Informaatiotutkimus 1–2(42), 31–50

KATSAUS

Tutkimusdatan saatavuustiedot tieteellisissä artikkeleissa: Raportti Data Availability Statementien käytöstä Tampereen yliopistossa

Tomi Toikko
Tampereen yliopisto
tomi.toikko@tuni.fi
https://orcid.org/0000-0001-8176-0148

Kaisa Kylmälä
Tampereen yliopisto
kaisa.kylmala@tuni.fi
https://orcid.org/0000-0001-8227-6817

Selvityksessä kartoitettiin kuinka yleinen Data Availability Statement (DAS) on Tampereen yliopiston tutkijoiden julkaisemissa vertaisarvioiduissa artikkeleissa (n = 2085) sekä millaista tietoa tutkimusdatasta ja sen jakamisesta saadaan DAS-ilmoitusten avulla. Data Availability Statement löytyi noin joka neljännestä artikkelista. DAS-ilmoitusten yleisyyttä ja datan avoimuutta tarkasteltiin muun muassa kustantajan kansainvälisyyden, Julkaisufoorumi-tasojen ja julkaisun avoimuuden kautta. Selvityksessä havaittiin, että avoimesti julkaistuissa artikkeleissa on suhteellisesti enemmän DAS-ilmoituksia ja avointa dataa kuin maksumuurin takana olevissa artikkeleissa. Data on tallennettu arkistojen lisäksi usein artikkelin yhteyteen. Lisäksi monet tutkijat säilyttävät dataa itse ja lupaavat jakaa sitä pyydettäessä. Tutkijan jakaessa dataa lähes aina rajoitteeksi on mainittu ”reasonable request”. Datan jakamattomuudelle yleisimpiä perusteluja ovat sensitiivinen data, lainsäädäntö ja omistajuus.

Asiasanat: artikkelit (julkaisut), avoin tiede, data, saatavuus, tiedelehdet, tutkimusaineisto

Pysyvä osoite: https://doi.org/10.23978/inf.126098

Lähtökohdat

Tampereen yliopiston kirjastossa haluttiin selvittää, kuinka yleinen Data Availability Statement (DAS) on Tampereen yliopiston tutkijoiden julkaisemissa tieteellisissä artikkeleissa. Lisäksi selvitettiin, millaista tietoa tutkimusdatasta ja sen jakamisesta saadaan DAS-ilmoitusten avulla. Nimensä mukaisesti Data Availability Statement antaa tietoa artikkeliin liittyvän tutkimusdatan saatavuudesta. Viime vuosina DAS-ilmoitukset ovat yleistyneet ja monet kustantajat ja rahoittajat ovat laatineet ohjeistuksia tai suosituksia niitä varten. Yliopistojen laatimat ohjeet ovat kuitenkin yhä harvinaisia. Selvityksen avulla pyrittiin hahmottamaan myös mahdollista tarvetta yliopiston omalle tuelle tai suositukselle DAS-ilmoitusten laatimiseen.

Selvitystä varten Tampereen yliopiston tutkimustietojärjestelmä TUNICRISistä otettiin yliopiston tutkijoiden julkaisut vuosilta 2020 ja 2021 ja ne ryhmiteltiin tiedekunnittain. Julkaisujoukkoa rajattiin ottamalla mukaan ainoastaan opetus- ja kulttuuriministeriön julkaisutiedonkeruun mukainen A1-julkaisutyyppi eli tieteellinen vertaisarvioitu alkuperäisartikkeli. Näistä 4922 A1-artikkelista otettiin 40 prosentin satunnaisotos tiedekunnittain ja vuosittain. Läpikäytyjä artikkeleita oli siten yhteensä 2085 eli hieman yli 42 prosenttia koko joukosta. Tiedekuntien väliset yhteisjulkaisut on otettu huomioon myöhemmin raportissa yliopistotason lukuja raportoitaessa.

DAS-ilmoitusten löytämiseksi artikkelit avattiin yksitellen ja niistä etsittiin mainintaa tutkimusdatasta ja sen saatavuudesta. Jokaisen artikkelin kohdalta kirjattiin ylös, oliko DAS ylipäänsä olemassa sekä missä kohdin artikkelia tai minkä otsikon alla se sijaitsi. Selvityksessä siis ei kartoitettu pelkästään Data Availability Statement -otsikon alla olevia tietoja, vaan datan saatavuuteen liittyviä teemoja etsittiin myös muiden otsikoiden alta. DAS-ilmoituksen sisällöstä kirjattiin ylös tieto datan saatavuudesta, sijainnista, avoimuudesta ja mahdollisista käytön rajoituksista. Selvityksessä saatiin näin DAS-ilmoitusten yleisyyden hahmottamisen lisäksi runsaasti arvokasta tietoa yliopiston tutkijoiden tuottamasta tutkimusdatasta.

Tässä selvityksessä tutkimusdatan määritelmän ulkopuolelle on jätetty tutkimuksessa tuotettu koodi. Lisäksi satunnaisotannan tuottamaa artikkelijoukkoa on jouduttu muokkaamaan sellaisissa tapauksissa, joissa artikkeli on ollut kielellä, jota analysoijat eivät ole ymmärtäneet. Lisäksi selvityksen ulkopuolelle jätettiin artikkelit, joista ei päästy lukemaan kokotekstiä. Yhteensä tällaisia sivuutettuja artikkeleita oli vain muutamia. Selvitys

keskittyy nimenomaan tutkijoiden näkökulmaan datan jakamisessa eli siihen, millaisia periaatteita ja rajoituksia tutkimukseen osallistuneet tahot datan käytölle ovat määrittäneet. Toisin sanoen tässä selvityksessä ei dokumentoida esimerkiksi Tilastokeskuksen tai THL:n kaltaisten toimijoiden datan jakamiseen liittyviä käytäntöjä.

Tästä artikkelista on jätetty pois Tampereen yliopiston tiedekuntakohtaiset luvut ja analyysit.

Tulokset

Selvityksen tulokset on ryhmitelty kolmeen alalukuun:

- Data Availability Statementien esiintyvyys
- datan saatavuus ja rajoitukset
- datan avoimuus.

Data Availability Statementien esiintyvyys

Analysoiduista 2085 artikkelista Data Availability Statement löytyi 540:stä eli noin 26 prosentista. Tarkastellessa DAS-ilmoitusten määriä vuosikohtaisesti, voidaan havaita selvää kasvua lyhyellä aikavälillä. Vuonna 2020 julkaistuista artikkeleista DAS löytyi noin 20 prosentista, kun taas vuoden 2021 artikkeleissa DAS-ilmoitus löytyi jo noin 31 prosentista.

Artikkeleiden DAS-ilmoitusten otsikoinnit ja sijainnit vaihtelivat suuresti. Analyysissa löydettiin yhteensä 39 erilaista otsikointia tai sijaintia datan saatavuustiedolle. Selvästi yleisin oli Data Availability Statement, joka oli otsikkona 230 (n. 43 %) julkaisussa. Seuraavaksi yleisimpiä otsikoita olivat Data Availability (147 kpl, n. 27 %) ja Availability of Data and Material(s) (66 kpl, n. 12 %). Suurin osa DAS-ilmoituksista oli otsikoitu ymmärrettävästi ja otsikoissa toistui useimmin sanat data ja availability erilaisissa variaatioissa. Joissain tapauksissa datan saatavuudesta kerrottiin varsinaisen DASin sijaan muissa artikkelin osissa, kuten Notes, Footnotes, Acknowledgements tai Supplementary Data. Kaikki analyysissa löydetyt DAS-sijainnit ovat tämän raportin liitteenä (liite 1).

DAS-ilmoitusten yleisyyttä tarkasteltiin eri muuttujien suhteen vuosien 2020 ja 2021 yhdistetystä julkaisujoukosta koko yliopiston tasolla. Käytettävissä olevat muuttujat olivat kustantajan kansainvälisyys, julkaisun kieli, julkaisukanavan Julkaisufoorumi-taso (JUFO) ja julkaisun avoimuus.

Kustantajan kansainvälisyys ja julkaisun kieli vaikuttivat selvästi Data Availability Statementien yleisyyteen. Kaikki löydetyt DAS-ilmoitukset löytyivät kansainvälisten kustantajien englanninkielisistä artikkeleista. Jos huomioidaan pelkät kansainvälisten kustantajien julkaisut, DAS löytyi noin 28 prosentissa otoksesta (taulukko 1). Englanninkielisistä julkaisuista DAS löytyi niin ikään noin 28 prosentissa (taulukko 2). On huomioitava, että rajauksena käytetyssä A1-julkaisuluokassa valtaosa kustantajista on Suomen ulkopuolelta ja siten julkaiseminen tapahtuu pääasiassa englannin kielellä. Otoksen artikkeleista vain hieman alle 8 prosenttia julkaistiin kotimaisten kustantajien kanavissa. Englanninkielisiä julkaisuja oli puolestaan lähes 93 prosenttia.

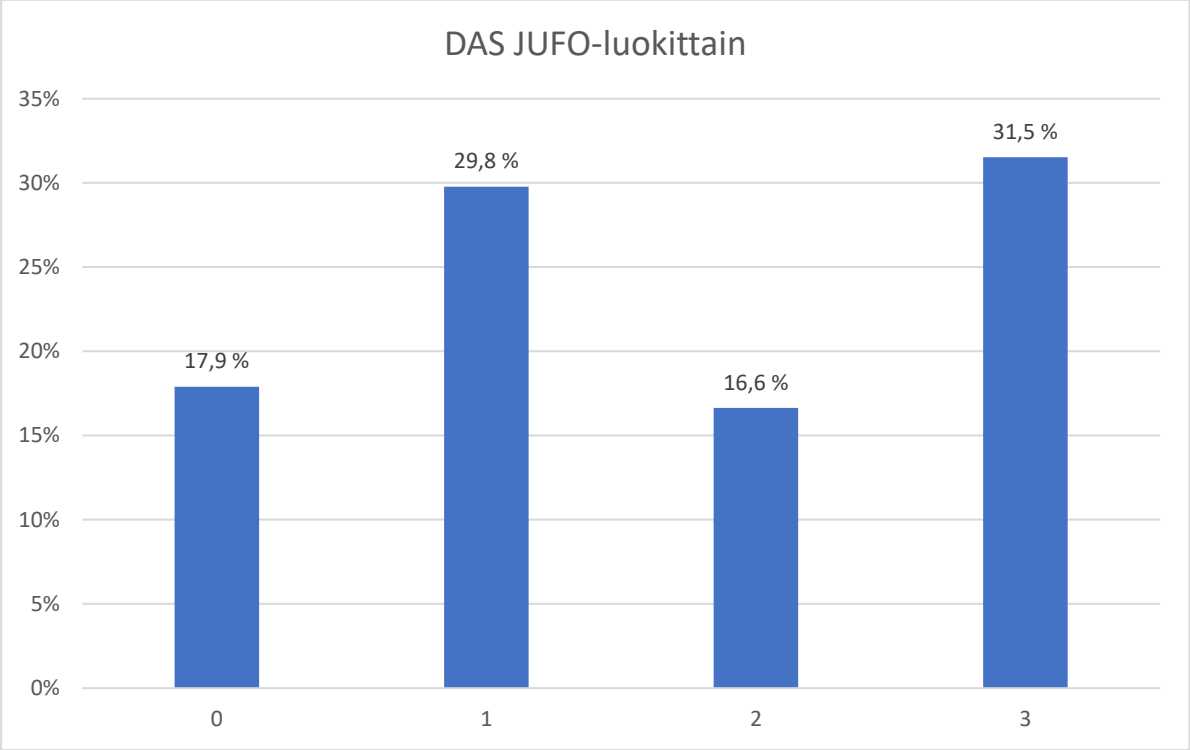
Taulukko 1. Data Availability Statementit kustantajan kansainvälisyyden perusteella.

Kustantaja	**DAS: ei**	**DAS: kyllä**	**Yhteensä**
Kansainvälinen	1381	540	1921
Kotimainen	164	0	164

Taulukko 2. Data Availability Statementit julkaisun kielen perusteella.

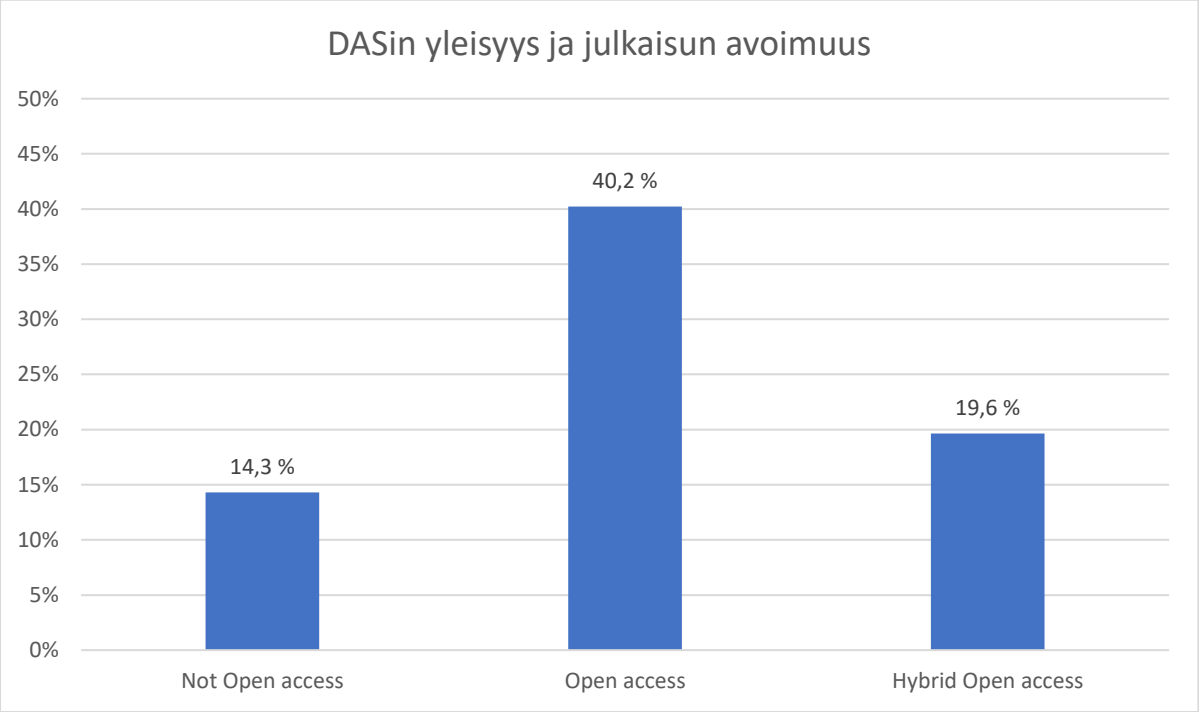
Julkaisun kieli	**DAS: ei**	**DAS: kyllä**	**Yhteensä**
Englanti	1390	540	1930
Suomi	137	0	137
Muut	18	0	18

Analyysi Data Availability Statementien ja JUFO-tasojen suhteesta tuotti hajanaisen tuloksen. Sekä JUFO1- että JUFO3-tasojen julkaisuista löytyi DAS selvästi useammin kuin niissä julkaisuissa, joissa JUFO-taso oli 0 tai 2 (kuvio 1). Vuosittaisessa tarkastelussa on myös merkittäviä eroja. Esimerkiksi vuonna 2020 JUFO1-tason julkaisuissa DAS löytyi noin joka viidennestä, kun vuonna 2021 vastaava luku oli jo yli 37 prosenttia. Koko aineistossa JUFO1-julkaisuja on 1253, JUFO2-julkaisuja 553 ja JUFO3-julkaisuja 184 kappaletta. JUFO-tasojen tuloksiin vaikuttavat ainakin julkaisukanavien kansainvälisyys ja avoimuus. Korkeimmalla JUFO-tasolla on otoksessa pelkästään kansainvälisiä julkaisuja, kun 1-tason julkaisuista kotimaisia on noin 9 prosenttia ja 2-tason julkaisuista noin 10 prosenttia. 1-tasolla julkaisujen avoimuus on selvästi muita luokkia yleisempää (noin 73 %).



Kuvio 1. Data Availability Statementien yleisyys JUFO-luokittain.

Tarkastellessa Data Availability Statementien yleisyyttä julkaisun avoimuuden näkökulmasta, voidaan löytää selviä eroja (kuvio 2). Avoimesti julkaistuista artikkeleista noin 40 prosentissa oli DAS. Hybridijulkaisuissa (n. 20 %) ja ei-avoimissa julkaisuissa (n. 14 %) luku oli huomattavasti alempi. Analyysin artikkeleista noin 39 prosenttia oli avoimia Open Access -julkaisuja, noin 32 prosenttia ei-avoimia ja noin 29 prosenttia hybridikanavissa julkaistuja artikkeleita. Tässä analyysissa ei huomioitu rinnakkaistallenteiden kautta syntyvää avoimuutta.



Kuvio 2. Data Availability Statementien yleisyys julkaisun avoimuuden perusteella.

Datan saatavuus ja rajoitukset

Yleisin DAS-ilmoituksissa kerrottu datan tallennuspaikka oli itse artikkeli (41 kpl). Seuraavaksi yleisimmin mainittuja tallennuspaikkoja tai arkistoja olivat Zenodo (12), Gene Expression Omnibus (11) & Github (9). Kokonaisuudessaan datan tallennuspaikkoja mainittiin 173 kertaa ja eri tallennuspaikkoja oli 64. Koko lista tallennuspaikoista on tämän artikkelin liitteenä (liite 2). Kaikkein yleisin datan jakamisen keino oli kuitenkin datan pyytäminen tutkimuksen tekijöiltä. Datan luvattiin olevan saatavilla tutkijalta pyydettäessä 288 artikkelissa eli noin 14 prosentissa analysoiduista.

Datan jakamiseen liittyviä rajoituksia mainittiin puolestaan 178 julkaisussa (taulukko 3). Selvästi yleisin mainittu rajoitus liittyi datan pyytämiseen tutkijalta. ”Reasonable request” oli vaatimuksena 134 julkaisussa. Sen lisäksi rajoituksina mainittiin esimerkiksi datan jakamiseen liittyviä sopimuksia ja erilaisia arviointi- ja lupaprosesseja.

Taulukko 3. Datan jakamisen rajoitukset.

Datan jakamisen rajoitukset	**Lukumäärä**
reasonable request	134
data-sharing agreement	9
application review	7
approval by the board/consortium/committee	6
permission from the data owner	6
qualified researcher	5
research criteria	5
ethical committee permission	4
GDPR	3
scientific collaboration	2
controlled access	2
terms and conditions	1
contract	1

153 artikkelin (noin 7,3 %) DAS-ilmoituksessa oli eksplisiittisesti mainittu, että data ei ole saatavilla. Voidaan perustellusti olettaa, että data ei todellisuudessa ole saatavilla huomattavasti useammassa julkaisussa, mutta tässä selvityksessä oltiin kiinnostuneita nimenomaan eksplisiittisistä maininnoista. Noin 69 prosentissa näistä maininnoista oli annettu jokin perustelu sille, ettei data ole saatavilla. Perusteluissa oli havaittavissa sekä vapaamuotoisia että standardinomaisia vastauksia. Nämä DASeissa annetut perustelut teemoiteltiin mielekkäiksi kokonaisuuksiksi analyysia varten (taulukko 4). Analyysissä pyrittiin noudattamaan mahdollisimman pitkälle perustelujen teksteissä käytettyjä muotoja. Useimmin toistuvat perustelut olivat datan sensitiivisuus (27 julkaisussa), lainsäädäntö (20 julkaisussa) ja datan omistajuus (17 julkaisussa). Näiden lisäksi aineistosta nousi esiin esimerkiksi tutkimuslupiin, etiikkaan ja yksityisyyteen liittyviä perusteluita. Perustelut ovat usein jossain määrin päällekkäisiä ja siten myös muodostetuissa teemoissa on päällekkäisyyttä. Tulokset antavat kuitenkin hyvän yleiskuvan ilmiöstä. Kyse on osittain myös tutkijan näkökulmasta. Esimerkiksi henkilötietoja

sisältävään dataan liittyvät rajoitteet voidaan mieltää muun muassa sensitiivisen datan, lainsäädännön tai eettisen arvioinnin teemoihin liittyviksi.

Taulukko 4. DAS-ilmoituksissa mainitut perustelut silloin, kun data ei ole saatavilla.

Data ei saatavilla: perustelut	**Lukumäärä**
sensitive data	27
legislation	20
data ownership	17
no permission	14
ethical reasons	14
privacy	12
GDPR	7
lack of informed consent	7
professional secrecy	6
technical or time limitations	3
confidentiality	3
IRB restrictions	3
data part of an ongoing study	1
hospital policy	1

Jotkut rajoitukset toistuivat artikkeleissa ikään kuin standardinomaisina vastauksina. Tällaisia olivat esimerkiksi perustelut, jotka oli muotoiltu muotoon ”privacy or ethical restrictions” tai ”technical or time limitations”. Analyysin ulkopuolelle jätettiin maininnat, jotka eivät sisältäneet informatiivista arvoa. Näitä ovat esimerkiksi perustelut, kuten ”data are not available” tai ”not applicable”.

Monessa DAS-ilmoituksessa datan jakamisen rajoitteissa oli erilaisia tasoja. Esimerkiksi koko datan jakamisesta voitiin mainita, että se ei ole mahdollista datan sensitiivisyyden takia tai siksi, että tutkittavilta ei ole kysytty lupaa datan jakamiseen. Monissa tällaisissa tapauksissa data oli kuitenkin saatavilla tutkijalta pyydettäessä, mahdollisten ehtojen täyttyessä, anonymisoidussa tai muutoin karsitussa muodossa.

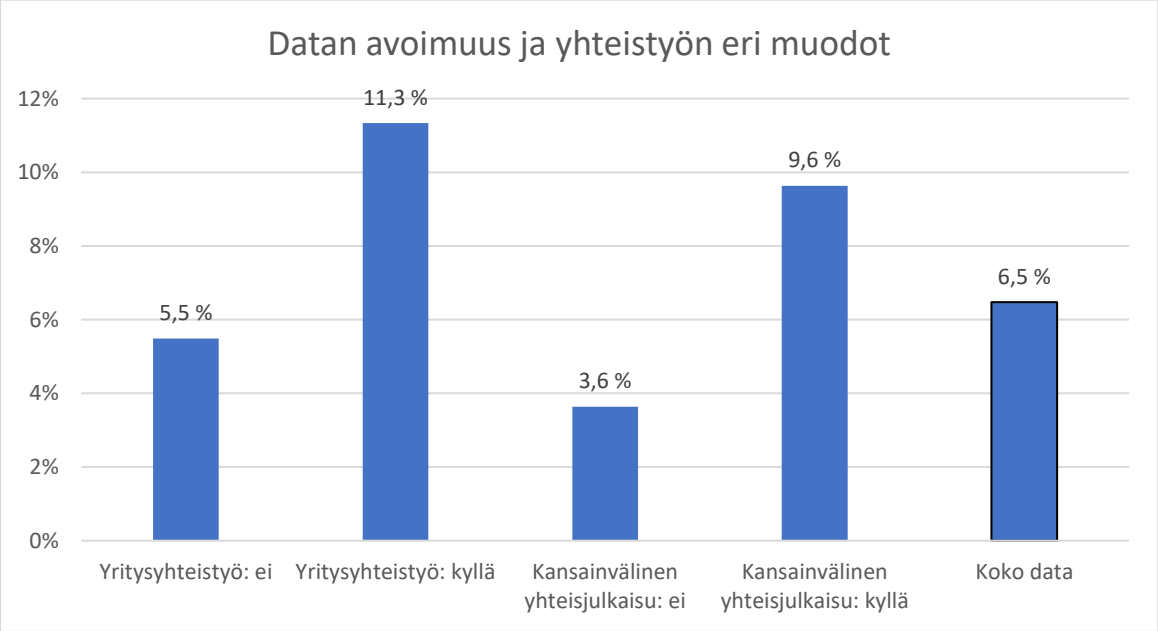
Datan avoimuus

Otoksen artikkeleista 135:ssa oli viittaus avoimeen tutkimusdataan. Se vastaa noin 6,5 prosenttia koko otoksesta. Tutkimusdatan avoimuutta tarkasteltiin yritysyhteistyön, kansainvälisen yhteistyön, julkaisun avoimuuden ja Julkaisufoorumi-tason kautta.

Yhteistyön eri muotoja ja avoimen datan yleisyyttä esitellään kuviossa 3. Yritysyhteistyönä tehdyissä artikkeleissa data oli avointa selvästi useammin (n. 11 %) kuin tutkimuksissa, jotka tehtiin ilman yritysyhteistyötä (n. 6 %). Koko otoksessa yritysyhteistyöjulkaisut olivat selvässä vähemmistössä, sillä niitä oli 353 kappaletta eli noin 17 % koko julkaisujoukosta.

Kansainvälisissä yhteisjulkaisuissa data oli avointa selvästi useammin (n. 10 %) kuin pelkästään suomalaisia organisaatioita sisältävissä julkaisuissa (n. 4 %). Kansainvälisiä yhteisjulkaisuja oli koko otoksesta noin 47 prosenttia. Kansainvälisellä yhteistyöllä tarkoitetaan tässä sitä, että vähintään yksi julkaisun kirjoittajista on affilioitunut ulkomaiseen yliopistoon.

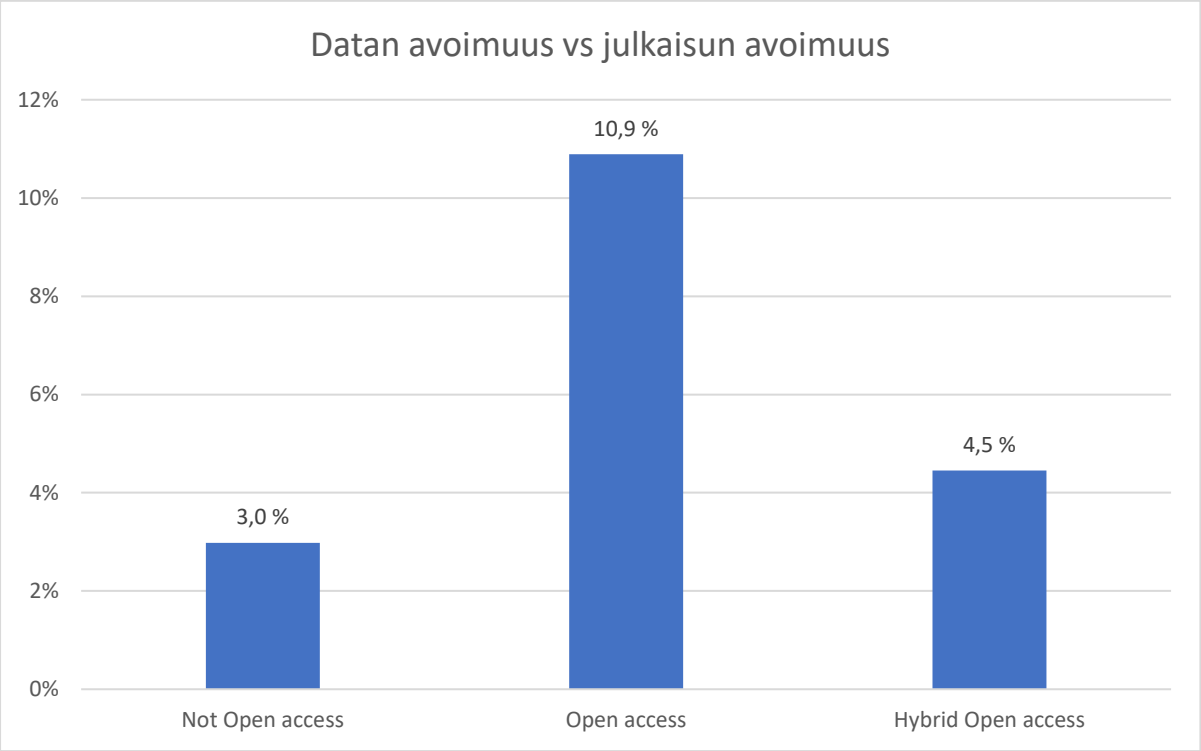
Osassa julkaisuja oli hyödynnetty useita datasettejä. Tässä selvityksessä datan avoimuutta tarkasteltiin julkaisukohtaisesti. Toisin sanoen, jos yksikin DASissa mainituista dataseteistä oli avoin, merkittiin kyseisen julkaisun kohdalla data avoimeksi.



Kuvio 3. Datan avoimuus yhteistyön näkökulmasta.

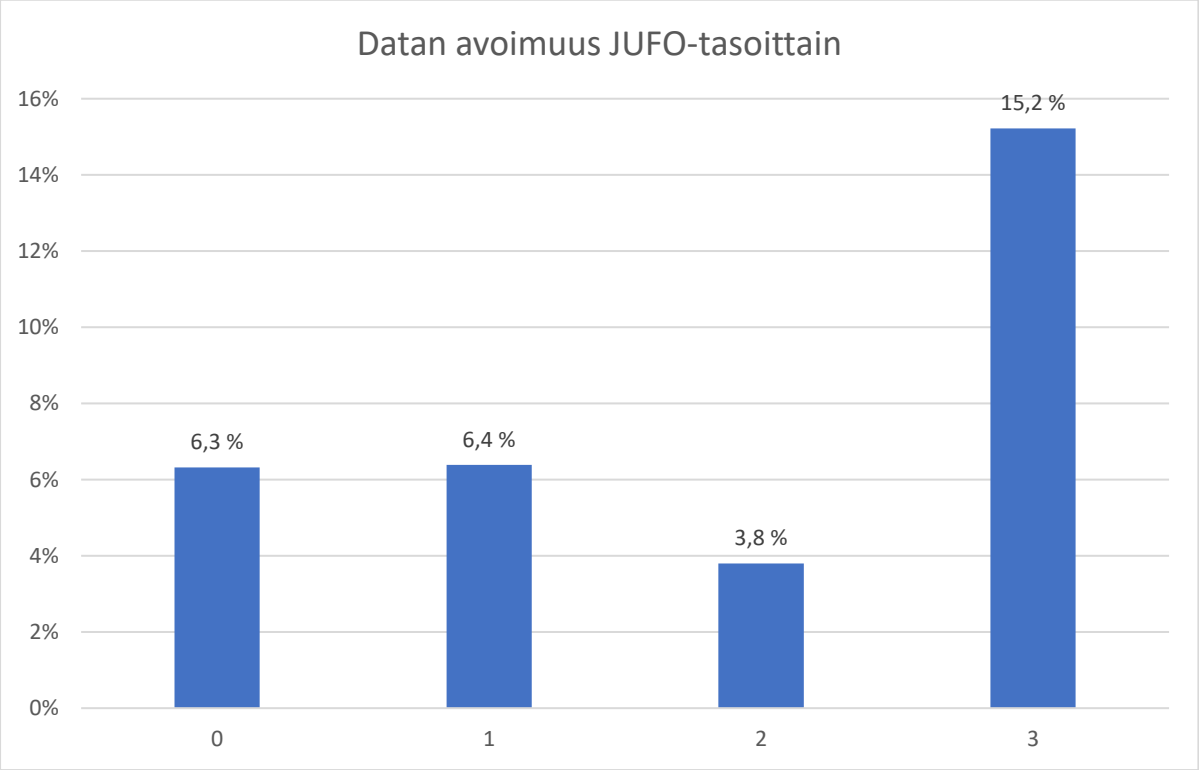
Datan avoimuutta analysoitiin myös julkaisujen avoimuuden näkökulmasta (kuvio 4). Avoin data oli selvästi yleisintä avointen Open Access -artikkelien kohdalla. Noin 11 prosenttia

artikkeleista, jotka oli julkaistu Open Access -kanavissa, sisälsivät viittauksen avoimeen dataan. Hybridijulkaistuissa artikkeleissa avointa tutkimusdataa oli vajaassa viidessä prosentissa ja täysin maksumuurin takana olevissa julkaisuissa noin kolmessa prosentissa.



Kuvio 4. Datan avoimuus julkaisun avoimuuden perusteella.

Datan avoimuuden vertailu julkaisukanavan JUFO-tasoon toi esiin mielenkiintoisen ilmiön (kuvio 5). Korkeimman JUFO-tason artikkeleista lähes joka kuudes sisälsi avointa tutkimusdataa. Tämä oli huomattavasti enemmän kuin muiden tasojen julkaisuissa. Etenkin 2-tason julkaisuissa avoimen datan määrä oli alhainen (3,8 %) suhteessa muihin. Tuloksia arvioidessa on huomattava, että JUFO3-tason artikkeleita on koko otoksessa selvästi vähemmän kuin 1- ja 2-tason julkaisuja. Esimerkiksi JUFO1-tason julkaisuista 80:ssa oli avointa tutkimusdataa, kun taas JUFO3-tason julkaisuista avointa tutkimusdataa löytyi 28:sta.



Kuvio 5. Datan avoimuus JUFO-tasoittain.

Johtopäätökset

Tutkimusdatasta kerrottiin DAS-ilmoituksissa moninaisin tavoin. Osa DASeista oli selkeitä ja pohdittuja kokonaisuuksia. Toisaalta joissain tiedot datasta ja sen saatavuudesta olivat pintapuolisia ja jopa ristiriitaisia. Joissain DASeissa puolestaan oli sellaista informaatiota, joka selvityksessä todettiin paikkaansa pitämättömäksi. Muutamissa DASeissa ongelmana olivat toimimattomat linkit, joiden takaa tutkimusdata olisi pitänyt löytyä. Ylipäänsä oli nähtävissä, että DASien määrä kasvoi selvityksen lyhyellä, kahden vuoden, aikajaksolla. Yleistymisestä huolimatta valtaosa artikkeleista ei silti sisältänyt DASia.

DAS-ilmoitusten sijainnista havaittiin, että tietyt otsikot ovat yleistyneet ja kustantajilla on tässä merkittävä rooli. Toisaalta koneluettavuutta haittaavat edelleen lukuisat eri otsikkovariaatiot, joiden alle datan saatavuudesta ja siihen liittyvistä rajoituksista kirjoitetaan.

Johtuen selvityksen rajaamisesta A1-luokan julkaisuihin oli otoksessa vain vähän suomalaisten kustantajien julkaisuja. Pienestä määrästä huolimatta voidaan todeta, että kotimaisilla kustantajilla on kehitettävää datan jakamiseen liittyvissä politiikoissaan ja ohjeistuksissaan, sillä yhdestäkään kotimaisesta julkaisusta ei löytynyt DASia.

Sekä Data Availability Statementien määrässä että datan avoimuudessa havaittiin julkaisun avoimuuden vaikutus. Open Access -julkaisuissa löytyi selvästi useammin DAS-ilmoituksia ja avointa dataa kuin maksumuurin takana olevissa julkaisuissa. Voitaneenkin ajatella, että avoimuus ruokkii avoimuutta. Toisin sanoen, avointa artikkelijulkaisemista korostavat toimijat kiinnittävät huomioita myös tutkimusdatan avoimuuteen ja datan saatavuuteen.

Käytetyin yksittäinen tallennuspaikka datalle oli hieman yllättäen artikkeli eikä mikään vakiintunut data-arkisto. Onkin relevanttia pohtia, onko datan julkaiseminen artikkelin ohessa kustantajan verkkosivuilla ylipäänsä kestävä tapa säilyttää tutkimusdataa. Toisaalta monessa DASissa mainittiin, että data on saatavilla pyydettäessä. Sekä artikkelissa julkaistun datan että pyydettäessä saatavilla olevan datan kohdalla voidaankin kysyä, tehdäänkö datalle tarvittavia kuratointitoimenpiteitä, jotta data on käyttökelpoista vuosienkin päästä.

Tutkijan säilyttämän datan suhteen nousee muitakin kysymyksiä. Voidaan esimerkiksi pohtia, kuinka kauan tutkijan on mahdollista säilyttää dataa siten, että se on todellisuudessa löydettävissä, onko tutkija itse saavutettavissa ja kuinka kauan lupaus datan jakamisesta pyydettäessä on voimassa. Datan jääminen tutkijan haltuun herättääkin merkittäviä kysymyksiä datan jakamisen FAIR-periaatteiden (Wilkinson et al. 2016) eli löydettävyyden, saavutettavuuden, yhteentoimivuuden ja uudelleenkäytettävyyden toteutumisesta.

Datan jakamiselle pyydettäessä mainittiin usein ehtoja, joista selvästi yleisin oli ”reasonable request”. Samankaltaisia tuloksia ovat saaneet esimerkiksi Gabelica, Bojcic ja Puljak (2022), jotka tutkivat Data Availability Statementeja ja niissä tehtyjä lupauksia. Analysoimistaan 3416 DAS-ilmoituksesta noin 52 prosentissa oli lupaus datan saamisesta pyydettäessä.

Ongelmaksi voi muodostua vaatimuksen ensimmäisen sanan merkitys. Reasonable voidaan kääntää suomeksi esimerkiksi kohtuulliseksi tai järkeväksi. Samalla se on kuitenkin sumea ja epäspesifi termi, jonka taakse voi piiloutua erilaisia merkityksiä. Datan uudelleen käytöstä kiinnostunut tutkija ei voikaan tietää, millaisia vaatimuksia tämä ”kohtuullinen pyyntö” todellisuudessa pitää sisällään.

Gabelican et al. (2022) tutkimuksessa huomattiin, että vain 14 prosenttia tutkijoista, jotka olivat luvanneet jakaa dataa pyydettäessä, edes vastasivat data-aiheiseen tiedusteluun. Sähköpostiviestiin vastanneista noin puolet jakoivat datan, joten datan jakaminen toteutui DASin mukaisesti vain noin seitsemässä prosentissa artikkeleista. Samankaltaisia havaintoja tekivät Tedersoo et al. (2021) tutkimuksessaan. He kartoittivat datan todellista saatavuutta DAS-ilmoitusten takaa ja havaitsivat useita ongelmia datan saatavuudessa tutkijalta. Ongelmakohtia nousi esiin muun muassa tutkijan tavoittamisessa, pitkässä kirjeenvaihdossa

tutkijan kanssa, riittävissä perusteluissa sekä kattavan ja käytettävän datan saamisessa. ”Reasonable request” yleisimpänä rajoituksena nostaakin esiin ajatuksen siitä, onko rajaus vaivaton tapa saavuttaa rahoittajan tai kustantajan vaatimukset datan saatavuudesta. Datan ei tarvitse tällöin olla heti julkaisukelpoista, kuvailtua tai arkistoitua, vaan datan avaamista voidaan siirtää ajassa eteenpäin lupaamalla toimittaa data tulevaisuudessa pyydettäessä.

Tutkimusdatan hallinnan ja jakamisen perusperiaatteisiin kuuluu se, että pyritään avoimuuteen, mutta jos dataa ei kyetä jakamaan, kerrotaan sille syyt. Otoksessa tämä toteutui vain harvoissa artikkeleissa. Datan jakamattomuudelle mainitut perustelut olivat pääasiassa datanhallinnan käytännön työssä ja kirjallisuudessa jo usein havaittuja syitä, kuten datan sensitiivisyys, lainsäädäntö ja omistusoikeudet.

Analyysissä havaittiin, että datan avoimuus on selvästi yleisempää sellaisissa artikkeleissa, joiden kirjoittajajoukkoon kuuluu henkilöitä joko yrityksistä tai Suomen rajojen ulkopuolelta. Yhtä lailla datan avoimuutta lisäsi merkittävästi se, jos artikkeli oli julkaistu avoimesti. Lisäksi korkeimman JUFO-tason lehdissä julkaistuissa artikkeleissa avointa dataa esiintyi huomattavasti useammin kuin muiden tasojen lehdissä. Kokonaisuutena avoimen datan määrä jäi otoksen artikkeleissa kuitenkin vähäiseksi.

Yksi selvityksen keskeisimmistä haasteista oli määrittää, mikä on tutkimusdataa. Data on hyvin tutkimus- ja tieteenalakohtaista ja näin sen muoto ja määrä vaihtelevat suuresti eri tutkimuksissa. Jotta olisi mahdollisuus ymmärtää tarkemmin artikkeliin liittyvää tutkimusdataa, tulisi ainakin jossain määrin ymmärtää kyseistä tutkimusta. Tästä syystä joissain tapauksissa oli vaikea hahmottaa, oliko mainittu data todella tutkimusdataa vai esimerkiksi jonkinlaista täydentävää dataa (supplementary data). Tällaista täydentäväksi nimettyä dataa esiintyi monissa selvityksen artikkeleissa. Joissain tapauksissa supplementary data -otsikon alle laitettu data arvioitiin tämän selvityksen tarkoittamaksi tutkimusdataksi. Useimmiten se oli kuitenkin nimenomaisesti tutkimusta täydentävää dataa, kuten dataa tiivistäviä tilasto-ohjelmien taulukoita. Joka tapauksessa, supplementary datan ja varsinaisen datan suhde on huomioon otettava asia, kun datan avaamisesta tai DAS-ilmoituksista tehdään tutkimusta.

Selvityksen datassa ei ollut käytössä julkaisukanavien kustantajatietoja. Analyysin ohessa tehtiin kuitenkin selvä havainto siitä, että kustantajien ohjeistuksilla ja vaatimuksilla on väliä. DAS esiintyikin järjestelmällisesti määrätyllä tavalla tiettyjen kustantajien julkaisuissa. Toisaalta myös rahoittajan vaatimuksilla on merkitystä. Esimerkiksi Wellcome Trust sisällytti vuonna 2016 avoimen tieteen politiikkaansa vaatimuksen datan saatavuuden maininnasta rahoittamissaan artikkeleissa. McIntoshin, Sumnerin ja Vitalen (2020) tekemässä selvityksessä havaittiin, että Data Availability Statementit olivat lisääntyneet merkittävästi

Wellcome Trustin rahoittamissa artikkeleissa. Vuonna 2016 DAS oli noin 22 prosentissa artikkeleista, kun vuonna 2019 sama luku oli noin 46 prosenttia. Huomioitavaa on kuitenkin myös se, että DAS-ilmoitusten merkittävästä yleistymisestä huolimatta suurin osa artikkeleista ei sisältänyt DASia, vaikka rahoittaja oli avoimen tieteen politiikassaan niin vaatinut. Toisaalta Gabelican et al. (2022) tutkimuksesta selviää, että BioMed Centralin kautta julkaisseet tutkijat noudattivat kustantajan vaatimuksia kuuliaisesti, kun tarkastelussa oli 3488 vuoden 2019 julkaisua. Otoksen julkaisuista lähes kaikki, 3416 (noin 98 %), sisälsi Data Availability Statementin.

Havainnot prosessista ja jatkosuunnitelmat

Selvitystyö oli antoisa oppimisprosessi. DAS-ilmoitusten etsintä ja niistä löytyneiden tietojen kirjaaminen manuaalisesti oli työmääränä iso. Vaikka valtaosasta läpikäytyjä artikkeleita puuttui DAS-ilmoitus, vaati niidenkin tarkistus merkittävästi työtä. Tietojen kartoittamista hidasti se, että DAS-ilmoitukset olivat vaihtelevissa kohdin artikkeleita, joskus tekstin seassa tai jopa liitetiedostoissa. Kiitos kirjaston harjoittelija Eemi Korkalle, joka tarjosi tärkeät lisäkäsiparit artikkelien läpikäyntiin.

Myös otoksesta saatujen tietojen analysointi ja laadunvarmistus vei runsaasti aikaa. Esimerkiksi datan jakamisen rajoituksiin liittyviä asioita kuvattiin hyvin moninaisesti artikkeleissa. Näiden yhtenäistäminen ja kokoaminen teemoihin vaati paljon työtä. Selvityksen aikana analysoiduista artikkeleista nousi esille muutamia uusia havaintoja, joiden mukaan DAS-ilmoituksista poimittavia tietoja tarkennettiin. Tämän vuoksi osa artikkeleista käytiin läpi kahteen kertaan.

Vastaavanlainen selvitys olisi mielenkiintoista toistaa uudemmalla pidemmän aikavälin datalla parin vuoden päästä. Tällöin olisi mahdollista saada tietoa muun muassa siitä, ovatko DAS-ilmoitukset yleistyneet ja ovatko niiden sisällöt kehittyneet. Toistaiseksi DAS-ilmoitusten vaihteleva sijainti artikkeleissa ja erilaiset otsikoinnit hankaloittavat DASien koneluettavuutta. Sen sijaan tiettyyn kustantajaan tai lehteen kohdistuvat selvitykset ovat jo nyt mahdollisia automatiikkaa hyödyntäen.

Toisaalta olisi mielenkiintoista selvittää, kuinka DAS-ilmoituksissa tehdyt lupaukset pitävät paikkaansa. Aiheesta on tehty muutamia tutkimuksia, joiden perusteella voi todeta, että DASissa annetun lupauksen ja todellisuuden välillä on vielä kuilu. Myös tässä selvityksessä havaittiin ristiriitoja DASiin kirjoitetun lupauksen ja todellisen tutkimusdatan saatavuuden välillä. Lisäksi joissain DASeissa oli havaittavissa tietynlaisia mallilauseiden tapaisia

ilmauksia, jotka toistuivat eri artikkeleissa. Tämä herättää ajatuksen siitä, onko datan jakamista tällöin pohdittu perusteellisesti vai vaan kopioitu valmiista mallista omia ajatuksia lähellä oleva lause.

Yliopistojen tutkimuksen tukipalveluille tämä raportti antaa mietittävää oman ympäristön ja DAS-ilmoituksiin liittyvän tuen tarpeista. Yhtäältä rahoittajat ja kustantajat laativat ohjeita ja esittävät vaatimuksia DASien tekoon, mutta toisaalta tutkijat eivät välttämättä hahmota DAS-ilmoituksen tärkeyttä tai millaisia tietoja siihen olisi hyvä kirjoittaa. Yliopistoissa tuleekin pohtia, voisiko omilla ohjeistuksilla tai suosituksilla lisätä tietoisuutta DAS-ilmoituksesta ja siten alleviivata sen merkitystä tutkimusdatan näkyvyyden lisäämisessä.

Ilmoitus datan saatavuudesta

Selvityksen data on julkaistu avoimesti CSV-tiedostona Zenodossa: https://doi.org/10.5281/zenodo.7564440. Datasetistä on poistettu Tampereen yliopiston tiedekuntatiedot.

Lähteet

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology, 150,* 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019

McIntosh, L., Sumner, J., & Vitale, C. (2020). Transparently Reported Research: An analysis of Wellcome-funding publications in 2016 and 2019. *RipetaReview*. https://doi.org/10.6084/m9.figshare.13810220.v1

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., . . . & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data, 8,* Artikkeli 192 (2021). https://doi.org/10.1038/s41597-021-00981-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3,* Artikkeli 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Liitteet

Liite 1. Data Availability Statementien sijainti artikkeleissa.

Sijainti	**Lukumäärä**	**Osuus**
Data Availability Statement	230	42,6 %
Data Availability	147	27,2 %
Availability of Data and Material(s)	66	12,2 %
Research Data for This Article	16	3,0 %
Data Sharing Statement	9	1,7 %
Data and Code Availability	9	1,7 %
Data Sharing	9	1,7 %
Acknowledgements	7	1,3 %
Supplementary Data	5	0,9 %
Notes	4	0,7 %
Footnote(s)	4	0,7 %
Methods	3	0,6 %
Data and Resource Availability	3	0,6 %
Related Research Data	2	0,4 %
Data Accessibility	2	0,4 %
Data Sharing and Declaration	2	0,4 %
Supplementary Material(s)	2	0,4 %
Funding, Data Sharing, and Potential Conflicts of Interests	1	0,2 %
Open Practices Statement	1	0,2 %
Availability of Supporting Data and Materials	1	0,2 %
Data, Code and Materials Availability Statement	1	0,2 %
Data Sharing and Data Accessibility	1	0,2 %
Dataset Repository and License	1	0,2 %

Materials and methods	1	0,2 %
Data-sharing Statement	1	0,2 %
Code & Datasets	1	0,2 %
Data and Materials Availability	1	0,2 %
Open Science and Reproducible Research	1	0,2 %
Results	1	0,2 %
Availability of Data Materials	1	0,2 %
Data Records	1	0,2 %
Software and Data Availability	1	0,2 %
Ethics, Data Sharing, Funding, and Potential Conflicts of Interest	1	0,2 %
Data	1	0,2 %
Data article	1	0,2 %
Declarations	1	0,2 %
Data Statement	1	0,2 %
Yhteensä	**540**	**100 %**

Liite 2. Artikkelien DAS-ilmoituksissa mainitut arkistot ja tallennuspaikat.

Arkisto/tallennuspaikka	**Lukumäärä**
Artikkeli	41
Zenodo	12
Gene Expression Omnibus	11
GitHub	9
Sequence Read Archive	7
FigShare	6
Mendeley Data	4
NIDDK Central Repository	4
Own website	4
PRIDE	4
European Genome-Phenome Archive	3
GenBank	3
gnomAD	3
LSHTM Data Compass	3
ArrayExpress	2
Dryad	2
FinBB	2
Global Health Data Exchange	2
GWAS	2
ICGC	2
OSF	2
Protein Data Bank	2
Synapse	2
BCAC	1

Bioproject	1
BitBucket	1
Cambridge Crystallographic Data Center	1
Cardiovascular Disease Knowledge	1
cBioPortal	1
CSC	1
dbGaP	1
EBAS Data Centre	1
Electron Microscopy Data Bank	1
Electron Microscopy Pilot Image Archive	1
ENA	1
ENCODE	1
ESS	1
Finnish Meteorological Institute	1
FSD	1
Genome	1
Gesis	1
GigaDB	1
GitLab	1
IDA	1
IEEE-dataport	1
ISSDA	1
MetaboLights	1
MHI-Humangenetics	1
NFBC	1
PGSCatalog	1
Protein Databank in Europe	1

RCSB	1
SASBDB	1
ScholarBank@NUS	1
SmartSMEAR	1
SNAP	1
TCGA database	1
THL	1
Tilastokeskus	1
UCSC Xena	1
UK Dataservice	1
United Nations Digital Library	1
University of Strathclyde data repository	1
Yareta	1