen rajoituksia sekä mahdollisia jatkotoimenpiteitä. Alaluvussa 7.1 käydään läpi johdannossa esitetyt tutkimuskysymykset ja vastataan niihin. Alaluvussa 7.2 arvioidaan tutkimusta suunnittelutieteen periaatteiden näkökulmasta. Alaluvussa 7.3 käsitellään vielä puutteita tutkimuksessa ja mahdollisia jatkotutkimuksen aiheita.

7.1 Tutkimuskysymykset

Tutkimuskysymys 1: Minkälaisia tarpeita kulttuuriperintöorganisaatioissa on täsmäytykselle ja mitkä vaatimukset nämä tarpeet asettavat toteutettavalle järjestelmälle?

Tämän tutkimuskysymyksen avulla oli tarkoitus kerätä tietoa Kansalliskirjaston tarpeista täsmäytykselle sekä vaatimuksista diplomityössä kehitettäville täsmäytysjärjestelmille. Tarpeet ja vaatimukset selvitettiin Kansalliskirjaston sanastotyön tekijöiden kanssa toteutetuilla haastatteluilla (haastattelu 1 ja haastattelu 2). Haastattelun 1 tavoitteena oli saada ymmärrys täsmäytyksen aiemmista työnkuluista kirjastolla sekä tarpeista siihen tulevaisuudessa. Siinä haastateltiin sekä teknisiä asiantuntijoita että sanastoja kehittäviä asiantuntijoita. Haastattelussa 2 syvennettiin ymmärrystä täsmäytyksen työnkuluista sekä aiempien ratkaisujen puutteista kontekstuaalisen haastattelun muodossa. Haastateltavana siinä toimi Kansalliskirjaston sanastokehittäjä.

Haastatteluissa selvisi, että Kansalliskirjastolla oli tehty täsmäytystä kohdistuen lähinnä ulkoisiin sanastoihin, pääasiassa Wikidataan. Täsmäytystyössä oli käytetty OpenRefine-työkalua, joka oli todettu hyväksi ratkaisuksi aiempiin täsmäytyksen ongelmiin. Haastateltavien mukaan tulevaisuudessa täsmäytystä on tarpeellista tehdä kohdistuen Kansalliskirjaston omiin sanastoihin, esimerkiksi YSO-ontologiaan. Täsmäytettäviä aineistoja olisivat esimerkiksi museoissa käytössä olevat sanalistat, jotka halutaan rikastaa linkitetyn datan tunnisteilla sekä linkitetyn datan sanastot, joiden käsitteet halutaan liittää johonkin Kansalliskirjaston sanastoon linkitetyn datan vastaavuussuhteilla.

Tunnistettujen tarpeiden perusteella asetettiin vaatimukset kehitettäville järjestelmille. OpenRefine-työkalun käyttö oli tärkeä osa aiempaa täsmäytystyötä ja se koettiin toimivaksi ratkaisuksi, joten yhteensopivuus sen kanssa oli yksi pääasiallisista vaatimuksista järjestelmille. Myös mahdollisuus tehdä täsmäytystä useille Kansalliskirjaston ylläpitämille sanastoille koettiin tärkeäksi, joten myös se asetettiin vaatimukseksi.

Tutkimuskysymys 2: Miten järjestelmä toteutetaan vastaamaan asetettuja vaatimuksia?

Toisen tutkimuskysymyksen tavoitteena oli selvittää, miten kehitettävät täsmäytysjärjestelmät rakennetaan, mitä toiminnallisuuksia ne tarjoavat ja mitä teknologioita niissä hyödynnetään. Tutkimuksessa selvitettiin kirjallisuudessa esitetyt aiemmat ratkaisut täsmäytysjärjestelmien tekemiseen sekä täsmäytyksessä ja tietuelinkityksessä käytetyt teknologiat. Myös haastatteluja käytettiin hyväksi teknisten ratkaisujen tekemisessä. Aiemmin mainitun haastattelun 2 lisäksi tähän käytettiin haastattelua 3, jonka avulla selvitettiin tarkemmin mitä ominaisuuksia työssä kehitetyillä järjestelmillä tulisi olla.

Työssä käsitellyistä, kirjallisuudessa esitetyistä täsmäytysjärjestelmistä mikään ei suoraan soveltunut muiden sanastojen tai organisaatioiden käyttöön. Myöskään olemassa olevat täsmäytyksen ja tietuelinkityksen ohjelmistot tai ohjelmakirjastot eivät soveltuneet täsmäytysjärjestelmän rakentamiseen Kansalliskirjastolle, joten oli tarve kehittää kokonaan uusi ratkaisu kirjaston sanastoihin kohdistuvaan täsmäytykseen. Kirjaston kehittämät ja ylläpitämät sanastotyössä käytetyt Skosmos- ja Annif-ohjelmistot mahdollistavat käsitteiden hakemisen linkitetyn datan sanastoista, mikä puolestaan mahdollisti täsmäytysjärjestelmien rakentamisen. Työssä kehitettiin siis kaksi prototyyppiä, yksi REST-rajapintakääreenä Skosmoksen REST-rajapinnan päälle ja yksi osaksi Annif-ohjelmiston REST-rajapintaa.

Vaatimus yhteensopivuudesta OpenRefine-ohjelmiston kanssa edellytti kehittämään järjestelmät, jotka toteuttavat W3C:n täsmäytysrajapintamääritelmän. Haastatteluissa 2 ja 3 selvitettiin rajapintamääritelmän ominaisuudet, jotka prototyypeissä toteutettiin. Kehitetty Skosmos-prototyyppi käyttää Finto-palvelun REST-rajapintaa toiminnoissaan, esimerkiksi rajapinnan hakumetodia käytetään täsmäytyshakujen tekemiseen. Haastattelun 3 perusteella prototyyppiin lisättiin myös etuliitehaku, käsitteiden esikatselu sekä toiminnot käsitteiden ominaisuuksien löytämiseen ja arvojen hakemiseen. Näissä toiminnallisuuksissa käytetään RDF-dataa, joka myös haetaan Finton REST-rajapinnasta. Annif-prototyyppi, joka käyttää Annifin sisäistä tekoälypohjaista käsitteiden ehdotustoimintoa, toteutettiin osana sen rajapintaa. Myös etuliitehakuja voidaan tehdä prototyypin avulla, mutta muita ominaisuuksia ei ollut mahdollista toteuttaa, sillä Annifin suorituskyky sanastojen käsittelyssä ei ollut riittävä.

Tutkimuskysymys 3: Miten järjestelmää arvioidaan ja miten sen toiminta vastaa arviointiperusteita?

Kolmannen tutkimuskysymyksen tavoitteena oli selvittää mitä arviointiperusteita käytetään kehitettyjen täsmäytysjärjestelmäprototyyppien arvioinnissa ja miten niiden toiminta vastaa näitä perusteita. Diplomityössä käytettiin kahta haastattelukierrosta (haastattelu 3 ja haastattelut 4A ja 4B)

arvioimaan toteutettujen järjestelmien käytettävyyttä ja soveltuvuutta Kansalliskirjaston tarpeisiin. Haastattelu 3, jonka tavoitteena oli arvioida prototyyppien ensimmäisten versioiden toimintaa testidatan avulla, toteutettiin puolistrukturoituna haastatteluna Kansalliskirjaston sanastokehittäjille. Haastattelut 4A ja 4B, joissa haastateltavat käyttivät lopullisia rajapintaprototyyppejä OpenRefine-ympäristössä, puolestaan toteutettiin kontekstuaalisina haastatteluina samoille kehittäjille. Neljänsien haastattelujen tuloksia käytettiin lisäksi arvioimaan prototyyppien avulla tuotettujen linkitysten laatua. Haastatteluiden ja linkitysten analysoinnin avulla arvioitiin prototyyppejä itsenäisesti ja niiden toimintaa verrattiin toisiinsa.

Haastattelussa 3 selvisi, että molemmista rajapinnoista puuttui vielä useita tärkeitä ominaisuuksia, jotka vaikeuttivat niiden käyttöä, esimerkiksi mahdollisuus hakea käsitteitä yksitellen sanastosta ja esikatsella niiden ominaisuuksia. Rajapintaprototyyppien lopullisiin versioihin pyrittiin lisäämään haastattelussa ilmenneet puuttuvat toiminnot ja näin syntyneitä lopullisia prototyyppejä arvioitiin seuraavissa haastatteluissa. Haastateltavat arvioivat, että Skosmoksen lopullisen prototyypin täsmäytyshakujen tulokset olivat hyödyllisiä ja niistä oli helppo löytää oikea käsite. Sen muut ominaisuudet toimivat myös odotetulla tavalla ja nopeuttivat täsmäytysprosessia. Haastateltavat kokivat, että Annifin lopullisen prototyypin hakutulokset olivat semanttisesti monipuolisempia ja siten mielenkiintoisempia kuin Skosmoksen prototyypin. Ne olivat kuitenkin osittain epäoleellisia ja siksi hankaloittivat oikean käsitteen löytymistä. Kehittäjät toivoivat, että Annifin prototyypin etuliitehaku olisi toiminut kuten Skosmoksen. Haastateltavien mukaan myös käsitteiden esikatselu voisi olla osana rajapintaa. Vaikka haastateltavat näkivät molempien prototyyppien käytön täsmäytystyössä mahdollisena, pitivät he Skosmoksen prototyyppiä valmiimpana käyttöönottoon, kun taas Annifin prototyyppi vaatisi heidän mukaansa vielä kehitystä ennen laajempaa käyttöönottoa.

Haastatteluissa 4A ja 4B tuotettuja linkityksiä arvioitiin laadullisen analyysin avulla haastattelujen jälkeen. Analyysin tulokset mukailivat haastateltavien arviota prototyypeistä. Skosmoksen rajapinnalla kyettiin löytämään lähes kaikki vastaavat käsitteet molemmilla testiaineistoilla. Annifin rajapinta suoriutui miltei yhtä hyvin kuin Skosmoksen, mutta osalle testiaineiston termeistä ei kyetty tekemään linkityksiä. Yleisesti ottaen Skosmoksen rajapinta tarjosi luotettavammin tuloksia, mutta Annifin rajapinta tuotti monipuolisempia ehdotuksia ja kykeni löytämään käsitteitä myös semanttisten vastaavuuksien perusteella.

7.2 Tutkimusprosessin arviointi

Tämä diplomityö toteutettiin noudattaen suunnittelutieteen menetelmiä. Tutkimusprosessin aikana pyrittiin noudattamaan Hevnerin ym. (2004)

esittämiä suunnittelutieteen ohjesääntöjä, jotka on esitetty taulukossa 5.1. Ohjesääntöjen toteutumista diplomityössä on arvioitu taulukossa 7.1.

Taulukko 7.1. Hevnerin ym. (2004) esittämien suunnitelutieteen ohjesääntöjen toteutuminen diplomityössä

Ohjesääntö	**Toteutuminen diplomityössä**
1. Artefaktin tuottaminen	Työssä tuotettiin kaksi instanssia Skosmoksen ja Annifin täsmäytysrajapintaprototyyppien muodossa.
2. Ongelman merkitys	Kansalliskirjaston sanastotyön ohjelmistot eivät mahdollistaneet automaattisen täsmäytyksen tekemistä sen omiin sanastoihin, joten täsmäytykseen oli kehitettävä erilliset järjestelmät. Tarkat tarpeet täsmäytyksen toteutukselle sekä vaatimukset järjestelmille selvitettiin käyttämällä haastatteluja Kansalliskirjaston sanastotyön tekijöiden kanssa.
3. Artefaktin arviointi	Prototyyppien tarpeita ja toteutuksia arvioitiin Kansalliskirjaston sanastokehittäjien kanssa suoritetulla puolistrukturoidulla ja kontekstuaalisella haastattelulla sekä analysoimalla haastatteluissa tuotettua dataa laadullisesti.
4. Tutkimuksen kontribuutio	Olemassa olevat ratkaisut eivät tutkimuksen mukaan olleet riittäviä vastamaan Kansalliskirjaston tarpeisiin, joten oli kehitettävä itsenäiset järjestelmät täsmäytyksen tekemiseen. Ne demonstroivat, että on mahdollista kehittää täsmäytysjärjestelmä, joka täyttää kulttuuriperintöorganisaation tarpeet linkitetyn datan sanastojen täsmäytykseen. Tutkimuksessa saatiin myös tietoa siitä, miten tekoälyyn ja yksinkertaiseen merkkijonojen vertailuun perustuvat lähestymistavat täsmäytyksessä vaikuttavat täsmäytysprosessin sujuvuuteen sekä linkityksen laatuun. Lisäksi tässä työssä keskityttiin järjestelmien kehittämisessä niiden käytettävyyteen loppukäyttäjien näkökulmasta, mitä ei ole painotettu aiemmassa tutkimuksessa.
5. Tutkimuksen tarkkuus	Olemassa olevat ratkaisut täsmäytykseen käytiin systemaattisesti läpi ja ne todettiin soveltumattomiksi ratkaisemaan esitettyä ongelmaa. Kehitettävien järjestelmien vaatimusten selvittämiseen käytettiin useita haastatteluja loppukäyttäjien kanssa. Järjestelmien toimintaa arvioitiin erilaisten haastattelujen sekä laadullisen data-analyysin avulla.
6. Suunnittelu hakuprosessina	Artefaktit toteutettiin kahdessa iteraatiossa, jotka koostuivat rakennusvaiheesta ja arviointivaiheesta.

Ohjesääntö	Toteutuminen diplomityössä
	Artefaktit pyrittiin rakentamaan asetettuja vaatimuksia vastaaviksi rakennusvaiheessa ja vaatimuksia päivitettiin arviointivaiheen yhteydessä.
7. Tutkimuksesta tiedottaminen	Ensisijainen tapa tiedottaa tutkimuksesta on tämän diplomityön julkaisu. Toteutetuista prototyypeistä on viestitty myös sisäisesti Kansalliskirjastolla. Lisäksi Skosmoksen ja Annifin käyttäjille on viestitty prototyypeistä ohjelmistojen GitHub-sivuilla[36, 37] . Myös prototyyppien koodi on saatavilla julkisesti GitHub-sivuilla.

7.3 Tutkimuksen rajoitukset ja jatkotutkimuksen aiheet

Tutkimuksessa nousi esiin joitakin rajoituksia, jotka vaikuttavat tulosten luotettavuuteen. Useimmissa haastatteluissa ei ollut mahdollisuutta haastella enempää kuin kahta sanastokehittäjää, mikä todennäköisesti tarkoittaa, että kaikki käyttötapaukset ja loppukäyttäjien näkökulmat eivät tulleet esiin tutkimuksessa. Järjestelmät kehitettiin yhteistyössä Kansalliskirjaston kanssa sen tarpeisiin, joten muiden Skosmoksen ja Annifin käyttäjien näkökulmat eivät myöskään näy tutkimuksessa. Lopullisen data-analyysin tuloksia tarkastellessa on otettava huomioon, että erilaiset käytännöt täsmäytyksen tekemisessä vaikuttavat tehtyihin linkityksiin. Kehittäjillä ei ollut ennalta sovittuja yhteisiä käytäntöjä, joten tuotetut linkitykset eivät välttämättä kerro koko totuutta rajapintojen toiminnasta eivätkä analyysin tulokset siten täysin yleisty kehitetyille järjestelmille.

Tulevaisuudessa tutkimuksen tulosten perusteella voidaan toteutetut järjestelmät jatkokehittää prototyyppivaiheesta osaksi Skosmos- ja Annif-ohjelmistoja. Järjestelmien toimintoja tulisi ennen tätä kehittää edelleen tutkimuksessa saadun tiedon perusteella. Annifin rajapinnan etuliitehakua on tarpeen kehittää vastaamaan rajapintamääritelmää läheisemmin ja rajapintaan tulisi lisätä käsitteiden esikatselu. Skosmoksen täsmäytysrajapinta voidaan puolestaan integroida osaksi Skosmoksen omaa REST-rajapintaa, jotta se mahdollistaisi täsmäytyksen tekemisen kaikilla Skosmos-asennuksilla. Molempien rajapintojen toiminnallisuutta voidaan lisäksi kehittää vastaamaan paremmin myös muiden Skosmoksen ja Annifin käyttäjien tarpeisiin sekä Kansalliskirjaston sisällä että muissa organisaatioissa. Tämän saavuttaminen vaatisi jatkotutkimuksia. Toinen jatkotutkimuksen kohde olisi täsmäytysrajapintamääritelmän kehittäminen vastamaan paremmin erilaisten järjestelmien tarpeisiin. Määritelmän seuraavasta versiosta on luotu luonnos (Delpeuch, 2024), jossa on jo otettu huomioon tässä tutkimuksessa ilmennyt

[36] https://github.com/NatLibFi/Skosmos
[37] https://github.com/NatLibFi/Annif

ongelma palvelumanifestin ”identifierSpace”- ja ”schemaSpace”-kenttien pakollisuudesta. Määritelmää on tärkeää kuitenkin kehittää edelleen vielä jatkossa.

Lähdeluettelo

Berners-Lee, T. (2006). Linked Data—Design Issues. Haettu 6.4.2024 osoitteesta http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 34–43. https://doi.org/10.1038/scientificamerican0501-34

Binding, C., & Tudhope, D. (2016). Improving interoperability using vocabulary linked data. International Journal on Digital Libraries, 17(1), 5–21. https://doi.org/10.1007/s00799-015-0166-y

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22. https://doi.org/10.4018/jswis.2009081901

Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1. Haettu 6.4.2024 osoitteesta https://www.w3.org/TR/2014/REC-rdf-schema-20140225/

Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection (1. p.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31164-2

Christophides, V., Efthymiou, V., & Stefanidis, K. (2015). Entity Resolution in the Web of Data (1. p.). Springer Cham. https://doi.org/10.1007/978-3-031-79468-1

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. Haettu 6.4.2024 osoitteesta https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Delpeuch, A. (2019). A survey of openrefine reconciliation services. Computing Research Repository (CoRR), abs/1906.08092. https://doi.org/arXiv:1906.08092

Delpeuch, A. (2020). Running a reconciliation service for wikidata. Teoksessa L. A. Kaffee, O. Tifrea-Marciuska, E. Simperl, & D. Vrandecic (Toim.), Proceedings of the 1st Wikidata Workshop (Wikidata 2020) (Vsk. 2773). CEUR-WS.org.

Delpeuch, A. (2022). Overview of Wikibase support. Haettu 6.4.2024 osoitteesta https://openrefine.org/docs/manual/wikibase/overview

Delpeuch, A. (2023). Reconciling. Haettu 6.4.2024 osoitteesta https://openrefine.org/docs/manual/reconciling

Delpeuch, A., Pohl, A., Steeg, F., Guidry Sr, T., & Suominen, O. (2023). Reconciliation Service API v0.2. Haettu 6.4.2024 osoitteesta https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/

Delpeuch, A., Pohl, A., Steeg, F., Guidry Sr, T., & Suominen, O. (2024). Reconciliation Service API - Draft Community Group Report. Haettu 21.4.2024 osoitteesta https://reconciliation-api.github.io/specs/draft/

Garcia, G. (2023). Getty Vocabularies OpenRefine Tutorial. Haettu 6.4.2024 osoitteesta https://www.getty.edu/research/tools/vocabularies/obtain/openrefine_tutorial.pdf

Garlik, S. H., & Seaborne, A. (2013). SPARQL 1.1 Query Language. Haettu 6.4.2024 osoitteesta https://www.w3.org/TR/2013/REC-sparql11-query-20130321/

Getty Research Institute. (n.d.). Getty Vocabularies. Haettu 6.4.2024 osoitteesta https://www.getty.edu/research/tools/vocabularies/index.html

Gu, L., & Baxter, R. (2004). Adaptive filtering for efficient record linkage. Teoksessa M. W. Berry, U. Dayal, C. Kamath, & D. B. Skillicorn (Toim.), Proceedings of the Fourth SIAM International Conference on Data Mining (ss. 477–481). SIAM. https://doi.org/10.1137/1.9781611972740.50

Harper, C. A., & Tillett, B. B. (2007). Library of congress controlled vocabularies and their application to the Semantic Web. Cataloging and Classification Quarterly, 43(3–4), 47–68. https://doi.org/10.1300/J104v43n03_03

Harpring, P. (2010). Introduction to controlled vocabularies: Terminology for art, architecture, and other cultural works (1. p.). Getty Publications.

hbz. (2024a). Lobid-gnd. Haettu 8.4.2024 osoitteesta https://github.com/hbz/lobid-gnd/tree/master

hbz. (2024b). Lobid-organisations. Haettu 8.4.2024 osoitteesta https://github.com/hbz/lobid-organisations

hbz. (n.d.). Lobid-organisations. Haettu 6.4.2024 osoitteesta https://lobid.org/organisations

Hedden, H. (2008). Controlled vocabularies, thesauri, and taxonomies. The Indexer, 26(1), 33–34. https://doi.org/doi:10.3828/indexer.2008.8

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly: Management Information Systems, 28(1), 75–105. https://doi.org/10.2307/25148625

Holtzblatt, K., Wendell, J., & Wood, S. (2005). Rapid Contextual Design. Elsevier Inc. https://doi.org/10.1016/B978-0-12-354051-5.X5000-9

Hyvönen, E. (2002). Semantic Web—The new Internet of meanings. Teoksessa E. Hyvönen (Toim.), Semantic Web Kick-Off in Finland-Vision, Technologies, Research and Applications (ss. 3–26). Helsinki Institute for Information Technology.

Hyvönen, E. (2021). Sammon taontaa semanttisessa webissä (Forging Sampos on the Semantic Web). Tekniikan Waiheita, 39(2), 87–105. https://doi.org/10.33355/tw.102864

Hyvönen, E., Viljanen, K., Tuominen, J., & Seppälä, K. (2008). Building a National Semantic Web Ontology and Ontology Service Infrastructure—The FinnONTO Approach. Teoksessa S. Bechhofer, M. Hauswirth, J. Hoffmann, & M. Koubarakis (Toim.), The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008 (ss. 95–109). Springer. https://doi.org/10.1007/978-3-540-68234-9_10

Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman II, C. (2014). Five stars of Linked Data vocabulary use. Semantic Web, 5(3), 173–176. https://doi.org/10.3233/SW-140135

Johannesson, P., & Perjons, E. (2014). An introduction to design science (1. p.). Springer Cham. https://doi.org/10.1007/978-3-319-10632-8

Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment, 3(1–2), 484–493. https://doi.org/10.14778/1920841.1920904

Levenštein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics-Doklady, 10(8), 707–710.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Meghini, C., Scopigno, R., Richards, J., Wright, H., Geser, G., Cuy, S., Fihn, J., Fanini, B., Hollander, H., Niccolucci, F., Felicetti, A., Ronzino, P., Nurra, F., Papatheodorou, C., Gavrilis, D., Theodoridou, M., Doerr, M., Tudhope, D., Binding, C., & Vlachidis, A. (2017). ARIADNE: A research infrastructure for archaeology. Journal on Computing and Cultural Heritage, 10(3), 1–27. https://doi.org/10.1145/3064527

Miles, A., & Bechhofer, S. (2009). SKOS Simple Knowledge Organization System Reference. Haettu 6.4.2024 osoitteesta https://www.w3.org/TR/2009/REC-skos-reference-20090818/

OpenRefine (2023). (Versio 3.7.4) [Tietokoneohjelmisto]. https://openrefine.org

Papadakis, G., Fisichella, M., Schoger, F., Mandilaras, G., Augsten, N., & Nejdl, W. (2023). Benchmarking Filtering Techniques for Entity Resolution. 2023 IEEE 39th International Conference on Data Engineering (ICDE), 653–666. https://doi.org/10.1109/ICDE55515.2023.00389

Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and Filtering Techniques for Entity Resolution: A Survey. ACM Computing Surveys, 53(2), 1–42. https://doi.org/10.1145/3377455

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design science research evaluation (K. Peffers, M. Rothenberger, & B. Kuechler, Toim.; ss. 398–410). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29863-9_29

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pohl, A., Steeg, F., & Christoph, P. (2018). lobid – Dateninfrastruktur für Bibliotheken. Informationspraxis, 4(1), 17. https://doi.org/https://doi.org/10.11588/ip.2018.1.52445

Reconciliation service test bench. (n.d.). Haettu 6.4.2024 osoitteesta https://reconciliation-api.github.io/testbench/

Semantic Web Standards. (2019). Haettu 6.4.2024 osoitteesta https://www.w3.org/2001/sw/wiki/Main_Page

Smith, C. (2021). Controlled Vocabularies: Past, Present and Future of Subject Access. Cataloging and Classification Quarterly, 59(2–3), 186–202. https://doi.org/10.1080/01639374.2021.1881007

Steeg, F., & Pohl, A. (2019). GND reconciliation for OpenRefine. Haettu 6.4.2024 osoitteesta https://blog.lobid.org/2018/08/27/openrefine.html

Steeg, F., Pohl, A., & Christoph, P. (2019). Lobid-gnd–Eine Schnittstelle zur Gemeinsamen Normdatei für Mensch und Maschine. Informationspraxis, 5(1), 25. https://doi.org/10.11588/ip.2019.1.52673

Suominen, O., Inkinen, J., & Lehtinen, M. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing. JLIS.It, 13(1), 265–282. https://doi.org/10.4403/jlis.it-12740

Suominen, O., Pessala, S., Tuominen, J., Lappalainen, M., Nykyri, S., Ylikotila, H., Frosterus, M., & Hyvönen, E. (2014). Deploying National Ontology Services: From ONKI to Finto (A. Polleres, A. Garcia, & R. Benjamins, Toim.).

Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., & Retterath, A. (2015). Publishing SKOS vocabularies with Skosmos [Julkaisematon käsikirjoitus].

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. European Journal of Information Systems, 25(1), 77–89. https://doi.org/10.1057/ejis.2014.36

A. Rajapintaprototyyppien palautusarvoja

Tähän liitteeseen on kerätty esimerkit Skosmos- ja Annif-rajapintaprototyyppien palvelumanifesteista JSON-muodossa.

Skosmos-prototyyppien palvelumanifestit

Kuvassa A.1 on esitetty ensimmäisen Skosmos-prototyypin palvelumanifesti ja kuvassa A.2 lopullisen Skosmos-prototyypin manifesti.


```
{
  "defaultTypes": [
    {
      "id": "http://www.w3.org/2004/02/skos/core#Concept",
      "name": "Käsite"
    },
    {
      "id": "http://www.w3.org/2004/02/skos/core#Collection",
      "name": "Kokoelma"
    },
    {
      "id": "http://purl.org/iso25964/skos-thes#ConceptGroup",
      "name": "Käsiteryhmä"
    },
    {
      "id": "http://purl.org/iso25964/skos-thes#ThesaurusArray",
      "name": "Sisarkäsitteiden joukko"
    },
    {
      "id": "http://www.yso.fi/onto/yso-meta/Concept",
      "name": "Yleiskäsite"
    },
    {
      "id": "http://www.yso.fi/onto/yso-meta/Individual",
      "name": "Yksilökäsite"
    },
    {
      "id": "http://www.yso.fi/onto/yso-meta/Hierarchy",
      "name": "Hierarkisoiva käsite"
    }
  ],
  "identifierSpace": "http://www.yso.fi/onto/yso/",
  "name": "Skosmos reconciliation service for YSO - Yleinen suomalainen ontologia",
  "schemaSpace": "",
  "view": {
    "url": "{{id}}"
  }
}
```


Kuva A.1. Ensimmäisen Skosmos-prototyypin palvelumanifesti suomenkieliselle YSO:lle.

```
{
  "defaultTypes": [
    {
      "id": "http://www.w3.org/2004/02/skos/core#Concept",
      "name": "Käsite"
    },
    {
      "id": "http://id.nlm.nih.gov/mesh/vocab#TopicalDescriptor",
      "name": "Aihe"
    },
    {
      "id": "http://id.nlm.nih.gov/mesh/vocab#PublicationType",
      "name": "Julkaisutyyppi"
    },
    {
      "id": "http://id.nlm.nih.gov/mesh/vocab#GeographicalDescriptor",
      "name": "Maantieteellinen käsite"
    }
  ],
  "extend": {
    "property_settings": [
      {
        "default": 0,
        "help_text": "Maximum number of values to return per row",
        "label": "Limit",
        "name": "limit",
        "type": "number"
      }
    ],
    "propose_properties": {
      "service_path": "/propose_properties",
      "service_url": "<täsmäytyspalvelu-URL>/mesh/fi/reconcile"
    }
  },
  "identifierSpace": "http://www.yso.fi/onto/mesh/",
  "name": "Reconciliation service for MeSH / FinMeSH (fi)",
  "preview": {
    "height": 100,
    "url": "<täsmäytyspalvelu-URL>/mesh/fi/reconcile/preview?id={{id}}",
    "width": 300
  },
  "schemaSpace": "",
  "suggest": {
    "entity": {
      "service_path": "/suggest/entity",
      "service_url": "<täsmäytyspalvelu-URL>/mesh/fi/reconcile"
    }
  },
  "view": {
    "url": "{{id}}"
  }
}
```

Kuva A.2. Lopullisen Skosmos-prototyypin palvelumanifesti suomenkieliselle MeSH-sanastolle. Palvelun verkko-osoite on korvattu ”<täsmäytyspalvelu-URL>”-tekstillä.

Annif-prototyyppien palvelumanifestit

Kuvassa A.3 on esitetty ensimmäisen Annif-prototyypin palvelumanifesti ja kuvassa A.4 lopullisen Annif-prototyypin manifesti.

```
{
  "defaultTypes": [
    {
      "id": "default-type",
      "name": "Default type"
    }
  ],
  "identifierSpace": "",
  "name": "Annif Reconciliation Service for TF-IDF Finnish",
  "schemaSpace": "http://www.w3.org/2004/02/skos/core#Concept",
  "versions": [
    "0.2"
  ],
  "view": {
    "url": "{{id}}"
  }
}
```

Kuva A.3. Ensimmäisen Skosmos-prototyypin palvelumanifesti suomenkieliselle YSO-pohjaiselle projektille.

```
{
  "defaultTypes": [
    {
      "id": "default-type",
      "name": "Default type"
    }
  ],
  "identifierSpace": "",
  "name": "Annif Reconciliation Service for YSO suomi (2023.6.Hypatia)",
  "schemaSpace": "http://www.w3.org/2004/02/skos/core#Concept",
  "suggest": {
    "entity": {
      "service_path": "/suggest/entity",
      "service_url": "<täsmäytyspalvelu-URL>/v1/projects/yso-fi/reconcile"
    }
  },
  "versions": [
    "0.2"
  ],
  "view": {
    "url": "{{id}}"
  }
}
```

Kuva A.4. Lopullisen Annif-prototyypin palvelumanifesti suomenkieliselle YSO-pohjaiselle projektille. Palvelun verkko-osoite on korvattu ”<täsmäytyspalvelu-URL>”-tekstillä.

B. Haastattelupohjat

Tähän liitteeseen on kerätty haastatteluissa 2, 3 ja 4 käytetyt haastattelupohjat.

Haastattelun 2 kysymykset

- Minkälaista dataa Wikidatan rajapinnalla linkitetään?
  - Mitä sanastoja linkitetään?
  - Missä muodossa ne tuodaan OpenRefine-ympäristöön?
- Kuinka paljon linkitettävää dataa on?
- Miten päätös linkitettävästä käsitteestä tehdään?
  - Vaikuttaako kandidaattien pisteytys päätökseen? Miten?
  - Miten käsitteitä tarkastellaan OpenRefine-ympäristön ulkopuolella?
- Miten toimitaan, jos täsmäytyshaku ei palauta vastaavaa käsitettä?
  - Jos vastaava käsite on olemassa, mutta rajapinta ei löydä sitä?
  - Jos vastaavaa käsitettä ei ole olemassa?
- Kuka käyttää linkityksessä syntynyttä dataa ja mihin tarkoitukseen?
- Mitä puutteita Wikidatan rajapinnassa on?
- Mikä Wikidatan rajapinnassa toimii hyvin?
- Miten linkitettävien käsitteiden ja kohdesanaston käsitteiden kielet tulisi ottaa huomioon täsmäytysjärjestelmässä?

Haastattelun 3 kysymykset

- Mikä toteutetuissa rajapintaprototyypeissä toimii hyvin?
- Mitä toiminnallisuuksia prototyypeistä vielä puuttuu?
  - Kuinka kriittisiä ne ovat linkitysprosessin kannalta?
- Toimivatko kaikki prototyyppien ominaisuudet odotetulla tavalla?

- Kuinka tärkeää on, että ne toimivat odotetulla tavalla?

- Kuinka hyödyllisiltä täsmäytyshaun tulokset vaikuttavat?
  - Onko niistä helppoa löytää oikeaa käsitettä?

- Miten hakutuloksia pitäisi pystyä rajaamaan?
  - Voidaanko käsitteiden ominaisuuksia tai tyyppejä käyttää täsmäytyksessä?

- Tarvitaanko muita rajapinnan toiminnallisuuksia (etuliitehaku, esikatselu, ja konseptien ominaisuuksien arvojen hakeminen)?

- Miten linkitettävien käsitteiden ja kohdesanaston käsitteiden kielet tulisi ottaa huomioon?

Haastattelujen 4A ja 4B kysymykset

Haastatteluissa 4A ja 4B haastateltavat toivat tehtäväosiossa testiaineistot OpenRefine-ympäristöön ja tekivät erilliset linkitykset molemmille aineistoille. Tehtäväosio suoritettiin siis kaksi kertaa molemmissa haastatteluissa. Tehtäväosioiden jälkeen suoritettiin vielä haastatteluosio.

Tehtäväosio:

1) Tuodaan testiaineistotiedosto OpenRefine-ympäristöön uuteen projektiin

2) Tehdään täsmäytyshaku molemmilla rajapinnoilla kaikille kielille
   a. Tehdään sopivat valinnat haun parametreille

3) Käydään läpi kaikki käsitteet molemmilla rajapinnoilla ja tehdään niille linkitykset suomenkieliseen sarakkeeseen
   a. Käytetään ainakin osassa linkityksistä hyväksi etuliitehakua molemmilla rajapinnoilla
   b. Käytetään ainakin osassa linkityksistä hyväksi esikatselua Skosmoksen rajapinnalla

4) Kun linkitysvalinnat on tehty, haetaan linkitettyjen käsitteiden ominaisuuksien arvoja Skosmoksen rajapinnan extend-toiminnallisuudella ja tarkastellaan sen tuloksia

5) Tallennetaan projekti ”OpenRefine project archive”-tiedostomuodossa

Haastatteluosio:

- Mikä rajapintaprototyypeissä toimii hyvin?
- Mitä toiminnallisuuksia prototyypeistä vielä puuttuu?
- Miten hyödyllisiä täsmäytyshaun tulokset olivat?
  - Onko niistä helppoa löytää oikeaa käsitettä?
  - Mikä tekisi niistä hyödyllisempiä?
- Ovatko lisätoiminnallisuudet (etuliitehaku, esikatselu, ja konseptien ominaisuuksien arvojen hakeminen) hyödyllisiä?
  - Löytyikö etuliitehaulla muutoin löytymättömiä käsitteitä? Miten helppoa niitä oli löytää?
  - Auttoiko esikatselu linkityspäätösten tekemisessä? Puuttuiko siitä tietoa?
  - Olivatko extend-toiminnallisuuden tarjoamat ominaisuudet hyödyllisiä? Puuttuiko niistä jotain?
- Käyttäisitkö rajapintoja tällaisenaan työssäsi?