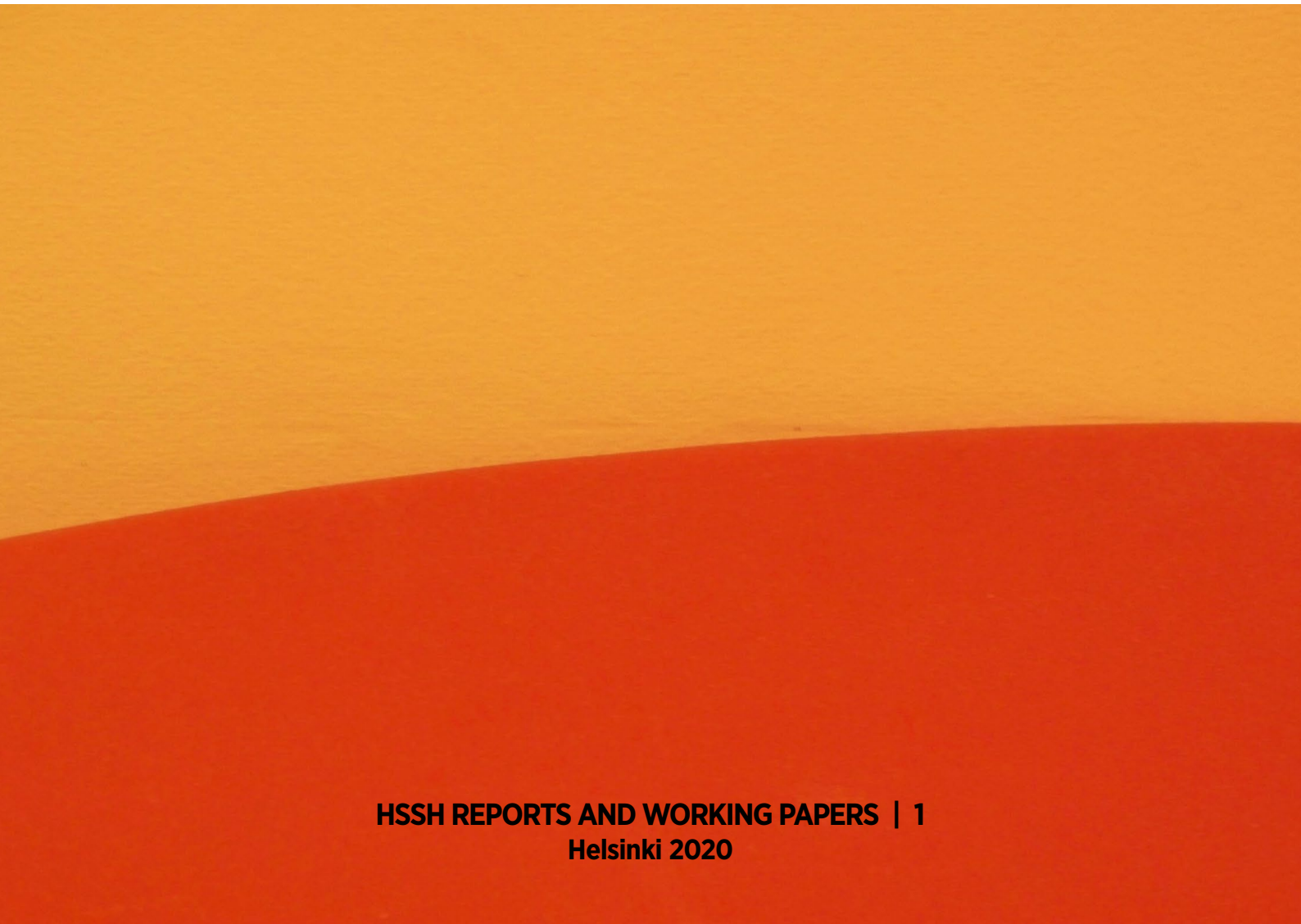
SOCIAL SCIENCES AND HUMANITIES

Katri Huutoniemi & Tuija Kajoskoski

THE NEED FOR RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES AND HUMANITIES AT THE UNIVERSITY OF HELSINKI

RESULTS OF A SURVEY FOR PRINCIPAL INVESTIGATORS



HSSH REPORTS AND WORKING PAPERS | 1
Helsinki 2020

Katri Huutoniemi & Tuija Kajoskoski

The Need for Research Infrastructures in the Social Sciences and Humanities at the University of Helsinki – Results of a survey for principal investigators

HSSH Reports and Working Papers 1
Helsinki Institute for Social Sciences and Humanities

Helsinki 2020

Taitto: Mari Karjalainen

ISBN 978-951-51-6533-6
ISSN 2670-3882

CONTENTS

EXECUTIVE SUMMARY

The planning group preparing the launch of the *Helsinki Institute for Social Sciences and Humanities* HSSH (1.1.2020) conducted a survey for principal investigators at the City Centre Campus in autumn 2019. The purpose of the survey was to establish an overview of the current situation and development needs of research infrastructures in the social sciences and humanities at the University of Helsinki. The term 'research infrastructure' was used according to the definition by the Academy of Finland and the Ministry of Education and Culture, i.e. as *a reserve of instruments, equipment, information networks, databases, materials and services enabling research at various stages. Research infrastructures may be based at a single location (single-sited), scattered across several sites (distributed), or provided via a virtual platform (virtual). They can also form mutually complementary wholes and networks*. The survey aimed at examining the need, usage and development of research infrastructures from the perspective of researchers, whereas infrastructure service providers and their perspectives were not surveyed.

We asked researchers what kind of data they use in their research work, and what kind of equipment and services they need to produce, acquire, process, use and share research data. Most questions contained both structured questions and open-ended questions, the latter allowing for more in-depth information. In addition, we asked for background information about the respondents and their research, their attitude towards sharing research data, their views on and examples of the development of research infrastructures in their own field, and their suggestions for persons participating in the development of research infrastructures in SSH at the University of Helsinki.

The total number of responses was 356, of which 190 were fully completed and 166 partially filled in. The number of principal investigators at the City Centre Campus is about 650, hence the response rate was 30—50% (depending on how it is calculated), which can be considered excellent. What is notable, besides the response activity, is that many open-ended questions were responded to by as many as 50—60 participants, and some of the responses were very thorough and contained useful points for the Institute's research infrastructure work. Responses were received from all target faculties and units (the Faculties of Arts, Social Sciences, Educational Sciences, Law, and Theology, as well as the Swedish School of Social Sciences and the Helsinki Collegium of Advanced Studies) and from a few other units. Researchers from the Faculty of Arts were clearly the most active in responding.

Physical research equipment relevant to the social sciences and humanities include *mobile devices* for collecting data (e.g. cameras, video cameras, audio recorders, scanning equipment, eye tracking devices, psychophysiological sensors, EEG devices) and processing data (e.g. headphones and loudspeakers for editing audio materials), as well as *special facilities* for producing interview, observational and experimental data (e.g. facilities for video interviews, silent space for observation studies, cubicles for test persons, facilities for simulation exercises). Some researchers also need actual *laboratories with related equipment* (e.g. laboratories for eye tracking research, laboratories suitable for studying phonetics, archaeology laboratories, brain imaging laboratories, laboratories with a driving simulator). The needs for mobile devices and research facilities are mostly specific to SSH

fields and thus should be met within the City Center Campus, whereas a suitable laboratory infrastructure is to some extent available at the Meilahti and Kumpula Campuses.

Services for the purchase, upkeep, storage and use of physical research equipment do not exist, but these remain the responsibility of individual research projects. The lack of coordination results in overlapping purchases, missing equipment, and poor condition of devices. The current situation also leads to inequality between researchers: existing research equipment either moves with researchers from project to project, or remains underused among individual researchers, while other researchers may often not have access to research equipment.

An important part of the research infrastructure is **IT for research**, which consists of both hardware and software services. On the hardware side, there are increasing needs in SSH fields especially for *storing and sharing research data* with personal identifiers, for *remotely accessing data* located elsewhere, and for computers with sufficient *computational capacity* (e.g. for editing video and audio data and for computation). Services for these needs are available from the UH Centre for Information Technology, from CSC, and from national research infrastructures, but finding and using the services is somewhat difficult, especially for new users. The IT solution consultation as well as the Helpdesk of the UH Centre for Information Technology help researchers with their IT problems, but they cannot support by finding discipline-specific solutions or services. It is difficult for researchers to know who to turn to for help in different matters, especially when they cannot properly define their problems in IT terms. Respondents would hope to see more ready-made solutions and guidance for storing and sharing SSH research data.

There are increasing needs for research software, too. Most researchers need *software for analysing* e.g. text, video, audio, image, observation and statistical data. *Software for editing* e.g. video and audio data is also needed. However, there are shortages in the availability and up-to-datedness of the software. As with physical research equipment, overlapping purchases are common in software licences; the purchase and circulation of licences would require a centralized service. In addition, researchers would need *training and support in the use of the software*, which is often not available. Along with the proliferation of digital and digitized data, there is an increasing need for software designed for the *automatized processing and recognition of data* (optical character recognition, speech recognition, automatic lemmatization etc.). The availability of these programs and support for using them should be improved.

Common investments in *developing software and methods* would also be needed. Of particular importance would be collaboration between researchers and program developers, but this is challenging and currently only a very small group of researchers are actually doing it.

The most frequent needs for services highlighted by the survey concern the **management and use of research data** – given especially the simultaneous transition to open science and the implementation of General Data Protection Regulation (GDPR). The majority of research data in SSH fields include personal identifiers, and a significant part of them are classified as sensitive personal data under GDPR. For the collection, processing, storing and sharing of these data, researchers need not only secure IT solutions (see above), but also legal services and research ethics counsel, which are insufficiently resourced by the university.

Legal services are needed in applying for research permissions and preparing research agreements, and increasingly also in determining the ownership of and access rights to

research materials. This concerns especially collaborative projects as well as the use of various derivative data sets. *Research ethics counselling* would be needed in designing data collection and processing, especially when dealing with new kinds of digital data (such as social media data) or data with strong personal identifiers (such as video data). Although legal services are available, they are concerned with reacting to individual cases (like in IT services), and furthermore, the services are constantly jammed. Research ethics counselling, in turn, is limited to the statements given by the ethical review board. According to the survey responses, researchers hope to get stronger guidance from the university, such as ready-made models for the resolution of juridical and ethical issues in different kinds of situations, as well as training for researchers.

Whenever not restricted by legal or ethical considerations, researchers are, in principle, willing to share their research data with others. However, *the cleansing, describing, curating and anonymization of data,* required for the reuse of data, are labour and service intensive. The UH research data services, coordinated by the library, offer a centralized channel for questions regarding research data management, but the university lacks the means to allow for the management of data according to FAIR data guidelines.

Important service providers in data management, which emerged from the survey responses, are the national research infrastructures Language Bank (service package of FIN-CLARIN) and the Finnish Social Science Archive (FSD), which however have limited resources to help researchers with the actual conduct of the work. Researchers would expect the university to offer them training on research management so as to make the reuse of data possible, as well as practical help to accomplish the work.

The results of the survey show that there is an evident need for developing the local research infrastructure in the social sciences and humanities at the University of Helsinki. The proliferation of digital resources and tools, combined with the movement towards open science, is profoundly changing the environment in which these fields operate. Research is increasingly based on multiple empirical source materials and new computational and mixed methods, and data are collected and used in heterogeneous research constellations. However, these developments have not permeated the practices of SSH fields as much as they could, and many researchers hope for a better-equipped research environment to harness the new opportunities effectively. This concerns not only the availability of modern technologies, tools, methods and data, but also shared rules and protocols for using them for different scholarly purposes.

The need for centrally managed solutions for collecting and managing research data is heightened by the new demands from legislation and regulation. One of the biggest challenges is the implementation of GDPR, which concerns a major part of research in SSH fields. Both the juridical interpretation of GDPR with regard to specific data sets, and its demands on the technologies for handling and transferring sensitive data, remain outside the scope of researchers' expertise. While national SSH research infrastructures have formulated guidelines and instruction, local services and solutions are urgently needed.

Based on the survey results, some of the most important actions for HSSH to take are the following:

- Mapping the current reserve of relevant research instruments, equipment, databases, materials, software and facilities as well as related services available for researchers either in UH or e.g. via national research infrastructures, and **establishing a portal or "one-shot-shop service"** for finding or accessing them. A related action is

collectively defining the concept of research infrastructure and identifying its relevant components in SSH fields.

- **Gathering together the existing (mobile) research equipment** and related software that are currently spread out across the campus, arranging their shared use and upkeep, and coordinating or centralizing the purchase of new equipment and licences.
- Strengthening the juridical, ethical and technical **guidance and services for research data management**. The existing support from the UH Research Services, the Centre for Information Technology and Helsinki University Library is not sufficient, but requires additional resources for addressing these problems from the perspective of SSH research. This involves e.g. centrally managed solutions for collecting, storing and sharing different types of data with personal identifiers, as well as the curation and anonymization of valuable data sets. Part of these needs can be covered by offering training to researchers, perhaps in collaboration with relevant national research infrastructures.
- Participating in the development of relevant **research data infrastructures**, i.e. digital infrastructures promoting data sharing and consumption, in collaboration with organizations that collect and keep data. Datasets of particular interest among SSH researchers are public collections (e.g. those of the National Library of Finland, archives, museums), registers and statistics produced by the public sector (e.g. Statistics Finland, government research institutes, cities), and media and social media material. Workable data infrastructure includes technology related to data usage (e.g. centralized access), processes, organization and social networks.
- Strengthening the UH **research environment for data-intensive SSH research**. This includes the arrangement of up-to-date equipment, tools and services, as well as the maintenance of technical and methodological expertise needed for the extensive use and development of the infrastructure for specified research purposes. This can mean, for example, permanent "staff scientist" positions, intensive collaboration with data scientists in e.g. the Faculty of Science, and a training programme for interdisciplinary mixed methodologies. Data-intensive SSH research fields that emerged from the survey include the following:

    * **Digital humanities**, including e.g. computational linguistics
    * **Social data science**, including e.g. survey- and register-based research and computational social sciences
    * **Audio-visual research,** including e.g. linguistic, anthropological, ethnographic and interaction research based on video, audio, image or multimodal data
    * **Experimental research**, including e.g. social, behavioural, cognitive and educational research based on laboratory or field experiments

TIIVISTELMÄ (SUMMARY IN FINNISH)

Helsingin yliopiston humanistis-yhteiskuntatieteellisen instituutin (Helsinki Institute for Social Sciences and Humanities, HSSH) toiminnan käynnistämistä (1.1.2020) suunnitteleva työryhmä teki syksyllä 2019 keskuskampuksen vastuullisille tutkijoille (*principal investigators*) kyselyn, jonka tarkoituksena oli luoda kokonaiskuva humanististen ja yhteiskuntatieteellisten alojen tutkimusinfrastruktuurin nykytilanteesta ja kehittämistarpeista Helsingin yliopistossa. Tutkimusinfrastruktuurilla tarkoitettiin OKM:n ja Suomen Akatemian määritelmän mukaisesti *välineitä, laitteistoja, tietoverkkoja, tietokantoja ja aineistoja sekä palveluita, jotka mahdollistavat eri vaiheissa tapahtuvan tutkimuksen. Tutkimusinfrastruktuurit voivat olla keskitettyjä, hajautettuja tai virtuaalisia ja ne voivat muodostaa toisiaan täydentäviä kokonaisuuksia ja verkostoja.* Kyselyllä pyrittiin selvittämään tutkimusinfrastuktuuripalvelujen tarvetta, käyttöä ja kehittämistä nimenomaan tutkijoiden näkökulmasta, kun taas palveluja tuottavien toimijoiden kartoitus ja näkökulmat jätettiin kyselyn ulkopuolelle.

Kyselyllä kartoitettiin, millaista tutkimusaineistoa tutkijat käyttävät tutkimustyössään, ja minkälaisia välineitä ja palveluita he tarvitsevat aineistojen tuottamiseen, hankintaan, käsittelyyn, käyttöön ja jakamiseen. Suurin osa kysymyksistä sisälsi sekä määrämuotoisia monivalintakysymyksiä että niitä tarkentavia avokysymyksiä. Lisäksi kysyttiin taustatietoja vastaajista ja heidän tutkimuksestaan, asennoitumista tutkimusaineistojen jakamiseen, näkemyksiä ja esimerkkejä oman alansa tutkimusinfrastruktuurien kehittämisestä, sekä ehdotuksia infrastruktuurien kehittämiseen osallistuvista henkilöistä.

Vastauksia kertyi yhteensä 356, joista 190 oli loppuun asti täytettyjä (valmiita) ja 166 vain osittain täytettyjä (keskeneräisiä). Vastuullisia tutkijoita on kampuksella noin 650, joten vastausprosentti oli laskentatavasta riippuen noin 30–50 %, mitä voidaan joka tapauksessa pitää kyselytutkimuksessa erinomaisena. Vastausaktiivisuuden lisäksi huomattavaa on, että moniin avoimiin kysymyksiin saatiin jopa 50–90 vapaamuotoista vastausta, joiden mukana oli hyvinkin seikkaperäisiä ja instituutin infrastruktuurityön kannalta hyödyllisiä huomioita. Vastaajia oli kaikista keskustakampuksen yksiköistä (humanistinen, valtiotieteellinen, kasvatustieteellinen, oikeustieteellinen ja teologinen tiedekunta, sekä Svenska social- och kommunalhögskolan ja tutkijakollegium) sekä muutamasta muusta yksiköstä; selvästi aktiivisimmin vastasivat humanistisen tiedekunnan tutkijat.

SSH-alojen tutkimuksessa tarvittavaa **fyysistä välineistöä** ovat erityisesti *mobiilit laitteet* aineistojen keräämistä (esim. kamerat ja videokamerat, diginauhurit, skannauslaitteistot, katseenseurantalaitteet, ihon sähkönjohtavuusmittarit, älysormukset, EEG-laitteet) ja käsittelyä (esim. kuulokkeet ja kaiuttimet äänieditointiin) varten sekä *erityiset tilat* haastattelu-, havainto- ja testiaineistojen tuottamiseen (esim. videohaastattelutilat, hiljainen tila havaintokokeiden tekemiseen, testihenkilöiden kubiikkelit, tilannehuone simulaatioharjoituksiin). Osa tutkijoista tarvitsee myös varsinaisia *laboratorioita tutkimuslaitteineen* (esim. katseenseurantalaboratorio, ääntämisen tutkimiseen soveltuvat laboratorio-olot, arkeologian laboratorio, aivokuvantamislaboratorio, ajosimulaatiokokeet). Mobiililaitteita ja tiloja koskevat tarpeet ovat SSH-aloilla omanlaisiaan ja niihin kaivattaisiin lisäresursseja nimenomaan keskustakampuksella, kun taas soveltuvaa laboratorioinfrastruktuuria on jossain määrin tarjolla Meilahden ja Kumpulan kampuksilla.

Fyysisen tutkimusvälineistön *hankintaan, ylläpitoon ja käyttöön liittyviä palveluita* ei ole tai ne ovat niin hajallaan, että varsinaisesta tutkimusinfrastruktuurista ei oikeastaan voida puhua. Jonkin verran teknistä tukea välineistön käyttöön saadaan esim. lääketieteellisestä tiedekunnasta, tutkimusavustajilta (yksiköissä joilla näitä on) ja toisilta tutkijoilta. Sen sijaan välineistön hankinta, ylläpito ja säilytys ovat kokonaan yksittäisten tutkimushankkeiden vastuulla. Näihin liittyvä koordinaatio ja keskitetyt palvelut puuttuvat, joten hankintoja tehdään päällekkäin, olemassa olevia laitteita ei löydetä, ja laitteiden kunto on usein huono. Nykyisellään tilanne on myös tutkijoita epätasa-arvoistava: Käytössä olevat laitteet joko siirtyvät tutkijoiden mukana hankkeesta toiseen, tai ovat alikäytettyinä yksittäisten tutkijoiden keskuudessa, kun taas osalla tutkijoista ei ole pääsyä laitteisiin.

Tärkeän osan tutkimusinfrastruktuuria muodostavat tutkimuksen **tietotekniikkapalvelut,** jotka pitävät sisällään sekä ’hardware’ että ’software’ resursseja. Hardware -puolella SSH-aloilla on kasvavia tarpeita erityisesti henkilötietoja sisältävän tutkimusdatan *tallennus- ja säilytyskapasiteetille*, tutkimusorganisaatioiden *yhteiskäytössä olevien henkilötietoja sisältävien aineistojen säilyttämiselle*, muualla sijaitsevien aineistojen (tietoturvallisille) *etäkäyttöyhteyksille*, sekä riittävän *tehokkaille tietokoneille* (esim. video- ja äänieditointiin, laskentaan). Näihin on saatavissa palveluita yliopiston tietotekniikkakeskuksen sekä CSC:n ja kansallisten tutkimusinfrastruktuurien kautta, mutta palvelujen löytäminen ja käyttö on erityisesti uusille käyttäjille hankalaa. Yliopiston tietotekniikkakeskuksen ratkaisukonsultit ja Helpdesk auttavat tutkijoita tietotekniikkaongelmissa, mutta heiltä ei saa tukea tieteenalakohtaisten ratkaisujen tai palvelujen löytämiseen. Tutkijoiden on vaikea tietää, kenen puoleen kääntyä missäkin asiassa, erityisesti jos ei osata määritellä ongelmia tietotekniikka-asiantuntijoiden kielellä. Yliopistolta toivottaisiin keskitettyjä tallennus-, säilytys- ja tietoliikenneratkaisuja SSH-alojen tutkimusaineistoille sekä niiden käyttöä koskevaa ohjausta.

Myös software -puolella on lisääntyviä tarpeita. Enemmistö tutkijoista tarvitsee *analyysiohjelmistoja* mm. teksti-, video-, audio-, kuva-, havainto- ja tilastoaineistojen analysointiin. Esimerkiksi video- ja audiomateriaalin käsittelyyn tarvitaan lisäksi erityisiä *editointiohjelmia*. Ohjelmistojen saatavuudessa ja ajantasaisuudessa on puutteita, ja niiden hankinta ja lisenssien kierrätys kaipaisivat keskitettyä palvelua. Kuten fyysisten välineiden osalta, myös ohjelmistolisenssien osalta tehdään päällekkäisiä hankintoja. Monien ohjelmien käytössä tarvittaisiin koulutusta ja tukea, mutta tätä ei useinkaan ole tarjolla. Digitaalisten ja digitoitujen aineistojen lisääntyessä on kasvavaa tarvetta myös erilaisille aineistojen *automatisoituun käsittelyyn ja luentaan* tarkoitetuille ohjelmistoille (OCR-luenta, puheentunnistus, automaattilemmaus, jne.). Näiden ohjelmistojen saatavuutta ja käyttötukea pitäisi parantaa.

Yhteisiä panostuksia tarvittaisiin myös *ohjelmistojen ja menetelmien kehittämiseen*. Erityisesti tarvitaan SSH-alojen tutkijoiden ja ohjelmistokehittäjien välistä yhteistyötä, mikä on haastavaa ja toistaiseksi varsin pienen tutkijajoukon harteilla.

Yleisimmät kyselyssä esiin nousseet palvelutarpeet liittyvät **tutkimusaineistojen hallintaan ja käyttöön** – erityisesti tilanteessa, jossa ollaan siirtymässä avoimeen tieteeseen mutta samalla joudutaan täyttämään tietosuoja-asetuksen vaatimukset. Valtaosa SSH-alojen tutkimusaineistoista sisältää henkilötietoja, joista merkittävä osa on tietosuoja-asetuksen määritelmän mukaan ”sensitiivistä” henkilötietoa. Tällaisten aineistojen keräämiseen, käsittelyyn, tallentamiseen ja jakamiseen tutkijat tarvitsevat

paitsi tietoturvallisia tietotekniikkaratkaisuja (ks. edellä), myös tutkimusjuridisia palveluita ja tutkimuseettistä neuvontaa, joiden resursointi yliopistolla on riittämätöntä.

Juridista apua tarvitaan tutkimuslupien ja –sopimusten laadinnassa, sekä lisääntyvässä määrin aineistojen omistus-, käyttö- ja hallintaoikeuksia koskevien kysymysten ratkaisemisessa. Tämä koskee erityisesti yhteistyöhankkeita sekä käytettäessä erilaisia johdannaisaineistoja. *Tutkimuseettistä neuvontaa* tarvittaisiin tutkimusasetelmien sekä aineistonkeruun ja –käsittelyn suunnittelussa, erityisesti vahvasti tunnisteellisia aineistoja (esim. videoaineistot) tai uudentyyppisisiä digitaalisia aineistoja (esim. sosiaalisen median aineistot) käytettäessä. Vaikka juridista neuvontaa on jossain määrin tarjolla, se on tietotekniikkapalvelujen tapaan reagointia yksittäistapauksiin, ja lisäksi palvelut ovat jatkuvasti ruuhkautuneet. Tutkimuseettiset palvelut puolestaan rajoittuvat ihmistieteiden eettisen ennakkoarvioinnin toimikunnan antamiin lausuntoihin. Kyselytulosten perusteella tutkijat toivoisivat yliopistolta selkeämpää ohjausta, kuten valmiita toimintamalleja tutkimusjuridisten ja -eettisten kysymysten ratkaisemiseen erityyppisissä tilanteissa, sekä tutkijoille suunnattua koulutusta.

Juridisten ja eettisten rajoitusten salliessa tutkijat ovat halukkaita jakamaan aineistojaan muiden kanssa, mutta aineistojen *siivoamiseen, kuvailuun, kuratointiin* sekä tunnisteellisten aineistojen *anonymisointiin* liittyvä työmäärä ja palvelutarve on suuri. Yliopiston kirjaston koordinoima datatukiverkosto tarjoaa keskitetyn palvelukanavan aineistonhallintaan liittyvissä kysymyksissä, mutta yliopisto ei pysty tarjoamaan suurta osaa aineistonhallintasuunnitelmien toteuttamiseen tarvittavista palveluita. Palvelujen tarjoajina kyselyvastauksissa nousivat esille erityisesti kansalliset tutkimusinfrastruktuurit Kielipankki (FIN-CLARINin palvelukokonaisuus) ja Yhteiskuntatieteellinen tietoarkisto (FSD), joilla on kuitenkin rajalliset resurssit auttaa tutkijoita varsinaisen työn tekemisessä. Yliopistolta toivottiin paitsi tutkijoille suunnattua koulutusta aineistojen käsittelyyn jatkokäytön mahdollistavalla tavalla tavalla, myös konkreettista apua työn tekemiseen.

Kyselyn tulokset osoittavat, että Helsingin yliopiston SSH-aloilla on selvää tarvetta paikallisen tutkimusinfrastruktuurin kehittämiseen. Tutkimuksen digitaalisten resurssien ja työkalujen lisääntyminen sekä suuntaus kohti avointa tiedettä ovat perustavanlaatuisella tavalla muuttamassa näiden alojen toimintaympäristöä ja käytäntöjä. Tutkimus perustuu yhä useammin moniin eri aineistolähteisiin ja uusien laskennallisten menetelmien käyttöön tai menetelmien yhdistelyyn, ja aineistoja kerätään ja käytetään monenlaisissa yhteistyörakenteissa. Muutokset eivät kuitenkaan ole levinneet SSH-alojen käytäntöihin niin laajasti kuin olisi ollut mahdollista, ja moni tutkija toivoo tutkimusympäristöltä parempaa tukea uusien mahdollisuuksien hyödyntämiseksi. Tämä ei koske ainoastaan uusien teknologioiden, työkalujen, menetelmien ja aineistojen saatavuutta, vaan myös yhteisiä säännöstöjä ja toimintatapoja niiden käyttämiseksi erilaisiin tieteellisiin tarkoituksiin.

Tarve keskitetyille ratkaisuille tutkimusaineistojen keräämisessä ja käsittelyssä on voimistunut lainsäädännön ja muun ohjauksen tuomien uusien velvoitteiden takia. Yksi suurimmista haasteista on EU:n tietosuoja-asetuksen toimeenpano, joka koskee valtaosaa SSH-alojen tutkimusaineistoista. Asetuksen juridinen tulkinta suhteessa erilaisiin aineistoihin sekä sen asettamat vaatimukset sensitiivisen datan käsittelyyn ja siirtämiseen käytettävälle teknologialle jäävät tutkijoiden osaamisen ulkopuolelle. SSH-alojen kansalliset tutkimusinfrastruktuuritoimijat ovat laatineet ohjeita ja suosituksia, mutta paikallisia palveluita ja ratkaisuja tarvitaan kipeästi.

Samalla lainsäädäntö ja muu kansallinen ja kansainvälinen ohjaus ovat tuoneet tutkimusaineistojen keruuseen ja käsittelyyn uusia velvoitteita, joiden täyttämiseksi tarvitaan SSH-aloille vahvempaa ohjausta ja palveluinfrastruktuuria. Helsingin yliopiston tutkimuspalvelut sekä erilliset palvelulaitokset (tietotekniikkakeskus ja kirjastot) tarvitsevat lisäresursseja ja paneutumista SSH-alojen tutkimustoiminnan erityiskysymyksiin.

Kyselyn perusteella tärkeimpiä toimenpiteitä Helsingin yliopiston SSH-alojen tutkimusinfrastruktuurin kehittämisessä ovat:

- Tutkijoiden käytettävissä (yliopistossa tai esim. kansallisissa tutkimusinfrastruktuureissa) olevien tutkimusvälineiden, -laitteistojen, -aineistojen, -ohjelmistojen ja –tilojen sekä niihin liittyvien palvelujen kartoittaminen ja **yhteisen portaalin tai ”yhden luukun palvelun”** luominen niiden löytämiseksi. Tähän liittyy myös tutkimusinfrastruktuurin käsitteen yhteinen määrittely ja siihen kuuluvien asioiden tunnistaminen SSH-aloilla.
- Keskustakampuksella olevien **tutkimusvälineiden, laitteiden ja niiden käyttöön liittyvien ohjelmistojen kokoaminen yhteen**, niiden yhteiskäytön ja ylläpidon järjestäminen, sekä uusien hankintojen koordinointi tai keskittäminen.
- **Tutkimusaineistojen hallintaan** liittyvän juridisen, eettisen ja teknisen ohjauksen sekä palvelujen vahvistaminen. Helsingin yliopiston tutkimuspalvelut sekä erilliset palvelulaitokset (TIKE ja kirjastot) tarvitsevat lisäresursseja ja paneutumista SSH-alojen erityiskysymyksiin. Tähän liittyy mm. henkilötietoja sisältävien aineistojen keskitetyt keruu-, tallennus-, säilytys- ja jakamisratkaisut sekä arvokkaiden aineistojen kuratoinnin ja anonymisoinnin palvelut. Osaan palvelutarpeesta voidaan vastata tutkijoille suunnatulla koulutuksella, mahdollisesti yhteistyössä relevanttien kansallisten tutkimusinfrastruktuurien kanssa.
- **Tutkimusdatainfrastruktuurien** eli (olemassa olevan) datan jakamiseen ja käyttämiseen tarkoitettujen digitaalisten infrastruktuurien vahvistaminen yhteistyössä aineistoja säilyttävien organisaatioiden kanssa. SSH-aloille tärkeitä, yliopiston ulkopuolella olevia aineistoja ovat mm. julkiset kokoelmat (esim. Kansalliskirjasto, arkistot, museot), viranomaistoiminnassa syntyvät rekisteri- ja tilastoaineistot (esim. Tilastokeskus, valtion tutkimuslaitokset, kaupungit) sekä median ja sosiaalisen median aineistot. Toimiva tutkimusdatainfrastruktuuri pitää sisällään datan käyttöön liittyvän teknologian (mm. keskitetty pääsy), prosessit, organisaation ja sosiaaliset verkostot.
- Ns. **dataintensiivisen tutkimuksen kehitystä tukevan tutkimusympäristön** vahvistaminen. Tähän sisältyy paitsi ajanmukaisten välineiden ja palvelujen järjestäminen, myös näiden tutkimuskäyttöön ja kehittämiseen tarvittavan teknisen ja menetelmäosaamisen turvaaminen. Tämä voi tarkoittaa esim. ”staff scientist” -tyyppisten pysyvien tehtävien avaamista, intensiivistä yhteistyötä mm. matemaattis-luonnontieteellisen tiedekunnan datatieteilijöiden kanssa sekä tieteidenvälisen, mixed method -menetelmäkoulutuksen järjestämistä. Kyselyssä esiin nousseita data-intensiivisiä SSH-tutkimusaloja ovat:
    * **Digitaaliset ihmistieteet**, ml. kieliteknologia
    * **Yhteiskuntadatatiede**, ml. laajoihin rekisteri- ja kyselyaineistoihin perustuvautkimus sekä laskennallinen yhteiskuntatiede
    * **Audiovisuaalisiin aineistoihin** perustuva tutkimus, ml. kielitieteellinen, antropologinen, etnografinen ja vuorovaikutukseen kohdistuva video-, ääni- ja kuva-aineistoja hyödyntävä tutkimus

* **Kokeellinen tutkimus**, ml. sosiaali-, käyttäytymis- ja kognitiotieteellinen sekä oppimistutkimus joka perustuu laboratorio- tai kenttäkokeisiin

1 INTRODUCTION

In February 2019, the Rector of the University of Helsinki decided to establish a new unit, the *Helsinki Institute for Social Sciences and Humanities* (HSSH), which is a joint enterprise of all units operating in social sciences and humanities (SSH) at the City Centre Campus: the Faculty of Arts, the Faculty of Social Sciences, the Faculty of Educational Sciences, the Faculty of Law, the Faculty of Theology, the Swedish School of Social Sciences, the Helsinki Collegium for Advanced Study, and the Doctoral School for Social Sciences and Humanities. As stated in the Rector's decision, the mission of the Institute is to: (1) strengthen basic research and research infrastructures in SSH, (2) expand research collaboration across SSH to create new multi- and cross-disciplinary research projects; (3) support SSH research and teaching that utilize digital technology; (4) improve the scientific and societal impact as well as the national and international visibility of research in SSH; and (5) enhance opportunities for acquiring external research funding for SSH, particularly from international sources.

One of the key duties of HSSH relates to the coordination, maintenance and development of research infrastructures for SSH at the University of Helsinki. Research infrastructures exist in a large variety of forms and structures and make an important contribution to the advancement of knowledge in all scientific areas. In many SSH disciplines, however, research infrastructures have traditionally not played a significant role, with the exception of libraries, archives and other collections. Only recently, the implementation of digital technologies, methods and protocols, together with the increasing demand for interdisciplinary integration and efficiency in scientific knowledge production, have heightened the importance of research infrastructures in SSH, too.

In order to establish an overview of the current situation and development needs for research infrastructures in SSH at the University of Helsinki, the planning group preparing the launch of HSSH operations conducted a survey for principal investigators at the University of Helsinki City Centre Campus in autumn 2019. The purpose of the survey was to gather information that would help in planning and implementing the research infrastructure mission of HSSH. The survey was conducted with the online open-source survey software LimeSurvey.

In the survey, the term 'research infrastructure' was used in accordance with the definition of the Academy of Finland and the Ministry of Education and Culture:

> *Research infrastructures form a reserve of research facilities, equipment, materials and services. As such, they enable research and development at various stages of innovation, while supporting organized research, researcher training and teaching. They also support and develop research and innovation capacity. Research infrastructures consist of equipment, knowledge networks, databases, multidisciplinary research centres, research stations, collections, libraries and related user services, where these are fundamental to research. Research infrastructures can be centralized, that is, based in a single location. They can also be distributed or virtual, and can form mutually complementary wholes and networks.*

The target group of the survey consisted of principal investigators at the City Centre Campus. According to the UH's definition, a principal investigator at the University of Helsinki is typically a person who (i) independently steers and leads research, (ii) has completed an applicable doctoral degree and become qualified as an independent researcher, (iii) has access to the necessary resources (facilities, funding, equipment) for independent research, (iv) supervises doctoral students and/or mentors postdoctoral researchers as well as (in applicable research fields) leads a research group, and (v) is placed on the third or fourth level in the four-level career path for researchers. The number of principal investigators at the UH city centre campus is about 650-700.

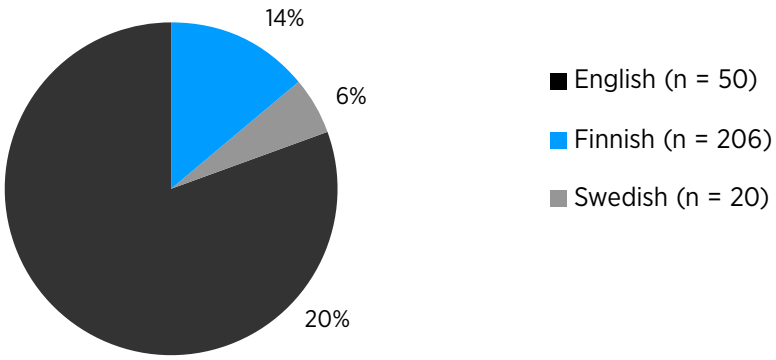
Surveying the practices, needs and views of principal investigators was the first step (step 1) in a three-step process of planning and implementing the research infrastructure mission of HSSH. The next steps will be to communicate the results of step 1 to existing and potential research infrastructure service providers within and outside of UH in order to investigate options to better match the demand and supply of research infrastructure services (step 2); and to consider viable alternatives to organizing, managing and financing the research infrastructures needed by SSH at UH (step 3). This report presents the results of step 1.

The structure of this report follows loosely the structure of the survey (see Chapter 2). Appendix 1 presents some basic information on the respondents (respondent's unit, position, fields of research, etc.). Chapter 3 gives an overview of the characteristics of the respondents' research in terms of types of research material or data, the producers of research material, sources of acquired research material, research methods, and the representation of "infrastructure-intensive" research areas. Chapter 4 looks at the research equipment and services needed by the respondents to produce, acquire or process research data. It presents response distributions to structured questions and analyses responses to open-ended questions. Chapter 5 deals with respondents' attitudes towards data sharing, as well as their reasons for not sharing data with others. Chapter 6 discusses respondents' views on the state-of-the-art and future development of knowledge production technologies in their fields. We clustered those views around four categories, each representing a broad yet distinctive "data-intensive" research area, and attempted to create an overview of the required actions and investments in those areas. Chapter 7 draws some conclusions and makes recommendations for HSSH in implementing its research infrastructures tasks.

2 BASIC INFORMATION OF RESPONDENTS

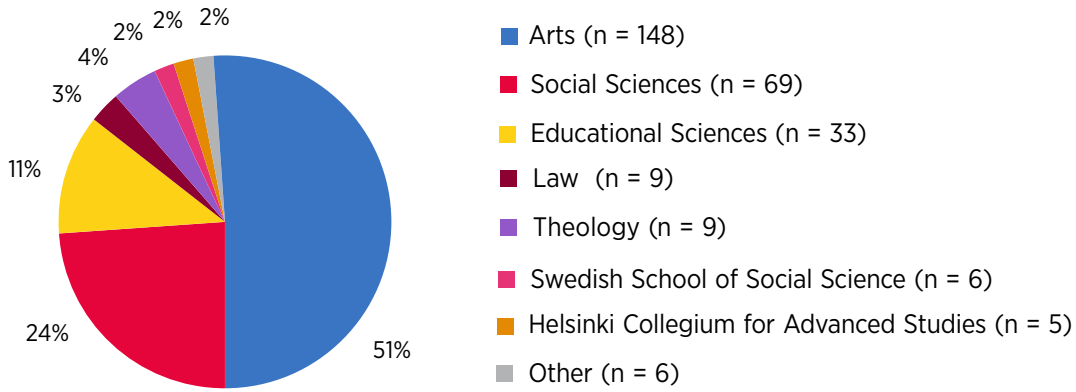
The total number of responses received was 356, of which 190 were fully completed and 166 partially filled in. The survey was available in three languages: Finnish, English and Swedish. The distribution of response languages is presented in Figure 1.

Figure 1. Response language (N=356)



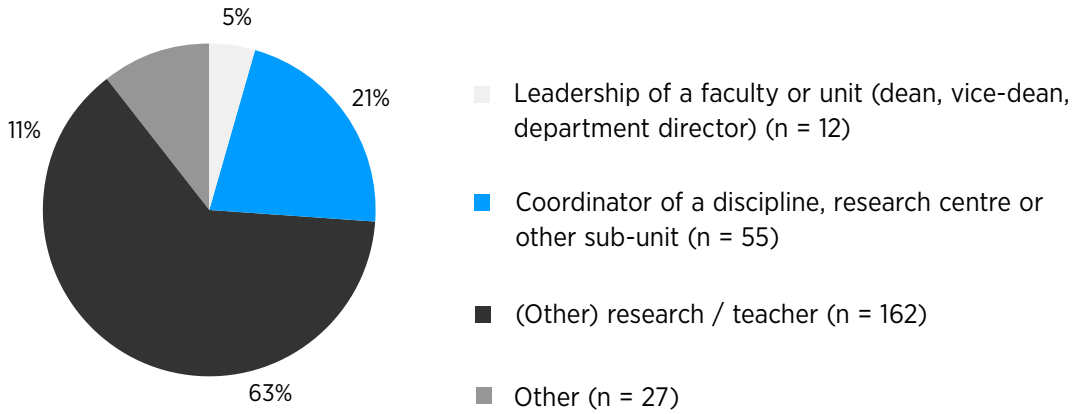
Responses were received from all target faculties and units. It was possible to select multiple faculties or units. Most respondents selected one unit (88%), 14 respondents selected two units, and two respondents selected three units. Researchers from the Faculty of Arts were the most active respondents. Respondents selecting an "Other" unit mentioned HELSUS (n=4), the Faculty of Science (n=2), INEQ initiative (n=1) and URBARIA (n=1) as their unit or faculty. The distribution of responses from different faculties or units is presented in Figure 2.

Figure 2. Faculty or unit of the respondent (N=289) (Note that the total number of faculty or unit representations exceeds the number of responses; some respondents represent more than one unit.)



The position of the respondent was divided into four categories: *Leadership of a faculty or unit* (*dean, vice-dean, department director*), *Coordinator of a discipline, research centre or other sub-unit*, (*Other) researcher/teacher*, and *Other*. The respondents were requested to give a free-form comment if they selected the category "Other". The most frequently selected category (63%) was *(Other) researcher/teacher*. The most common mentions in the category "Other" were PhD candidate (n=7), Emeritus/Emerita (n=6), and Docent (n=6). The distribution of respondents according to their position is presented in Figure 3.

Figure 3. Position of the respondent (N=289)



The respondents were also requested to specify the perspective from which they were responding to the survey. A list of units on behalf of which respondents were answering are listed in Appendix 2 (in cases where a free-form comment was provided).

The respondents were requested to specify the disciplines their research primarily represented. As answer options, we used the list of research fields in the University of Helsinki Research information system (Tuhat). Disciplines with at least five mentions and the percentage of respondents selecting them are presented in Table 1. The most common disciplines represented the humanities, *Languages* being the most commonly selected discipline. *Humanities* as a general category was the second most mentioned discipline, and *History and archaeology* the third most mentioned. *Sociology* was the most common discipline within the social sciences. Other categories with over 15 mentions were *Other humanities, Educational sciences, Political science, Literature studies, Philosophy*, and *Psychology*.

Table 1. Which discipline(s) does your research primarily represent?
Discipline with at least five representatives and the percentage of respondents selecting it (N=289) (The numbers before the disciplines refer to the research field classification in the University of Helsinki Research information system.)

Discipline	Count	Percentage of respondents
6121 Languages	69	24%
HUMANITIES	51	18%

615 History and archeology	47	16%
6160 Other humanities	33	11%
5141 Sociology	33	11%
516 Educational sciences	31	11%
517 Political science	28	10%
6122 Literature studies	18	6%
611 Philosophy	16	6%
515 Psychology	16	6%
615 Theology	15	5%
5200 Other social sciences	15	5%
513 Law	13	4%
5143 Social anthropology	13	4%
5202 Economic and social history	13	4%
113 Computer and information sciences	13	4%
SOCIAL SCIENCES	12	4%
5144 Social psychology	12	4%
5201 Political history	12	4%
5142 Social policy	11	4%
518 Media and communications	11	4%
6162 Cognitive science	10	3%
6131 Theatre, dance, music, other performing arts	8	3%
512 Business and management	7	2%
5203 Development studies	7	2%
6132 Visual arts and design	6	2%
6161 Phonetics	5	2%
519 Social and economic geography	5	2 %

3 CHARACTERIZATION OF RESPONDENTS`RESEARCH

The first substantial section of the survey inquired about the characteristics of respondents' research. Rather than starting by asking about research infrastructures per se, we thought we would be better off by asking questions about the kinds of data respondents are using in their research, the producers and sources of that data, and the methods they use for analysing the data. Given the relatively low level of conceptualization and institutionalization of research infrastructures in the social sciences and humanities, at least at the UH, we decided to focus on the data production and processing technologies that underlie the question of research infrastructures. At the end of this section, however, we also explicitly asked whether respondents would consider their research or research interests as part of one or more "infrastructure-intensive" research areas.

3.1 TYPES OF RESEARCH MATERIAL

The respondents were requested to select the research material or data types they primarily used in their research. It was also possible to specify the answer in an open field. Roughly half of all respondents indicated using *documents and archives (53%)*, *interview data (50%)* or *literature (49%)* as one of their primary research material. More than one third of the respondents mentioned using *writings or other products of research subjects (38%)*, *media material or social media data (38%)*, *observation data (36%)* or *survey data (36%)*. The number of selections for each research material or data type, as well as the percentage of respondents, are presented in Table 2.

Table 2. What kind of research material or data do you primarily use in your research? The number of selections for each material or data type and the percentage of respondents selecting it (N=242).

Research material or data type	Count	Percentage of respondents
Documents, archives	128	53%
Interview data	122	50%
Literature as empirical material	119	49%
Writings or other products of research subjects	92	38%
Media material, social media data	91	38%
Observation data	88	36%
Survey data	87	36%
Linguistic corpora etc.	65	27%
Images, objects, artefacts, etc.	62	26%

Register data, register-based statistics	54	22%
`Big data´	45	19%
Experimental data	33	14%
Other type of data, please specify	28	12%
No actual empirical material	5	2%

According to the free-form answers, *documents and archives* include both historical and contemporary documentation, such as policy documents, regulatory and legal documents, historical documents and government document archives.

Interview data include, among other things, individual and group interviews, structured, semi-structured and theme interviews, both written and spoken. Interview data often supplement some other data.

Literature as empirical material include e.g. scientific literature, fiction, and news and current affairs media writings.

Writings or other products of research subjects include pictures and video material, writings, texts, drawings and other products of children and pupils, artistic products, letters, speeches, dictums and other textual products, peer interviews, and the reflections of peer researchers.

Media material and social media data include e.g. news archives, news articles, radio programmes and documentaries, web pages, corporate profiles on social media, as well as commentary on discussion forums, social media (e.g. Facebook, Twitter, Instagram, WhatsApp) and chats.

Observation data include ethnographic field work, focus group observation, participatory observation, archaeological field observation, and classroom or day care observations. Observation data are typically linked with other data types, such as interviews.

As for *survey data*, respondents use both externally produced and self-acquired data. Survey data are often collected through electronic questionnaires, and used jointly with e.g. interview data.

Linguistic corpora are used in text and audio formats (recorded speech, video data). Corpora are often acquired from openly available sources such as Language Bank, or other CLARIN services.

As for *images, objects, artefacts, etc.* as research material, respondents mentioned using e.g. photographs, satellite images, maps, images of events, archaeological data, art images and artefacts, historical samples and historical images.

Register data or register-based statistics mentioned by respondents include statistics about socio-economics, tax registers, student grades, PISA research data and geographic information system data.

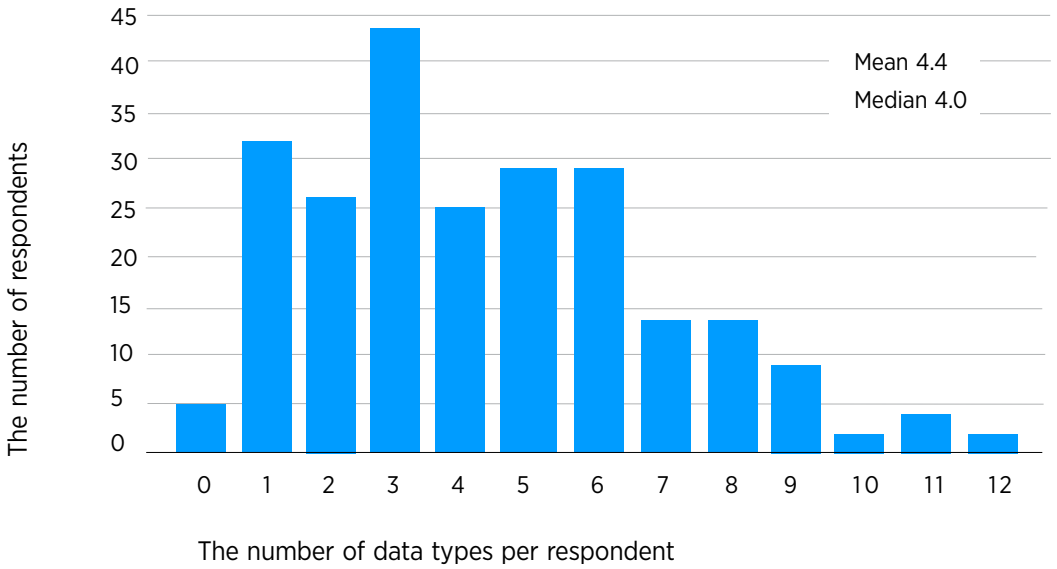
As for *big data*, respondents mentioned language corpora, speech corpora, news corpora, GIS data, linked data (RDF-data bases) and digitated data (e.g. ECCO, EEBO, ESTC).

As for *experimental data*, respondents mentioned field experiments, data measuring research subjects' activity and reactions, as well as behavioural data such as gameplay data, eye tracking, psychophysiology, speech recordings or targeting phonetic research.

Other types of data mentioned were, among other things, physiological data (e.g. EEG, heart rate), data on the built environment and commercials.

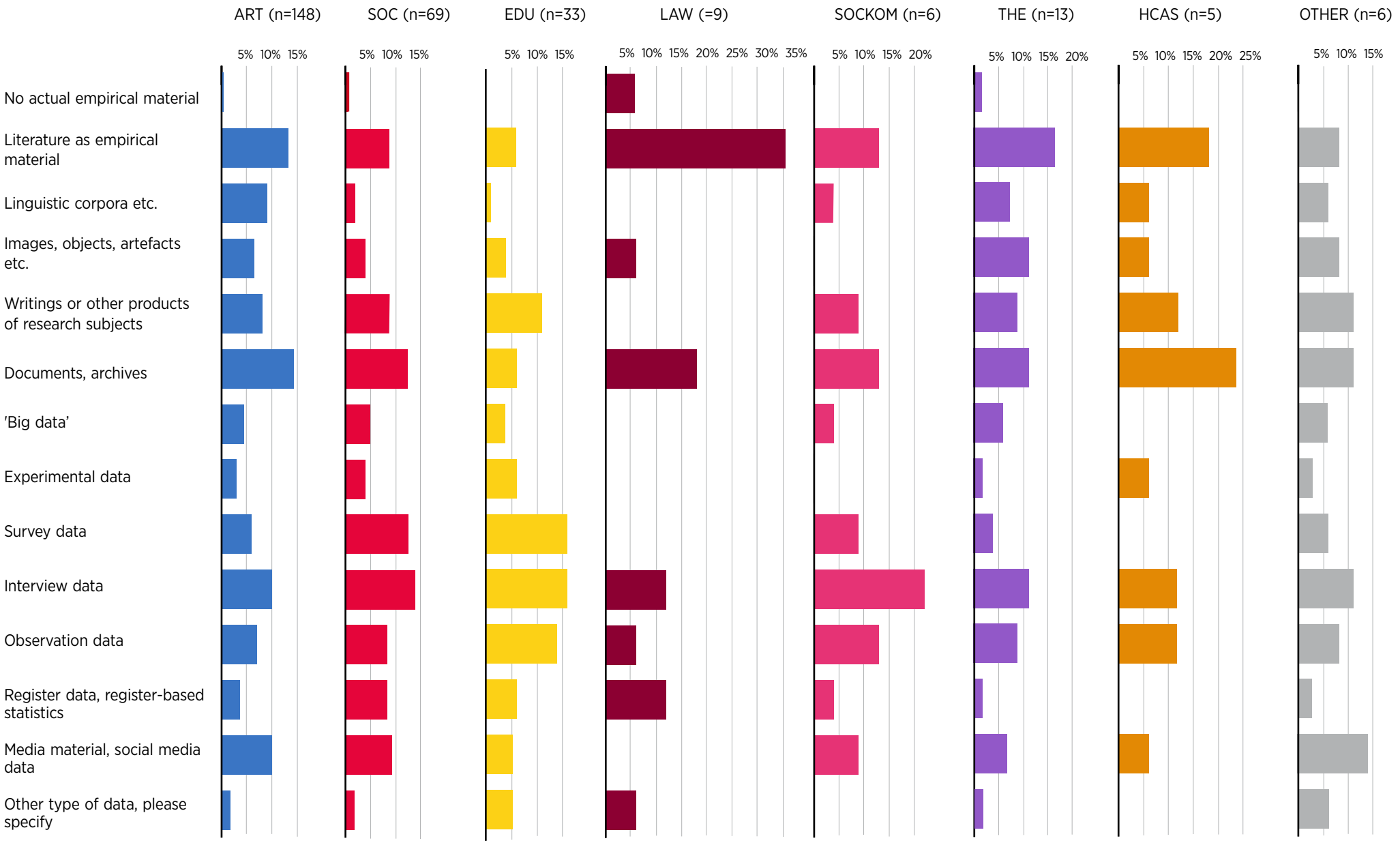
Respondents typically use several different types of data, as can be observed from Figure 4. On average, respondents selected 4.4 different research material or data type categories. The most common count of data types selected (n=44) was three categories.

Figure 4. The number of different research material or data types selected by respondents (N=230).



The distribution of various research material or data types used by respondents from different faculties and units is presented in Figure 5. The profiles are quite similar across units. In the Faculty of Arts, the Faculty of Social Sciences, the Swedish School of Social Science, the Faculty of Theology and the Helsinki Collegium for Advanced Studies, *documents and archives, interview data, literature as empirical material, media material and social media data* and *observation data* are widely used. In the Faculty of Theology, among the most popular data types are also *images, objects, artefacts, etc.* In the Faculty of Educational Sciences, *interview data, survey data, observation data* and *writings or other products of research subjects* were the most popular data types. In the Faculty of Law, the most popular data type is *literature as empirical material*, but *documents and archives, register data* and *interview data* are also common.

Figure 5. The distribution (%) of different research material or data types used by respondents from each faculty and unit.

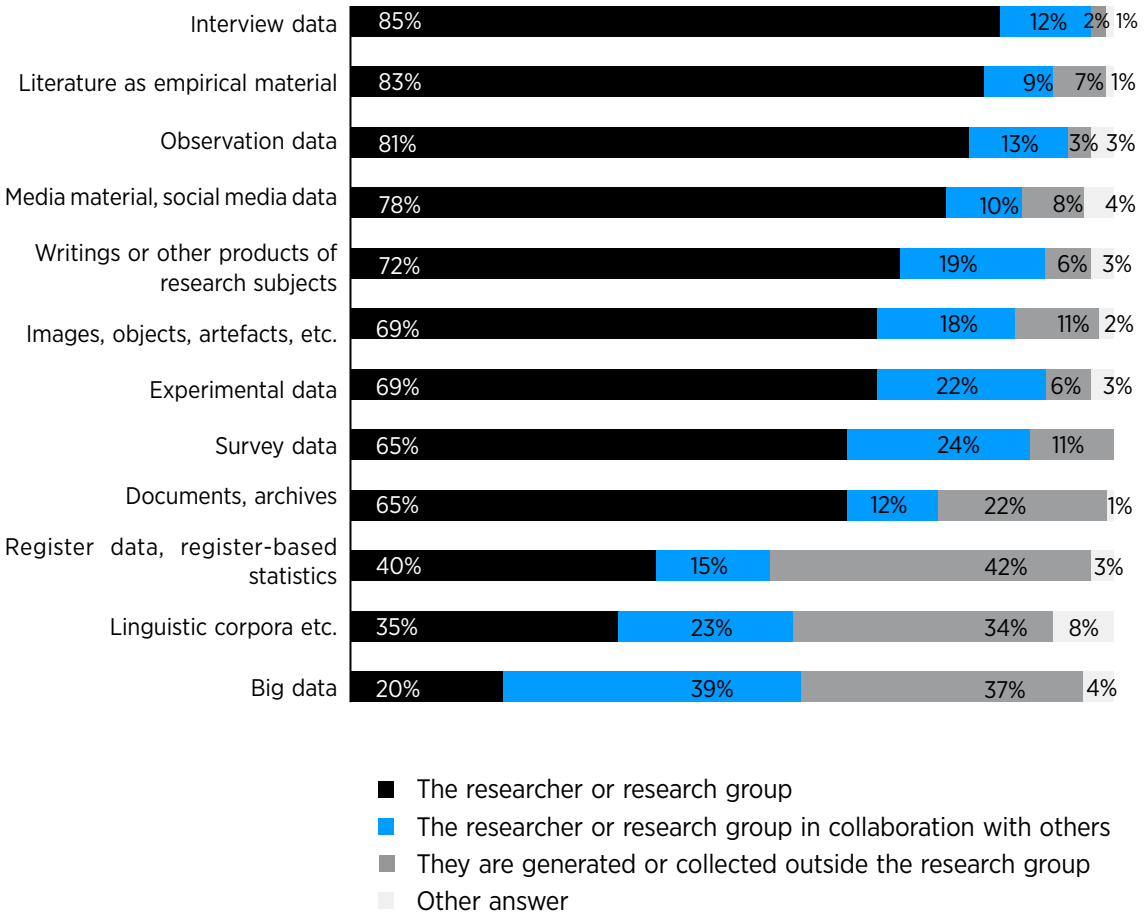


ART = Faculty of Arts, SOC = Faculty of Social Sciences, EDU = Faculty of Educational Sciences, LAW = Faculty of Law, SOCKOM = Swedish School of Social Science, THE = Faculty of Theology, HCAS = Helsinki Collegium for Advanced Studies, OTHER = Other faculty or unit

3.2 PRODUCERS OF RESEARCH MATERIAL

For most research material or data types, the data is generated or collected by the researcher or the research group. *Register data or register-based statistics* and *Linguistic corpora, etc.* are often generated or collected also outside the research group. *Big data* are mainly generated or collected either in collaboration with outside partners, or completely outside the research group. The distribution of the primary generator or collector of research data for each data type is presented in Figure 6.

Figure 6. Who primarily generates or collects your research data? (N=237)



3.3 SOURCES OF ACQUIRED RESEARCH MATERIAL

The participants were asked about from where or with whom they generate data outside their own research group. The number of selections for each partner type and the percentage of respondents selecting them are presented in Table 3. The participants were also requested to comment on their selections in an open field. The free-form comments are summarized in Appendix 3.

Table 3. If you acquire research data (or material) from outside your research group or generate data in cooperation with others, from where or with whom? The number of selections for each partner type and the percentage of respondents selecting it (N=237).

Partner	Count	Percentage of respondents
Foreign universities or research institutes	102	43%
Other Finnish universities or research institutes	72	30%
Open data repositories	64	27%
Other units at the University of Helsinki	63	27%
Memory organizations	63	27%
Public authorities	54	23%
Research infrastructures (e.g. CLARIN, ESS)	38	16%
Companies	23	10%
Others, please specify	28	12%

The most common partners were *foreign universities or research institutes*. The free-form answers showed that the respondents acquire data from (or with) a large number of different universities and other research institutes around the globe. Universities and other research institutes in Europe were most commonly mentioned. Common European countries were the UK and Sweden. In addition to Europe, research data and material was often acquired from the United States, Australia, Russia, and Asian countries such as China and Japan.

Other Finnish universities and research institutes, such as the Institute for the Languages of Finland (Kotus), the Finnish Institute for Health and Welfare (THL) and Statistics Finland, were also mentioned as common sources of research data and material.

Open data repositories was the third biggest category. Respondents mentioned among the most common Statistics Finland, Finnish Social Science Data Archive (FSD) and a number of foreign data archives.

Memory organizations, such as archives and museums, *other units at the University of Helsinki* (e.g. the University Library, the National Library, the Department of Psychology and Logopedics, the Faculty of Medicine and the BioMag Laboratory) and *public authorities* (e.g. the Ministry of Education and Culture, the Population Register Centre, hospitals and KELA) were also commonly mentioned as partners.

The most commonly mentioned *research infrastructures* were (FIN-)CLARIN, Language Bank, CSC, and the European Social Survey.

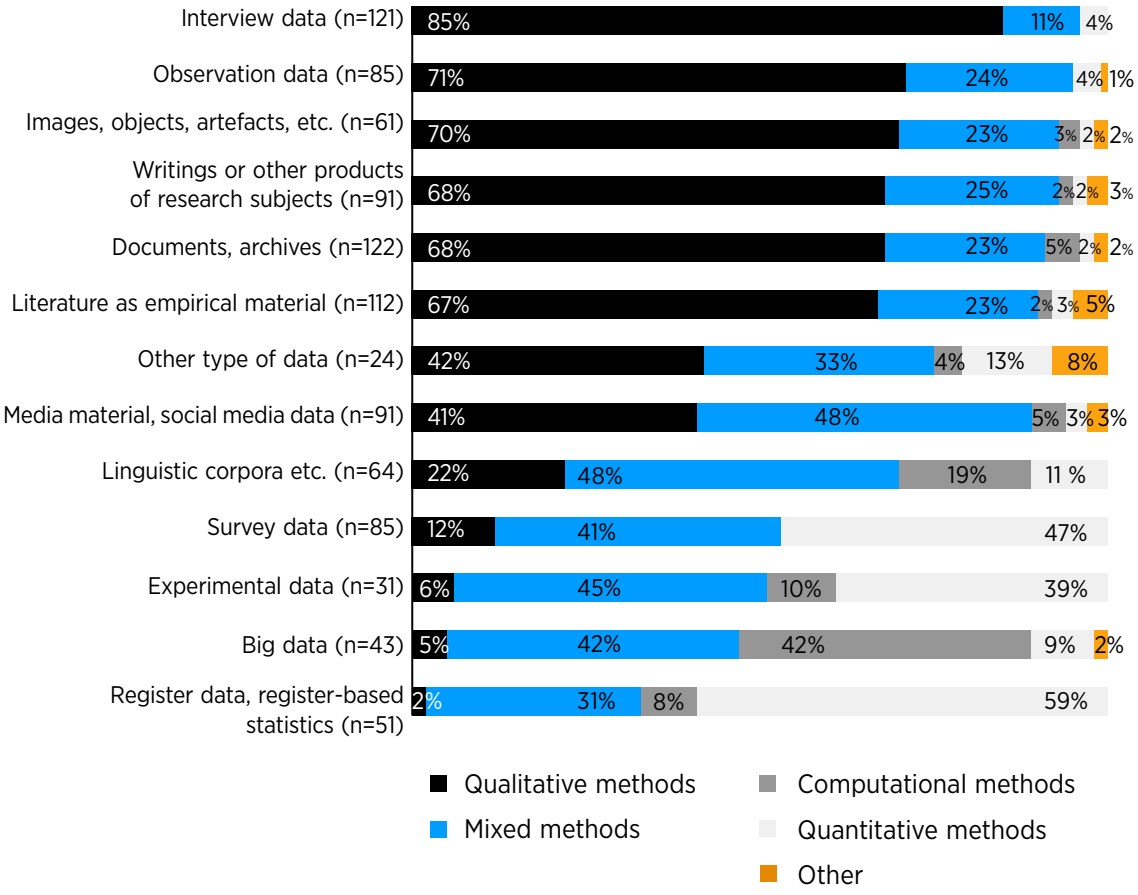
Companies included media companies, private hospitals, publishing companies and private research organizations such as Taloustutkimus and the Research Institute of the Finnish Economy (ETLA).

Other partners mentioned were other researchers, Finnish and international third sector organizations, cities and municipalities, schools and libraries.

3.4 RESEARCH METHODS

The participants were requested to select the primary methods they use in analysing research data. The question of methods was tied to the previous question of data types; respondents were asked to select methods for each type of data they use. The distribution of primary methods used in analysing different types of data is presented in Figure 7.

Figure 7. What kind of methods do you primarily use in analysing your research data? (N=237)



Qualitative methods are the most popular methods used in analysing interview data (85%), observation data (71%), images, objects and artefacts (70%), writings or other products of research subjects (68%), documents and archives (68%) and literature as empirical material (67%).

Mixed methods are used to some extent for these data types, and constitute the most popular approach for analysing media material and social media data (48%), linguistic corpora (48%), experimental data (45%) and big data (42%).

Computational methods are primarily used for analysing big data (42%), but to some extent also for linguistic corpora (19%), experimental data (10%) and register data and register-based statistics (8%).

Quantitative methods constitute the most frequently used approach in analysing register data and register-based statistics (59%) and survey data (47%) and is also commonly used with experimental data (39%).

3.5 REPRESENTATION OF "INFRASTRUCTURE-INTENSIVE" RESEARCH AREAS

The participants were asked to state if their research interests could be considered part of one or more "infrastructure-intensive" research areas. There were eight answer options: *computational linguistics / language technology; (other) digital humanities; archaeology; experimental social and behavioural sciences; register-based social sciences; regular large-scale social surveys; computational social sciences / social data science* and *other*. This list was based on our preliminary understanding of what research traditions there are at UH that might require some infrastructural support. This understanding, in turn, was created on the basis of e.g. the self-assessment reports of the recent Research Assessment exercise at UH (2018-2019) as well as the letters of intent submitted to the recent Finnish Research Infrastructure (FIRI) calls for funding applications. In addition to select relevant categories, respondents were also requested to specify their selections with a free-form comment. 53% of the respondents (n=128) selected at least one area. The number of selections for each category and their share of the total number of respondents are presented in Table 4.

Table 4. Can your research or research interests be considered part of one or more of the following "research infrastructure-intensive" areas? The number of selections for each area and the percentage of respondents selecting it (N=242).

Infrastructure-intensive research area	Count	Percentage of respondents
(Other) digital humanities	55	23%
Experimental social and behavioural sciences	34	14%
Computational linguistics / Language technology	32	13%
Regular large-scale social surveys	26	11%
Computational social sciences / Social data science	16	7%
Archaeology	13	5%
Register-based social sciences	12	5%
Other; please specify	20	8%
(No selections)	(114)	(47%)

The most frequently selected category was *(other) digital humanities* (23%). Other frequent selections were *experimental and behavioural sciences* (14%), *computational linguistics / language technology* (13%), and *regular large-scale social surveys* (11%).

According to the free-form comments, *(other) digital humanities* deal with large digital data such as media data, texts, interview data, internet data and data bases, which are analysed with a variety of different methods: computational methods (e.g. computational history research), network analysis, and many non-quantitative methods.

Experimental social and behavioural sciences is relevant in areas such as social and moral psychology, moral cognition, experimental philosophy, experimental cognitive science and cognitive neuroscience, behavioural and brain research, phonetic experimental research, social use of languages, didactics and policy exercises. Data types mentioned by the respondents include audio-visual data, interview data, eye tracking data and EEG and other physiological data.

Computational linguistics / language technology typically deals with large data collections and corpora. Several respondents stated developing or using language technologies.

In the free-form comments of *regular large-scale social surveys*, respondents mentioned the following research areas: social psychology and moral psychology, cognitive sciences, research on social structures and divisions, school research and political participation. Data types mentioned by the respondents include longitudinal survey data, interview data and experimental data.

As *other* infrastructure-intensive research areas, the respondents mentioned e.g. conversation analysis, which requires equipment for data gathering, software for editing and analysing, and large storage capacity, especially for video material. Onomastics, anthropology, qualitative language studies and papyrus research are other examples of mentioned areas.

We also looked at the characteristics of respondents within each "infrastructure-intensive" research area. Table 5 lists the most common disciplines of the respondents selecting an area. Table 6 shows the prevalence of using different types of research material among the respondents selecting an area. Table 7 presents the number of faculty or unit representations among the respondents selecting an area.

Table 5. The most common disciplines of the respondents who considered their research or research interests part of one or more "infrastructure-intensive" research areas listed in the questionnaire. (The numbers before the disciplines refer to the research field classification in the University of Helsinki Research information system.)

Research infrastructure-intensive area Disciplines (Top 5)	Count
Computational linguistics / language technology (n=32)	
6121 Languages	24
HUMANITIES	10
113 Computer and information sciences	8
615 History and archaeology	5
6160 Other humanities	4
(Other) digital humanities (n=55)	
6121 Languages	22
HUMANITIES	15
615 History and archaeology	15

6160 Other humanities	11
113 Computer and information sciences	8
Archaeology (n=13)	
615 History and archaeology	11
6121 Languages	6
HUMANITIES	2
614 Theology	2
6160 Other humanities	2
5202 Economic and social history	2
Experimental social and behavioural sciences (n=34)	
6162 Cognitive science	8
515 Psychology	8
516 Educational sciences	8
5144 Social psychology	6
615 History and archaeology	5
517 Political science	5
Register-based social sciences (n=12)	
517 Political science	5
5141 Sociology	2
515 Psychology	2
5200 Other social sciences	2
3142 Public health care science, environmental and occupational health	2
Regular large-scale social surveys (n=26)	
5144 Social psychology	6
515 Psychology	6
517 Political science	6
516 Educational sciences	5
6160 Other humanities	2
5141 Sociology	2
5200 Other social sciences	2
112 Statistics and probability	2

Computational social sciences / social data science (n=16)	
517 Political science	5
HUMANITIES	4
5200 Other social sciences	4
113 Computer and information sciences	4
6162 Cognitive science	2
515 Psychology	2
112 Statistics and probability	2
Other (n=20)	
6121 Languages	12
HUMANITIES	5
615 History and archaeology	4
6122 Literature studies	3
614 Theology	2
5141 Sociology	2
5143 Social anthropology	2

Table 6. The prevalence of using different types of research material or data among the respondents who considered their research or research interests part of one or more "infrastructure-intensive" research areas.

	Computational linguistics / language technology (n=32)	(Other) digital humanities (n=55)	Archaeology (n=13)	Experimental social and behavioural sciences (n=34)	Register-based social sciences (n=12)	Regular large-scale social surveys (n=26)	Computational social sciences / social data science (n=16)	Other (n=20)	(No selection, n=114)
Literature as empirical material (n=119)	18	30	9	11	3	5	8	8	(65)
Linguistic corpora etc. (n=65)	27	28	7	7	0	2	5	7	(19)
Images, objects, artefacts, etc. (n=62)	8	16	12	7	1	4	4	4	(30)
Writings or other products of research subjects (n=92)	14	24	3	15	3	8	7	7	(49)
Documents, archives (n=128)	19	34	8	13	5	10	10	11	(65)
'Big data' (n=45)	15	27	2	14	4	8	9	3	(3)
Experimental data (n=33)	6	16	0	20	2	8	4	2	(2)
Survey data (n=87)	11	23	2	23	10	24	8	4	(31)
Interview data (n=122)	12	25	3	19	6	18	9	10	(66)
Observation data (n=88)	6	18	4	15	4	9	6	7	(48)
Register data, register-based statistics (n=54)	7	8	3	7	10	11	8	3	(22)
Media material, social media data (n=91)	16	29	2	11	5	8	7	5	(46)
Other type of data (n=28)	3	7	1	7	0	5	2	6	(12)

Table 7. The faculties or units of the respondents who considered their research or research interests part of one or more “infrastructure -intensive" research areas.

	Computational linguistics / language technology (n=32)	(Other) digital humanities (n=55)	Archaeology (n=13)	Experimental social and behavioural sciences (n=34)	Register-based social sciences (n=12)	Regular large-scale social surveys (n=26)	Computational social sciences / social data science (n=16)	Other (n=20)	(No selection, n=114)
Faculty of Arts (n=148)	30	40	11	14	2	6	7	14	(58)
Faculty of Social Sciences Science (n=69)	1	8	0	10	7	13	9	3	(27)
Faculty of Educational Sciences (n=33)	1	5	0	11	2	9	1	0	(13)
Faculty of Law (n=13)	0	0	0	0	0	0	0	1	(6)
Faculty of Theology (n=13)	0	4	2	3	0	0	0	1	(7)
Soc & Kom (n=6)	0	1	0	1	1	0	0	0	(3)
HCAS (n=5)	0	1	0	0	0	0	0	1	(3)
Other faculty or unit n=6)	2	3	1	1	0	0	1	0	(2)

4 WHAT KIND OF RESEARCH EQUIPMENT AND SERVICES ARE NEEDED?

The main part of the questionnaire concerned the research equipment and services researchers need to produce, acquire, process, use and share research data. These needs were mapped with three (obligatory) structured matrix questions and combined with a set of (voluntary) open-ended questions that enabled the respondents to specify their selections. In addition, there was a separate open-ended question that inquired about other deficiencies or development needs related to the research infrastructures for the social sciences and humanities. In what follows, we first present the response distributions to the three structured questions, then summarize the content of the free-form answers.

4.1 RESPONSES TO STRUCTURED QUESTIONS

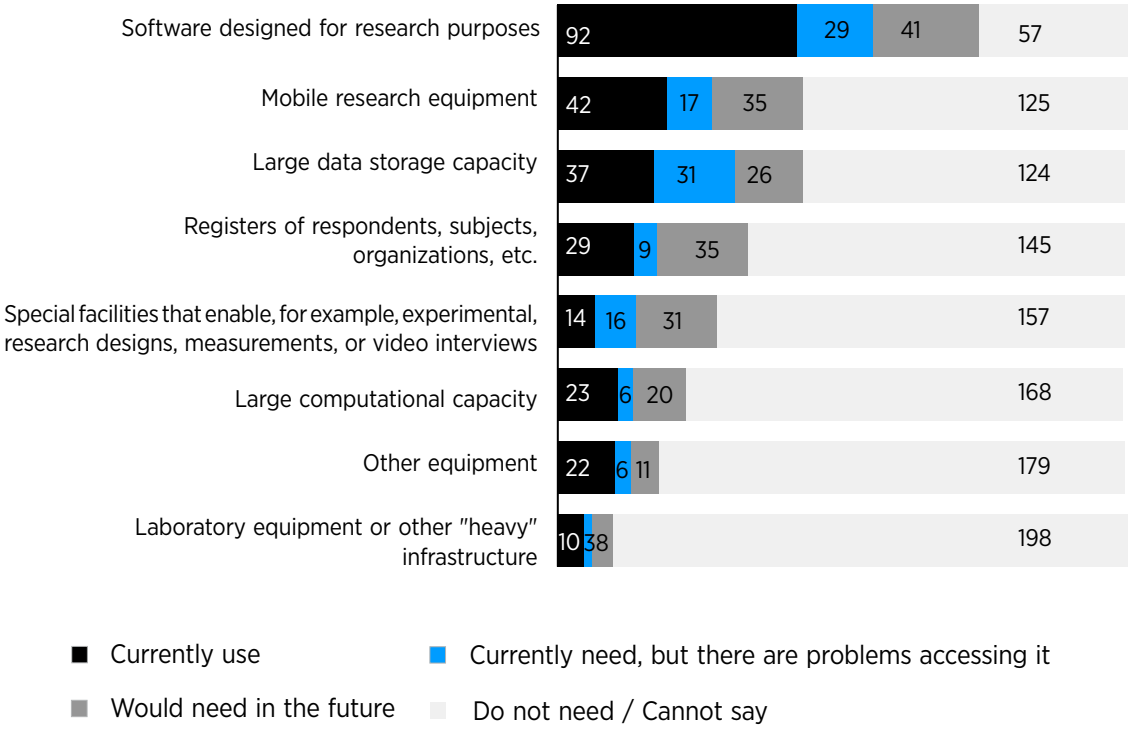
4.1.1 RESEARCH EQUIPMENT

The respondents were first asked what kinds of research equipment they needed to produce, acquire or process research data. There were eight sub-questions specifying the type of equipment, one of which was other *equipment,* and four answer options for each sub-question: *currently use, currently need but there are problems accessing it, would need in the future,* and *do not need/cannot say.* Figure 8 presents the distribution of answers to each sub-question.

There are two categories of equipment, in particular, that respondents currently need but experience problems with accessing: *large data storage capacity* (n=31) and s*oftware designed for research purposes* (n=29). Problems are also experienced with access to *mobile research equipment* (n=17) and *special facilities that enable, for example, experimental research designs, measurements or video interviews* (n=16).

The largest equipment categories needed in the future are *software designed for research purposes* (n=41), *mobile research equipment* (n=35), *registers of respondents, subjects, organizations, etc.* (n=35) and *special facilities that enable, for example, experimental research designs, measurements, or video interviews* (n=31) and *large data storage capacity* (n=26).

Figure 8. What kinds of research equipment do you need to produce, acquire or process research data? (N=226) (Note that the number of responses may differ between sub-questions, because not everyone responded to all sub-questions.)



We also examined the responses within the three largest faculties: Arts, Social Sciences and Educational Sciences. A summary of the results is presented here, and full figures for each of the three faculties are presented in Appendix 4.

In the **Faculty of Arts**, respondents are currently experiencing problems accessing *Large data storage capacity* and *Software designed for research purposes*. Respondents also state that they need these in the future, along with *Mobile research equipment* and *Registers of respondents, subjects, organizations, etc.*

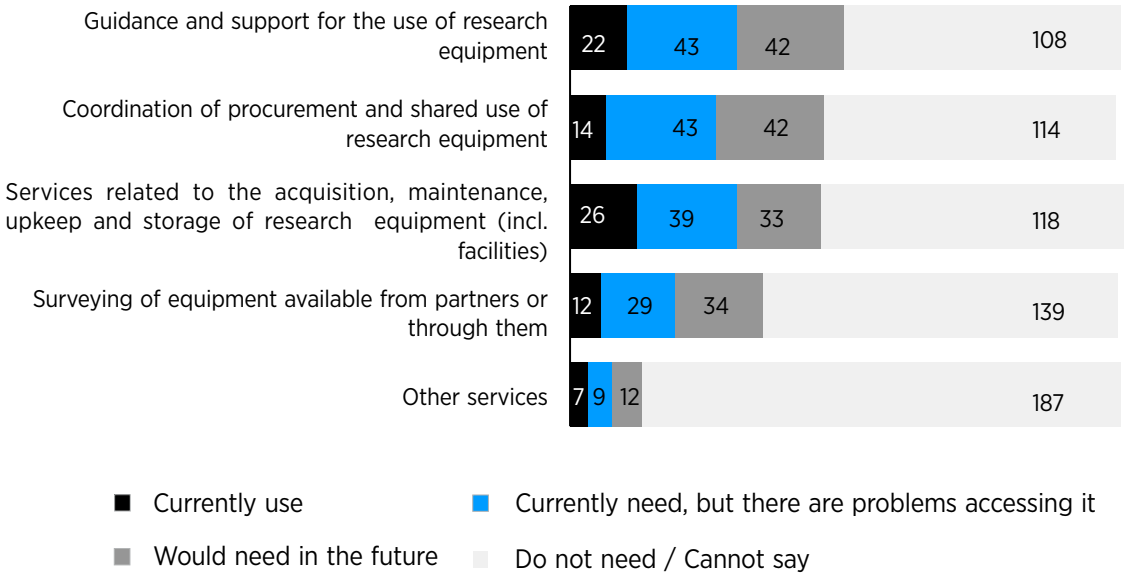
Similar to the Faculty of Arts, the respondents in the **Faculty of Social Sciences** currently have problems accessing *Software designed for research purposes* and *Large data storage capacity*. In the future, respondents need especially *Registers of respondents, subjects, organizations, etc., Special facilities that enable, for example, experimental research designs, measurements, or video interviews and Software designed for research purposes.*

In the **Faculty of Educational Sciences**, the respondents currently need but have problems accessing L*arge data storage capacity* and *Mobile research equipmen*t, which is also needed in the future. Other equipment needed in the future includes *Special facilities that enable, for example, experimental research designs, measurements, or video interviews* and *Registers of respondents, subjects, organizations, etc.*

4.1.2 SERVICES FOR ACQUIRING AND USING RESEARCH EQUIPMENT

Second, the respondents were asked what kinds of services they needed to acquire or use research equipment. There were five sub-questions specifying the type of service, one of which was *other services*, and four answer options for each sub-question: *Currently use, Currently need but there are problems accessing it, Would need in the future*, and *Do not need/cannot say*. Figure 9 presents the distribution of answers to each sub-question.

Figure 9. What services do you need to acquire or use research equipment? (N=226) (Note that the number of responses may differ between sub-questions, because not everyone responded to all sub-questions.)



According to the responses, there is a shortage of services in all categories. The most obvious shortage concerns *Coordination of procurement and shared use of research equipment*. Only 12 respondents reported that they currently use the service, while 43 respondents reported that they have problems accessing the services. Results in the other response categories (except for *other services*) were similar: the number of respondents who reported having problems with access to the services clearly exceeds the number of respondents who currently use the services.

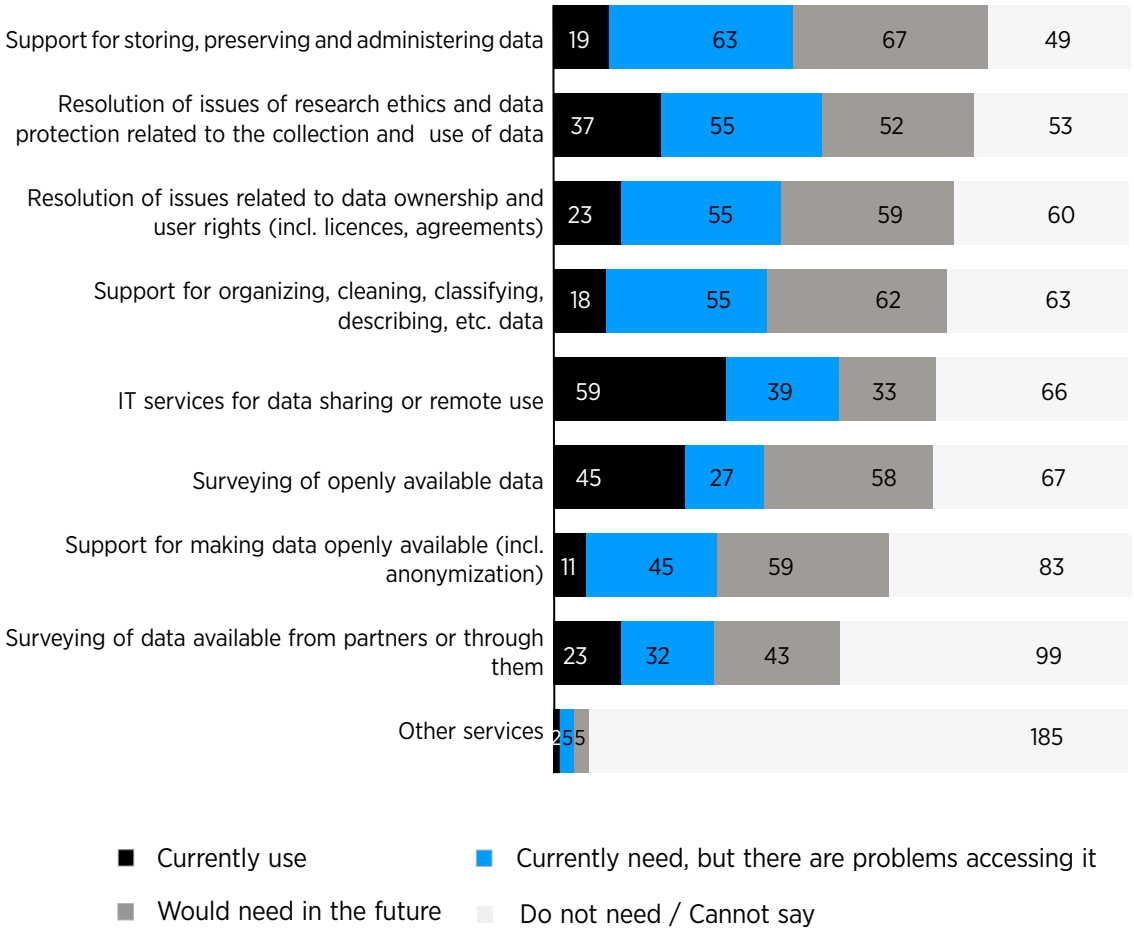
Again, we also examined the responses within the three largest faculties: Arts, Social Sciences and Educational Sciences. The response distributions in these faculties are similar to the overall results; there is a shortage of services in all categories. Similarly, the shortage is most obvious in *Coordination of procurement and shared use of research equipment* in all three faculties. The results of the **Faculty of Educational Sciences** differ from the other faculties in that the majority of respondents indicated needing the services either currently or in the future. Only two respondents altogether stated that they currently used any of the services, thus suggesting that although presently not available, there is a clear demand for these services. Full figures for each of the three faculties are presented in Appendix 5.

4.1.3 SERVICES FOR ACQUIRING AND USING RESEARCH DATA

Third, the respondents were asked what kinds of services they needed to acquire or use research data. There were nine sub-questions specifying the type of service, one of which was other services, and four answer options for each sub-question: *currently use, currently need but there are problems accessing it, would need in the future*, and *do not need/ cannot say*. Figure 10 presents the distribution of answers to each sub-question.

In eight out of the nine sub-questions, over 50% of the respondents stated that they either currently used, would currently need, or would need the service in the future. The most obvious shortage of services concerns the following: *Support for storing, preserving and administering data; Resolution of issues of research ethics and data protection related to the collection and use of data; Resolution of issues related to data ownership and user rights (incl. licences, agreements); Support for organizing, cleaning, classifying and describing data;* and *Support for making data openly available (incl. anonymization).* The number of respondents currently having problems with access to the needed services clearly exceeds the number of respondents presently using these services. Roughly a third of all respondents in most sub-questions state they would need the service in the future.

Figure 10. What services do you need to acquire or use data? (N=200) (Note that the number of responses may differ between sub-questions, because not everyone responded to all sub-questions.)



We also examined the responses within the three largest faculties, Arts, Social Sciences and Educational Sciences. A summary of the results is presented here, and full figures for each of the three faculties are presented in Appendix 6. The responses in the **Faculty of Arts** are similar to the overall results. A majority of the respondents *currently use,* would *currently need,* or *would need* the following services *in the future: Support for storing, preserving and administering data; Resolution of issues of research ethics and data protection related to the collection and use of data; Resolution of issues related to data ownership and user rights (incl. licences, agreements); Support for organizing, cleaning, classifying and describing data;* and *Support for making data openly available (incl. anonymization).* In the **Faculty of Social Sciences**, the shortage of services is most evident in the same sub-questions, but in general, a lightly smaller share of respondents stated they were either currently using, or would currently need the services than respondents in the Faculty of Arts. In the **Faculty of Educational Sciences**, there is an obvious shortage of services in all sub-questions except for *Surveying of openly available data*, which, on the other hand, roughly 50% of the respondents stated that they would need in the future.

4.1.4 TRAINING ON DATA ACQUISITION AND USE

The respondents were also asked if some of the data-related service needs should be addressed by providing training for researchers. Most of the respondents did not feel that the data-related service needs should be addressed by providing training. The number of selections for each service category and the share of respondents are presented in Table 8.

Table 8. Should some of the service needs you mentioned above be addressed by providing training for researchers? The count and percentage of respondents who considered training a required option for addressing a given service need (N=198).

Training need	Count	Percentage of respondents
Resolution of issues related to data ownership and user rights (incl. licences, agreements)	66	33%
Resolution of issues of research ethics and data protection related to the collection and use of data	63	32%
Support for storing, preserving and administering data	63	32%
Support for organizing, cleaning, classifying, describing, etc. data	55	28%
Support for making data openly available (incl. anonymization)	50	25%
Surveying of openly available data	46	23%
IT services for data sharing or remote use	41	21%
Surveying of data available from partners or through them	32	16%
Other services	9	5%

Roughly one third of respondents stated that training should be offered for *resolution of issues related to data ownership and user rights; storing, preserving and administering data*; and *resolution of issues of research ethics and data protection related to the collection and use of data*. Approximately one quarter of the respondents stated that training should be offered for o*rganizing, cleaning, classifying, describing, etc. data* and for *making data openly available (incl. anonymization).*

4.2 RESPONSES TO OPEN-ENDED QUESTIONS

As mentioned above, respondents were also requested to specify their responses to the structured questions with free-form answers and comments. We asked them to say more about the equipment and services they used or needed, and for those services and equipment they currently use, to also mention the location of the equipment or the service provider if they knew it (the number of comments were 371, 202 and 344, respectively). We also asked them to say what type of training should be offered to address the data-related service needs (the number of comments was 156). Finally, we asked them to freely comment on any (other) deficiencies or development needs related to the research infrastructures for the social sciences and humanities that could be addressed by the forthcoming HSSH (the number of comments was 58).

In what follows, we summarize the content of all free-form answers into three sections that deal with (1) physical research equipment and facilities, (2) IT services and tools, and (3) research data governance. This logic does not directly follow the original structure of the questions, as we noticed that the distinction between (sub)questions was neither entirely clear nor optimal for creating an overview of the current situation. The division of respondents' needs into the three categories, we think, allows for a more effective overview of the actions to be taken to improve the situation.

4.2.1 PHYSICAL RESEARCH EQUIPMENT AND FACILITIES

We identified three main types of physical research equipment and facilities relevant to social sciences and humanities: laboratories (and laboratory equipment), other special facilities, and mobile research equipment. The main services that are essential for using these physical infrastructures are guidance and training for the use of research equipment; upkeep, storage and shared use of research equipment; and coordination of information, purchase and partnering.

Laboratories are used for e.g. behavioural experimental research, eye tracking research, studying human movement and for archaeological research. Laboratory equipment mentioned by the researchers include EEG and brain imaging equipment (owned by the Department of Psychology); MEG (magnetoencephalography); ultrasound imaging; ED-XRF (Energy Dispersive X-ray Fluorescence); SEM (owned by the Department of Chemistry); mass spectrometry (in Kumpula); instrumented car for driver research; VR and screen-based laboratory simulators including driving simulator; desktop eye tracking, and 3D VR. Researchers and research groups use their own research unit's laboratories, laboratories shared with other units (e.g. in phonetics), or laboratories owned by other units (e.g. Cognitive Brain Research Unit and BioMag at the Faculty of Medicine).

Although laboratories are available, they often do not match current and/or future requirements, for instance by being too small or unprofessional. Respondents mentioned, for instance, that there is a need for laboratories suitable for behavioural research (experiments); larger laboratories for e.g. running several experiments at the same time or studying human movement; laboratories suitable for studying phonetics; and laboratories appropriate for eye tracking research (e.g. enabling standardized lighting conditions).

Other facilities respondents currently use include facilities for (video) interviews with individuals and groups; UH Unitube studios; and the Playful Learning Centre for observing children. According to both structured and free-form answers, there is a clear shortage of special facilities. Respondents mentioned spaces for observation studies, facilities for plain language research that are suitable for studying special groups (e.g. people with learning difficulties or memory disorders), facilities with high information security for studying vulnerable groups, and facilities for simulation exercises based on the Policy Operations Room model. In addition, several respondents highlighted the need for "cooperative facilities" to enhance collaboration between research groups on the campus.

The mobile research equipment used includes both small portable devices – such as cameras, audio recorders, microphones, tablets, scanners in mobile phones, smartphones – and larger portable devices, such as mobile EEG caps, audio visual field equipment, mobile eye tracking, stereomicroscope, ambulatory physiological censors, total station, laptops, accelerometer, Portable X-ray Fluorescence (pXRF) and mobile psychophysiology (MindMedia Nexus). Small mobile equipment is usually acquired for research projects and is stored in offices. Departments and other units (e.g. HCAS) own some equipment (e.g. cameras and microphones) that respondents are able to borrow. Faculties also have common equipment (e.g. three mobile EEG devices in the Faculty of Educational Sciences), but as their location is often unknown, similar equipment is acquired directly for research projects. Although currently widely used, several respondents commented on the shortage of video cameras, recorders, eye tracking devices and field laptops. Other mobile equipment that was needed (either currently or in the future) included portable scanning device, solar panels (for audio-visual fieldwork) and an underwater camera drone.

Guidance and training for the use of research equipment is currently scarcely available, although there is a clear need for these services, in particular with new devices with technological advances. Some respondents spend time learning to use the technologies themselves, while others have received help from assistants or colleagues in their unit, or from other units or even outside the university.

A shortage of services related to the **upkeep, storage and shared use of research equipment** is evident. While laboratory engineers do maintain laboratories and laboratory equipment, there appears to be a shortage of technical staff with professional knowledge able to maintain research equipment. This is especially the case with equipment acquired for specific research projects. Another problem is the lack of a shared use of equipment. Equipment is available, but researchers and groups are often unwilling to share it, or there is no infrastructure for sharing it. In general,respondents wish for more shared use of research equipment, e.g. in the form of an "equipment library", where trained staff are available for handing out equipment and providing guidance for using the equipment.

An even more obvious shortage has to do with **the coordination of information, purchase and partnering** related to research equipment. A centralized database of what equipment is available, and where, is mentioned by several respondents. Evidently, there is unused equipment in various locations around the campus, but no coordination

for their shared use. Coordination would also be beneficial in the purchase of equipment. Respondents would like to have technical professionals who would assist in purchasing the most suitable equipment and would have knowledge about what is already available.

4.2.2 IT SERVICES AND TOOLS

Respondents have urgent needs relating to various IT services and tools for acquiring, processing, using and sharing research data (see Figure 8 and Figure 10). Sufficient, safe and easy to access **data storage** is an essential service for researchers. Videos, images, archaeological data, and even large textual data require large data storage capacity, and respondents anticipate the need for storage space to grow in the future (e.g. with data mining and storing big data). Respondents report using data storage services offered by the university, but several respondents also raise the concern that the storage capacity offered is not sufficient for their research data, and acquiring additional space is not easy:

> *Already my library of pdf versions of publications needed in research was too large (87 Gb) for my personal storage space on the university server. Video recordings are also too large to fit there. It should be easier and cheaper to request more storage space.*

Several respondents are using storage capacity offered by external parties, such as the IT Centre for Science CSC (e.g. IDA Research Data Storage Service) and FIN-CLARIN. Other services and tools mentioned include OneDrive, external hard drives, memory sticks, services offered by foreign universities and commercial services. In addition to large data storage capacity, respondents face challenges with finding a secure storage for sensitive data. This issue is further emphasized by the new data protection legislation. There has been a shortage of secure data storage services at the UH and most respondents are either storing their data in unsecure locations or using outside services. One respondent mentioned using a secure service established for him/her by the UH Data Support, and another reported using "Umpio" (safe data storage offered by UH). As with other areas, respondents feel that information is not easily available and some kind of centralized services should be offered:

> *With the new legislation, data protection and data security have become very important elements of research data. It is unreasonable for every single researcher to think about where to lock his or her sensitive material. The UH should have a centralized service responsible for providing data retention so that security is taken care of and maintained.*

In addition, processing 'big data' with computational tools such as those in language technology, cognitive models or Bayesian methods, require large **computational capacity**. According to the responses, services provided by CSC (e.g. ePouta) fulfil respondents' computational capacity requirements quite well. Other solutions reported include computer clusters, power workstations, computation capacity of the Department of Computer Science, and the servers of the 'IT for Science' group at the Faculty of Science. One respondent noted that the issue is, however, not always a question of capacity, but rather of knowing the correct methodologies.

Research data sharing is essential both within the UH and with outside partners (e.g. between members of a consortium). Respondents listed several services or repositories they use for data sharing, such as services provided by CSC (e.g. IDA, FUNET), UH network drives (e.g. 'P-drive'), Google Drive, Zenodo, GitHub, CLARIN, File sender, and external hard drives. Data sharing within UH is mostly considered to function well, but several respondents reported that sharing data with researchers in other universities is often problematic. Respondents also mentioned a number of other problems they are currently experiencing in relation to data access and sharing. One respondent mentioned that direct cloud storage of recorded data from a cell phone is not possible, so the respondent uses Google instead, even though it involves risks. Another respondent noted that using VPN is always problematic. Sharing large and/or sensitive data (e.g. images and video material and other ethnographic data) is particularly challenging.

The majority of respondents use some **software for research purposes**. Software is used for collecting, transcribing, editing, coding and analysing various types of data, including metadata, as well as for sharing and citing research outputs. The survey tools mentioned by respondents included Survey Monkey and Qualtrics. Statistical analysis software such as SPSS, R, Python and STATA, and qualitative analysis tools Atlas.ti and Nvivo, are commonly used. Another large group of software relates to processing and analysing various audio-visual data: video analysis (e.g. Cinemetrix), image processing and analysis, analysis of eye tracking data, FaceRecognition, speech analysis, synchronizing speech and movement, voice editing, transcribing, analysing prosody, etc. Yet another group of software is used for analysing and processing language and textual data: Elan, Transana, Praat, Transkribus and qualitative analysis tools such as Atlas.fi. Other software mentioned included GIS software (ArcGIS, MapInfo), Presentation software, MATLAB, Mathematica, Oxygen-editor, BibleWorks, Accordance, and DSS Digital Library. Licences for software are acquired either by UH or by the research project, but in some cases software needs to be developed by the researchers or research projects. Some software is also available without licences. Although research software is widely available, respondents reported some problems with access to software; not all software is available at the UH, licences may be too expensive for individual researchers (e.g. Qualtrics), and a shortage of software developers affects the utilization of tailored software.

Respondents raised several needs related to **training for software and tools**. A commonly mentioned area concerns data analysis software (e.g. quantitative and qualitative). Researchers find that their knowledge on how to best use the software is not at the level it should be, and the benefits of the software are not fully utilized. Training is thus considered important, especially for junior researchers. Training should also be available when new updated versions of software are published. Another area brought up by the respondents was the digital humanities and computational methods. In addition to training on the software used in the digital humanities, researchers highlighted the importance of a wider understanding about the areas of application in this new and increasingly topical area.

Inadequate training and information about software is likely to result in unnecessary efforts, and this is time that could be spent on research work. Additionally, respondents feel that there is not enough information available at the UH about the software currently available. Moreover, one respondent noted that information about the availability of particular software is not enough, and instead a wider understanding about state-of-the-art software would be beneficial:

Knowledge of modern research software or research workflows is hardly circulating at all, and many researchers continue to do their work alone in their silos. This is likely to result in a fair amount of duplication of effort.

Training related to a specific type of software – survey tools – was mentioned by several respondents. One respondent suggested that to prevent "reinventing the wheel", an information package in the form of step-by-step instructions on how to conduct a survey would be useful. One respondent reported that support for a currently available survey tool, LimeSurvey, was dependent on the efforts of a few voluntary persons, and suggested that UH should acquire a licence for a software that has 24/7 online support, such as Qualtrics. Several respondents also felt that more training or better information should be available on how and where to store large amounts of data (especially if it is sensitive in nature).

4.2.3 RESEARCH DATA GOVERNANCE

Perhaps the most prevalent needs of the respondents concern the services and guidance on the handling of research data. In addition to the IT services and tools discussed above, respondents need legal advice especially on data protection and its implications for data collection and reuse; and rights, licences and agreements that regulate the ownership and use of data. Another type of required services is more technical advice or concrete help with organizing, describing and curating data.

Data protection and its implications for data collection, sharing and reuse are highly topical issues for the social sciences and humanities, as most data used by researchers are personal data. Yet, the practical implications of the new general data protection regulation (GDPR) are still unclear to many respondents. Information on GDPR has been available at UH, but it is rarely enough. The counsel of UH research lawyers is highly appreciated, but their services are often delayed. As a result, many respondents reported relying on colleagues' examples and practices rather than on professionals' statements. Respondents demand "active jurisprudence" for creating shared practices and preconfigured solutions in order to avoid the need for everyone to reinvent the wheel. GDPR is very problematic for foreign researchers:

Personalized advice on data management, transfer and storage [should be] made available to grant fund holders, rather than a one-size-fits-all approach. Appropriate training available at different levels (from novices through to those with existing advanced knowledge) on GDPR, the various platforms available for data transfer and sharing and different means of storing data. My impression coming here is that there is little understanding in the university of the challenges recently arrived foreign researchers have penetrating its systems of support – this needs to be vastly improved. The University of Oxford, where I was previously, is considerably ahead of UH in this area.

A common issue concerning the need for shared solutions is the status of personal data collected before the GDPR; many respondents wonder whether, how and by whom such data can be used. Resolutions are needed both for data collected by researchers themselves, and for data acquired from outside sources. Moreover, lawyers in different universities may interpret the legislation differently.

In general, the area of data protection is considered essential for every researcher, and training on the basics of GDPR and its implications for research are considered important. In addition to general training, many respondents would find general guidelines, model formulas and model contracts useful. One respondent suggested that training should include a workshop where participants could formulate the relevant documents. Data protection also raises very specific concerns where case-by-case assistance is required. Respondents mentioned, for instance, data protection implications for ethnographic data, or for using television broadcasts as research data. For ad hoc advice and case-by-case needs, the opportunity to directly contact specialists, e.g. in the form of on-call services, should be offered.

Other legal issues concerning research data include **copyrights, licences and agreements that regulate the ownership and use of data**. The practical implications of the regulation in these areas are complicated and the need for professional services is essential. Although respondents have received assistance from UH research lawyers, colleagues and national services such as CLARIN and CSC, there appears to be a need for additional services, especially case-by-case assistance as issues come up. Copyright and user right issues were among the most commented areas, and they touch upon the whole research process from data acquisition to publishing results and making data available for wider use. Respondents mentioned copyright issues related to, for instance, language corpora created from already published writings, publishing and using images, and international copyright issues. New forms of data, such as big data and co-creation are challenging:

> *Copyright issues related to co-development and co-creation of knowledge can be difficult. Universities, academic researchers, peer researchers and research subjects are all equally entitled to use the data.*

The minority of respondents considered training for researchers to be a solution to these issues. Instead of general training on legal issues, respondents emphasized consultations, workshops, training tailored to specific data types and information on who to turn to with these issues. One respondent commented that researchers should not be offered training at all, as it takes time away from research; instead, more hands-on service from legal professionals should be available. At the same time, some respondents would like to get training on topics such as immaterial rights, international copyright, using television broadcasts in research, using images, using data produced elsewhere and making agreements with partners.

Organizing, describing, and curating data is labour and time intensive, yet it is a very essential part of any research process with empirical data. Time-consuming data management tasks can be very frustrating to researchers who would rather spend the time on their subject matter.

Respondents commonly use, or would like to use, research assistants or other services for data management tasks, but it is not always possible due to e.g. insufficient funds or

the unavailability of services. Some areas of data management may also require special knowledge, which makes it even more challenging to find assistants. Support for organizing, cleaning, classifying and describing data will likely be of significance also in the future with continuing digitalization and changing requirements for open science, as one respondent pointed out:

> *It is certainly necessary in the future, as the potential for reuse of the data collected seems to be a competitive advantage in funding applications.*

Training could provide some solutions for these service needs. Although some respondents find this a technical and time-consuming task that could be done by research assistants (with training offered to them) rather than by researchers, some others feel that this is an essential area of knowledge for every researcher. One respondent commented that training in data handling should already be given during undergraduate studies. Information packages, workshops, "clinics" and personal assistance in more complicated issues were suggested as forms of training. This is yet again an area where "the wheel is reinvented" over and over again. Thus, information on what kind of support is available and sharing best practices within and between research areas could be useful. FIN-CLARIN and (FSD) were mentioned as providers of training in this area.

Making data openly available (incl. anonymization) is a new area to many researchers, but the theme is getting more topical due to increasing requirements towards open science and the digitalization of data. Researchers generally consider the opening of data a valuable aim, but they are currently facing many challenges. There is a lack of knowledge of what is a sufficient level of anonymization, especially with regard to sensitive data and video data, and in general, under what conditions research data can be made publicly available. Other challenges relate to questions of data ownership, which may prevent further usage of data, and the lack of suitable channels or services through which data could be made available. FIN-CLARIN and the Finnish Social Science Data Archive FSD are offering help with issues related to making data openly available, but respondents would like to receive assistance locally within the UH. Good tools for data management and anonymization are essential, but respondents would also require legal assistance with the more complicated issues, as well as technical assistance (e.g. in the form of research assistants) for anonymization.

A number of respondents would find general training on how to create open data and use databases, etc. beneficial, as this is a new field to many. Workshops and clinics could provide assistance with practical and more acute problems. Other respondents would rather purchase the service or train research assistants, because the tasks are time-consuming.

Surveying of data available from partners, open repositories, etc. is important to many researchers within the social sciences and humanities. Researchers use services offered by FIN-CLARIN and CSC (e.g. Language Bank) and many open repositories made available by research institutions and authorities, e.g. Statistics Finland, the Finnish Institute for Health and Welfare (THL), and the Digital and Population Data Services Agency (VRK), although this type of data can sometimes be expensive. According to the responses, services for improving data sharing between researchers, universities and many other organizations would be welcome. One problem taken up by the respondents is that sharing requires coordination, which is currently non-existent. This issue is relevant not only within the UH, but also between universities nationally and worldwide. Additionally,

mapping out and registering accessible data is essential. Sometimes information about data is available, but other issues make it difficult to obtain:

> *The materials are openly accessible in memory organizations, but in many cases detailed lists of, for example, archival collections, are still not available in digital format, which in practice means mapping the data by telephone and email inquiries. I would like to see support directed at memory organizations for publishing their materials.*

Researchers would also benefit from information about other researchers working with similar materials or research methods. Most respondents think they needed more information on what data was available and how to access them.

Research ethics is a central issue in the social sciences and humanities, as the research subjects often represent vulnerable groups, such as children, minorities and immigrants. New forms of data (e.g. social media and video data) require a novel understanding of their ethical implications in research. Some respondents find current requirements in research ethics hindering rather than being helpful and advancing research; for instance, ethical pre-evaluation of research is felt to be a slow process that is getting unnecessarily complicated. One respondent describes a complicated situation in the Finnish context:

> *Especially minors, particularly vulnerable young people, are very difficult to study in Finland, even if the purpose is to make their own voice heard. Part of the reason is the strict requirement for parental consent. For example, parents of young people with an immigrant background may not understand the consent letter (it costs enormously to translate it into all the languages needed), but in order to improve their education and adaptation, it would be important to get their voice heard. In other countries, the number of young people with an immigrant background already ensures that researchers get the necessary samples.*

Training on ethical issues already in early stages of a researcher's career (or even during undergraduate studies) is of importance, but such general training alone is not enough; every research project is different and different research areas have their special characteristics. Therefore, more tailored services, either in the form of training or case-by-case professional assistance, are also needed.

5 ATTITUDE TOWARDS DATA SHARING

This section of the questionnaire first surveyed researchers' attitudes towards data sharing in general, and then more closely the reasons why researchers have not been willing or able to share data. The first question, *If you produce research data (as a researcher or in your research group), what is your view on sharing data with others?,* consisted of four answer options: *We have shared data with others or have agreed on the shared use of data; We have not yet shared data, but are, in principle, willing to do so; We are not willing to share data with others;* and *Other view, please specify*. Respondents were able to choose multiple categories, and it was also possible to specify selections in an open comment field attached to each category.

The number of selections and the percentage of respondents in each category are presented in Table 9. The responses were also analysed according to the respondents' faculty or unit, but there appeared to be only minor differences between them. The distribution of answers according to faculties is presented in Appendix 7.

Table 9. If you produce research data (as a researcher or in your research group), what is your view on sharing data with others? The number of responses and their share of the total number of respondents (N=194).

View on data sharing	Count	Percentage of respondents
We have shared data with others or have agreed on the shared use of data	89	46%
We have not yet shared data, but are, in principle, willing to do so	65	34%
We are not willing to share data with others	21	11%
Other view; please specify	19	10%

46% of the respondents reported that they *had already shared data or have agreed on the shared use of data*. Several researchers report sharing data as a rule, unless there is a cogent reason not to do so (e.g. the research subjects have not given permission to use data in other studies), while a few researchers noted that the decision to publish data is based on their assessment of its quality or attractiveness to other researchers. Data is commonly made publicly available after the respondents' own research has been published. Respondents report that they have shared data openly through platforms such as Language Bank, figshare.org or the National Library, or to a restricted population, e.g. inside their own research group or consortium, or with graduate and postgraduate students.

34% of the respondents reported that they *had not yet shared data, but are, in principle, willing to do so*. The most common concerns among the respondents were the need for careful anonymization of the data, and the implications of GDPR on contractual matters. Some respondents commented that they would be willing to share data to a restricted audience if appropriate contracts were made, or that they would be willing to share some

part of their data. Researchers are also concerned with more practical matters, such as receiving support with data anonymization and publication.

Only 11% of the respondents reported that they *are not willing to share data*. The sensitive nature of data (e.g. health information, vulnerable research subjects) was the most common reason mentioned in the free-form comments. 10% of the respondents reported having an *other view* on data sharing. A few researchers commented that the ability to share data depended on the research project and the partners involved in it, and therefore the question was difficult to answer. For some respondents, data sharing has not been relevant so far.

The second question in this section of the survey, *If you have not shared research data with others, why not?,* consisted of seven answer options*: Data sharing involves legal or ethical problems; The reuse of data involves risks, please specify; We have not sufficiently familiarized ourselves with the legal, ethical, practical and other conditions of sharing data; Transforming data into a format that others can use is not worthwhile; It requires additional work for which we do not have sufficient resources; We wish to keep our data for ourselves;* and *Other reason, please specify*. Again, respondents were able to choose multiple categories, and it was possible to specify selections in an open comment field attached to each category. The number of selections and the percentage of respondents in each category are presented in Table 10, and the free-form comments are discussed thereafter.

Table 10. If you have not shared research data with others, why not? The number of responses and their share of the total number of respondents (N=194)

Reason for not sharing data	Count	Percentage of respondents
Data sharing involves legal or ethical problems	72	37%
It requires additional work for which we do not have sufficient resources	41	21%
We have not sufficiently familiarized ourselves with the legal, ethical, practical and other conditions of sharing data	27	14%
The reuse of data involves risks; please specify	18	9%
We wish to keep our data for ourselves	18	9%
Transforming data into a format that others can use is not worthwhile	17	9%
Other reason; please specify	14	7%

The most commonly chosen category with 37% of the respondents selecting it was *data sharing involves legal or ethical problems*. The vulnerability of research subjects and/or the sensitive nature of data limit the ability to provide data for further use. The ethics board may pose restrictions on the use of sensitive data. In some cases, it is not possible to provide a sufficient level of anonymization that would protect the privacy of the research subjects:

> *It's about ethics: the data are such that [the research subject's] identity may be revealed, at least to the person's next of kin, even if they are anonymous.*

Another commonly mentioned issue is that researchers do not have permission, either from research subjects or from other relevant parties, to share the data. Sharing register data, for example, is not allowed by most registrars. Archaeological data, as another example, are usually copyrighted by local authorities, and data sharing must be approved by them.

A little over one fifth of respondents reported that sharing data *requires additional work for which they do not have sufficient resources*. There are challenges especially with qualitative data, and the possibility of misinterpretation sets high requirements for preparing the data for further usage. Funding for this purpose is often not available, and new projects typically start right after previous ones. For these reasons, data is not shared, even though it is in the interests of researchers.

14% responded that they *had not sufficiently familiarized themselves with the legal, ethical, practical and other conditions of sharing data*. The respondents often lack time, or do not even know where to look for information.

9% of the respondents reported that *the reuse of data involves risks*. Two types of risks stood out. The first type of risks relates to the security and vulnerability of research subjects, involving the possibility of the misuse of information, discrimination and even physical threat to them. The second type of risks brought up by the respondents relates to the misunderstanding of the data. As mentioned earlier, qualitative data may be hard to make sense of after comprehensive anynomization. Quantitative data, in turn, requires at least statistical literacy, and sometimes understanding the underlying theories and methodologies behind the data is essential, too:

> *For example, if the data is intended to carry out a particular type of intervention, it may act against its purpose in the hands of someone who is not familiar with the theory, methods and practical skills of the intervention.*

9% of the respondents state that *they wish to keep the data for themselves*. According to the free-form comments, however, this is often the case only as long as they are using the data themselves. Some respondents again raised the necessity of understanding the nature of the data and the context in which they are collected, as well as the confidential nature of their data.

9% of the respondents state that *transforming data into a format that others can use is not worthwhile*. This can be due to the lack of resources, or because the transformation would not save any resources compared to the situation that other researchers would collect new datasets themselves.

7% of the respondents selected *other reason* for not sharing data. Most comments deal with issues already mentioned in other categories above. One respondent highlights an issue specific to archaeological data:

> *Understanding archaeological material without an interactive user interface is often unnecessarily challenging; simple file sharing is not user-friendly.*

Another respondent commented that in international cooperation, possibilities and restrictions for data sharing in different countries or universities must be determined.

6 OVERVIEW OF INFRASTRUCTURAL DEVELOPMENT IN SOME DATA-INTENSIVE RESEARCH AREAS

The last part of the questionnaire dealt with respondents' future visions, expectations, and benchmarks for research infrastructures in their fields. It consisted of three open-ended questions: *What future visions are there in your field for the production, acquisition, use or analysis of research data? How in the future could the Helsinki Institute for Social Sciences and Humanities and the research infrastructure activities pursued under its auspices promote the above visions? What international or national examples of good research infrastructures would you like to highlight in your research field?* The number of responses was 90, 74 and 58, respectively.

In what follows, we do not report the results question by question, but summarize them into domain-specific overviews of the state of affairs and future prospects in four overlapping clusters of data production and processing technologies, or data-intensive research areas: (1) digital humanities, (2) social data science, (3) audio-visual research, and (4) experimental research (Figure 11). The identification of these clusters draws on the responses to the three open-ended questions (above), combined with the responses discussed in section 3.5. Each cluster contains a range of methodological approaches, data types, etc., which can also be understood as distinctive data- or infrastructure-intensive areas themselves (cf. section 3.5).

These clusters are not defined by traditional disciplinary boundaries, but by research practice. As a result, many disciplines and research units can be considered belonging to several clusters. For example, research on language spans across three of the above clusters: Some dimensions of language research, such as phonetics, speech synthesis and theoretical modelling of language processing, are heavily dependent on *experimental research* and thus require a particular type of research infrastructure in order to conduct these experiments. Some other dimensions of language research, such as language technology and computational linguistics, are often considered part of the *digital humanities* to which they provide important methods and tools. Still other dimensions of language research, focusing on spoken interaction and conversation analysis, represent *audio-visual research*, as they work primarily with audio and video material, the production and analysis of which entails yet another type of research infrastructure.

The results reported in this chapter partly overlap with those reported in chapter 4, as respondents' views on infrastructural development are tightly coupled with their current needs and observed shortages (see Figure 12). With the risk of repeating some findings, this chapter discusses those needs and desires in the context of particular clusters of data production and processing technologies.

Figure 11. Four overlapping clusters of data-intensive research in social sciences and humanities

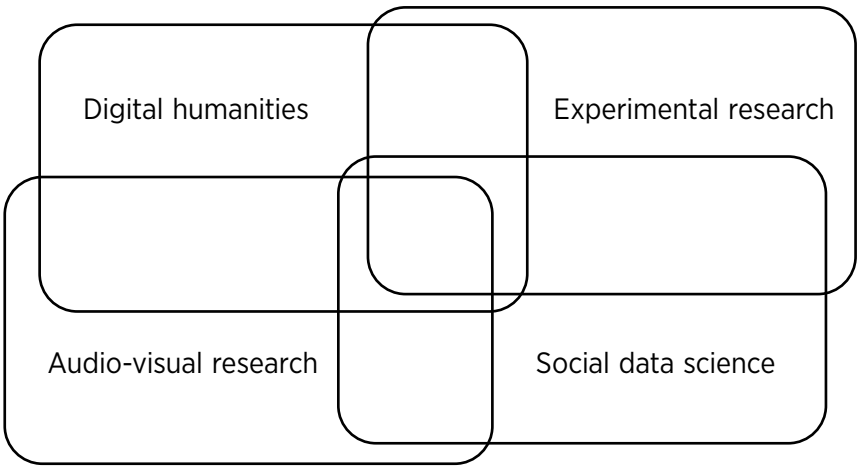
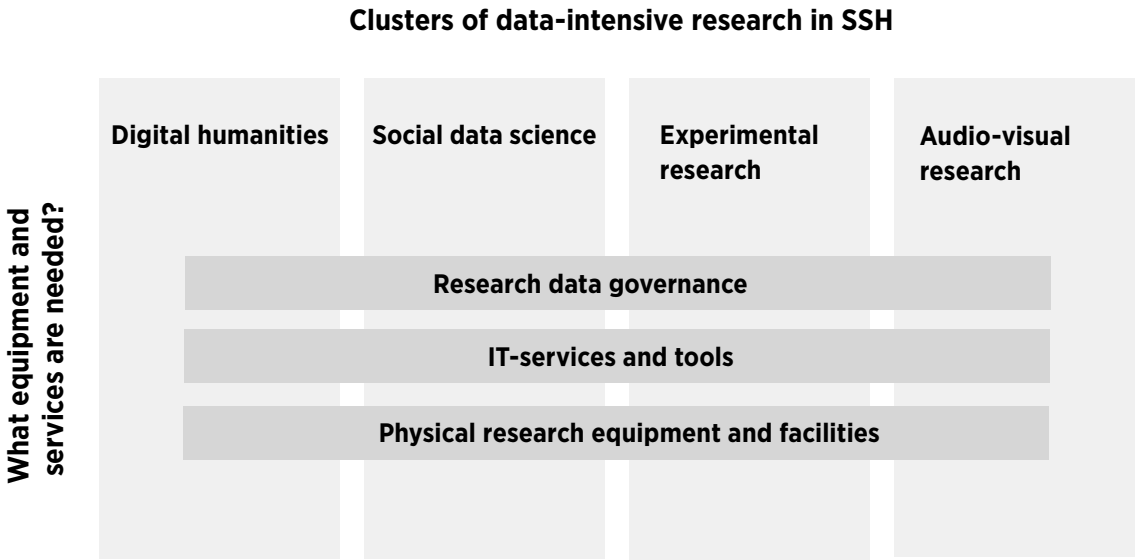


Figure 12. The specific needs for research infrastructures vary across the clusters of data-intensive research.



6.1 DIGITAL HUMANITIES

This set of data production and processing technologies has to do with many research fields in the humanities, covering a range of language studies, literature, history, archaeology, cultural studies, arts, etc. (see also Table 5). In addition, many branches of (interpretative) social sciences – including e.g. legal science, political science, and media studies – rely on approaches and tools developed in the digital humanities. As indicated above, "digital humanities" is an umbrella term that encompasses a multitude of approaches and respective research infrastructures, which are partly overlapping and can also form nested structures. Digital humanities is, nevertheless, an established concept in UH and elsewhere, and is institutionalized on campus in e.g. HELDIG (Helsinki Centre for Digital Humanities) and the Faculty of Arts' Department of Digital Humanities.

At its core, digital humanities is about the application of modern data processing to solve humanities research questions. Characteristic to this category is the use of digital tools and methods in studying humanities source materials, such as literature, documents and archives, interview data and archaeological objects (see also Table 6).

A necessary starting point for the progress of digital humanities is **the availability of source materials in a digital format**. Most humanities textual (and numerical) material in e.g. history still exists only in printed or handwritten format, but the development of tools for e.g. optical character recognition (OCR) and handwritten text recognition will soon radically change the situation. Digitization, however, is also labour-intensive work, which requires investments:

> *Digitalization is a major opportunity that will change the nature of the entire research field, as in principle all source material will be immediately accessible to the researcher, without having to spend hours in libraries and archives.*

One of the key factors enhancing the availability of digital material, and the development of software for transforming originals into digital format, is close **collaboration with organizations that collect, keep and deliver such materials**. Important partners are especially memory organizations and FIN-CLARIN, but to some extent also public authorities and third sector organizations (see Appendix 3). Also centrally managed (remote) access from the university to the data is hoped for:

> *The Institute should support and host national infrastructures for negotiating and managing access to datasets of common interest. In addition, those datasets need to be centrally documented and curated for biases and other aspects relevant for research use.*

The ongoing digitalization can also enhance collaboration between organizations in **linking data from various sources**. Global networks of infrastructures in language studies, for example, or conjoined digital artifact collections and archives, are envisaged. It is also anticipated that the university might play a role in publishing research materials.

Not all important materials will be available in digital format via memory organizations, etc., but researchers also need **digitization software** in their own computers:

Historians really need OCR software on their computers, and the ability to easily share photographic archives with each other. If such software and resources were to come out, historians would also need to be slightly educated in, for example, that each stored photograph must have metadata, e.g., so that the archive ordering tag of the photographed document appears in the image.

There are high expectations of **software for automatizing data processing**, for both digitized data and born-digital data. In the context of the digital humanities, respondents report progress especially with processing textual data. (For audio and video data, see section 6.3). Currently, the cleanup, refinement and enrichment of raw data constitutes a significant part of researchers' work. Automatizing and streamlining data processing would release researchers from much mechanical work and thus enable them to focus more on intellectual substance:

The [HSSH] infrastructure should focus as many resources as it can on providing solutions to the current effort overhead of researchers. Only when researchers are free to spend a lot of time on their core research questions can they produce the outputs which are counted in assessment exercises and tenure track evaluations.

In the future, research of history, and especially computational research of history, will increasingly focus on digital research data. Most people in the field have not yet fully realized that cleaning up, harmonizing and processing data make up the largest part of the research process. However, this is going to change, and scientific research that centres around empirical data will be successful in the SSH field. Open science will also further develop and enhance the research processes. In addition, methodological development in terms of data analysis and use will change in the future when it will be done on the conditions of SSH research rather than by borrowing methods from other disciplines.

An important step forward is the development of **speech recognition**, also for the Finnish language, which enables automatic transcription of audio-recorded interview material into text. Given the central role of interviews as source material across SSH (see Table 2), the development and mobilization of automatic transcription software is a huge opportunity to save researchers' time doing manual transcription work:

It is good to think about how it would be possible to conduct surveys and interviews at a reasonable price and within a reasonable time. The vision is that in the future there will be devices that record, transcribe and start the analysis, under the direction and control of a researcher, of course! Maybe such devices are already available.

Creating an infrastructure that will support and encourage the storage and **reuse of not only raw data, but also variously edited, annotated and enriched data** is a central area for future development in the digital humanities. A related challenge is the development of services that enable multimodal data to be linked together and visually

presented. While the potential of sharing, linking and reusing digital data is high, the shortage of technical solutions (e.g. storage space and tools for data cleanup) and funding are currently hindering the development of open science practices. Perhaps the biggest challenge, however, is the required change of research culture:

> *Over the last couple of decades, we have had masses of projects that have produced a variety of research materials. In my opinion, much of them are still underutilized, and we have not developed a widespread culture of using each other's materials. . . We actually have no examples of research data sets collected by one researcher but enriched by another. If the process were to function normally, new levels of analysis would emerge for the old data, which would also be archived and made available to others. In my opinion, this would be the most important single step forward: creating a culture where we regularly use each other's materials.*

Copyright issues arise throughout the research process (see section 4.2.3), but they are especially relevant when sharing data for further usage. A central infrastructure could provide efficient resolution of copyright issues with legal counselling services and easier access for researchers to relevant information.

Another huge opportunity in the digital humanities is the development and use of **computational methods and machine learning for the analysis of textual data**. The use of such methods often requires pre-processed text material and thus benefits from automatized data processing tools (see above). More importantly, however, it requires an understanding of the opportunities and limitations of such methods, and typically also collaboration with computer scientists. In this area, there is an obvious need for training staff to acquaint themselves with new methods.

> *The Institute should facilitate the renewal of research by developing an educational curriculum to inform staff of the potential of computational approaches. Further, instead of being done separately in a vacuum, to truly be of benefit, the needed work to develop computational tools and research protocols has to be done in close collaboration with the end-user researchers in the social sciences and humanities.*

> *Therefore, it is imperative that the Institute does not function merely as a technical support organization, but engages in and enables also transdisciplinary research into the novel computational and statistical methods needed for robust and trustworthy analysis targeting such complex data and questions.*

Examples of good research infrastructures in this domain, given by the respondents, are the pan-European infrastructures *CLARIN* and *DARIAH* and their national consortiums. International and national cooperatives within the domain of language studies are the *National Institute for Japanese Language and Linguistics (NINJAL)*, the *Japanese Historical Text Initiative, the Richly Annotated Cuneiform Corpus (Oracc)*, and the *Language Bank of Finland*. Other good examples given by the respondents include the international linked open data service providers the *Getty Foundation and the American*

Art Collaborative (in the domain of arts), and *Linked Data Finland LDF.fi* and *ONKI.fi* in Finland. Services that provide centralized access to various data collections include the National Library's *Finto-fi* and *Finna.fi*, as well as the *Europeana Collections* that contain digitized items from European archives, libraries and museums. *National Library's Fenno-Ugrica* offers a digitized collection of newspaper articles and monographs in the Uralic languages. *CSC IT Centre for Science* provides ICT expert services for higher education and research institutes, etc., which are developed through national cooperation.

6.2 SOCIAL DATA SCIENCE

Social data science is used here as an umbrella term that encompasses a number of data-intensive approaches in the social sciences. Among the target group, some important sources of social data are surveys, registers, social media applications and online platforms (see also Table 6). As was the case with the digital humanities, the definition of social data science also involves conceptual ambiguities. For example, the conduct and utilization of large-scale social surveys require tools and procedures that differ from those needed by e.g. register-based or computational social sciences. But at least in the context of UH, researchers using these approaches increasingly consider themselves to be doing social data science. An indication of this is the recent establishment of the Centre for Social Data Science (CSDS) in the Faculty of Social Sciences.

The prospects for infrastructural development in social data science are, to some extent, similar to those in the digital humanities. One of the most important progress areas is **improved access to restricted generally interesting data**:

> *Focus should be placed on materials useful to as many researchers as possible. This might take the form of centrally negotiated and managed research access to restricted generally interesting data, such as books, newspapers and media archives. For social media, this might mean centrally managed facilities for harvesting data from common platforms of interest.*

What is perhaps distinctive to social data science is its focus on accumulative, "living" data in its contemporary institutional and societal context. As more and more such "naturally formed" data on human and social phenomena are available, social scientists are increasingly inclined to harvest and utilize such data, instead of – or alongside of – generating new research data through e.g. questionnaires or field observations. As opposed to most humanist research materials, social data are often born digital, so infrastructural investments are not needed for digitization. Instead, new technical tools and services are needed for **harvesting relevant data** from e.g. registers, websites and online platforms, social media applications and through sensory capture. According to respondents,

> *Data gathering should be supported by a flexible physical and digital infrastructure, which again should minimize the effort and errors of researchers.*

One of the respondents' worries was that:

> *While there is more data existing, there is less data available – private corporations have huge budgets for data acquisition; we cannot compete. So we really would benefit from pooling resources and sharing – but this is time consuming and not merited enough.*

Data sharing is, indeed, one of the progress areas of social data science – and of many other domains, too, given the push for open science. This is expected to take place through the vehicle of joint, international platforms, and is also expected to be standardized and streamlined to minimize repeated effort and errors. Common shared aspects, like those covered by Data Management Plan, should be managed by a shared infrastructure as much as possible. On the analysis side, however, workflows and tools should be designed and applied in consort with their end users. According to one respondent:

> *The issue comes down to the fact that for interpretative social science and the humanities, based on "found" data not originally created for research, there exist no ready, tried and thorough research protocols. Instead, the field is still very much in an explorative state, with much need for genuine methodological development in addition to mere application and services.*

On the other hand, respondents call for a more critical attitude towards ever-increasing amounts of (social) data. It was stated that there is too much data, while the most important thing – interpretation – is often missing. Interpretation of data requires the cultivation of a **more critical, theoretically driven perspective to data analysis**. To this end, some foresaw a broader arsenal of quantitative methods in teaching programmes, including e.g. mixed methods approach and structural equation modelling in describing and modelling complex societal phenomena.

Another critical standpoint was about the sense of making data available:

> *Too much time is spent on making data accessible (metadata, DOIs, etc.) and listing them on multiple repositories, but in the end almost nobody uses most of these datasets. This process needs to be streamlined instead of building more and more tools to "make data available". The biggest challenge is to keep datasets available: What happens if a link to a dataset in a published research paper becomes unavailable? What happens if researchers move abroad and lose their permissions to edit university-internal websites or national service providers?*

Given that most scientifically interesting social data include **personal identifiers**, the use and reuse of such data causes a great deal of trouble and work (see section 4.2.3). For example, respondents are unsure about their legal rights and obligations in using data from social media platforms. To what extent are they entitled to use data from e.g. Facebook groups in their research? And how should or could they inform the research subjects about their study? Again, these aspects could be better managed by a shared infrastructure and instructions rather than by individual researchers.

As the data-intensive approach is becoming more common in fields less acquainted with empirical research, such as law, the need for infrastructural support for methodological, technical and management aspects of data processing proliferates:

> *The field is becoming increasingly empirical, therefore, legal data needs to be handled with even more care concerning data protection and processing. Court cases and case law require different methodological tools that need to be developed in relation to e-infrastructures (e-justice, for example).*

A particular source of social data is **registers**. Unlike with "new", often unstructured and unstable data from social media and online platforms, the tools and protocols for dealing with register data are well established. While register-based research is considered a special strength of the Nordic countries, Finnish research infrastructure in this area is lagging behind e.g. Sweden and Denmark. According to one respondent, the objective of such research infrastructure would be to link the most important Finnish population and administrative registers together. The creation of such a database would require collaboration between researchers from different fields, as well as close communication with the registrars, such as Statistics Finland, the Finnish Institute for Health and Welfare, Kela, VRK and the Finnish Defence Forces. Currently, researchers and research groups have access to various limited datasets, generated for specific research projects, but such limited datasets do not allow for comprehensive, population-level social scientific research:

> *The HSSH infrastructure could serve as a unifying structure for researchers doing register-based research at the City Centre Campus, within which the overall data set, described above, could be created in co-ordination with the registrars. For example, a data manager / statistician could be employed within SSH to support researchers in data management and analysis when the data becomes available.*

According to another respondent, a key problem with combining register data from multiple sources is lengthy and **complicated licence processes**, especially when register data is combined with interview data. In place of the prevailing licence system, they would hope for a “one-stop shop” for applying licences. Despite its current shortcomings, respondents state that register-based research is gaining popularity, especially as open access datasets become more frequent.

Another development related to data sources is that conducting **surveys** is becoming more common. As survey tools and knowledge on how to conduct a survey become essential to a growing number of researchers, well-maintained survey software and centrally offered training seems beneficial (as opposed to “reinventing the wheel” every time researchers have to familiarize themselves with conducting a survey).

Examples of good research infrastructures in this domain, given by the respondents, include linked population and administrative register data providers, such as the *Department of Medical Epidemiology and Biostatistics* at the Karolinska Institutet and *Lund University* in Sweden, and the *National Centre for Register-based Research* at Aarhus University in Denmark. Multinational surveys include the *European Social Survey (ESS)*, *PISA* and the *World Values Survey,* and *Suomi24* data is an example of social media data collection. *The Finnish Social Science Data Archive (FSD)* provides a national example

of a data archive within the field of social sciences. The *CSC IT Centre for Science and the UH Department of Computer Science* (in Kumpula) offer examples of infrastructures that help develop computational methods and tools.

6.3 AUDIO-VISUAL RESEARCH

Empirical research based on audio-visual materials – video and audio recordings, images, film, new media, etc. – emerged in the survey as an important domain which is highly dependent on local physical infrastructure, on the one hand, and on shared practices and tools for processing and managing data, on the other. Audio-visual materials are used in several fields that span the faculties and units of UH City Centre Campus, and thus the common needs for infrastructural support by researchers working with such materials have not as yet been heard. Audio-visual research was also not recognized as an "infrastructure-intensive" research area in the design of this survey (see section 3.5), but only emerged from the responses. Moreover, the use of audio-visual material was not explicitly inquired about, but it was hidden in the categories *observation data*, i*nterview data*, *media material*, *social media material*, and *other type of data* (see section 3.1).

Researchers actively using audio-visual materials in their knowledge production come from e.g. language studies, educational sciences, and social and cultural anthropology[1], and focus on phenomena such as spoken interaction, conversation dynamics, visual culture and (technology-mediated) social interaction.

A critical prerequisite for audio-visual research is the **availability of good equipment and training** for collecting materials. As indicated in section 4.2.1, there is a clear shortage of audio-visual equipment, and even more so of guidance and training in the use of such equipment. Special facilities for video interviews and video observations are lacking, too:

> *Video and audio recording are crucial elements of my research and central for my discipline, both for data collection and the creation of outreach materials (films and podcasts). The availability of good equipment and training are of prime importance.*

> *It would be great if in the future there were better and tailor-made facilities for collecting material (such as video observation and interviews) that could be reserved by researchers… . In addition, the opportunity to utilize the programming knowledge needed to code mobile research programs that can collect data would be great!*

Respondents highlight the need for **technical expertise** to take care of the purchase, upkeep, training and shared use of audio-visual equipment. This concerns not only physical

1 In autumn 2019, independently of this survey, the teaching/planning team of Visual Anthropology Visual Methods in the Social Sciences (YMV-A514) mapped the existing teaching and research infrastructure and praxis available at the Faculty of Social Sciences, as well as the future needs at the Faculty. They received about 30 responses, and found, among other things, that there is a clear shortage of good equipment, while the use of audio-visual materials and methods is increasing in most disciplines in the Faculty of Social Sciences.

equipment, such as cameras, video cameras, etc., but increasingly also the software for effectively using this equipment in e.g. novel research settings, such as synchronizing speech and movement, eye tracking, etc. (see section 4.2.2).

Perhaps the most urgently needed infrastructural support for audio-visual research is the development of shared practices, tools and instructions for **managing audio-visual material**. Currently, the lack thereof means that a great deal of valuable material is kept on researchers' own laptops, memory sticks, etc., with no opportunity to be effectively shared, stored or reused for cumulative advances of e.g. research on intersubjectivity in interaction. Thus, according to one respondent, the availability of Finnish-language speech material, especially conversational data, is quite limited at present.

The combination of large size (in terms of megabytes) and high sensitivity (in terms of GDPR and privacy) makes video and audio data very challenging to store, share and reuse. Firstly, **secure storage capacity** is needed for these data types:

> *Conversation analysis continues to be based on the analysis of authentic interaction situations, so there must be a place to store massive video footage... The Language Bank provides services for data sets that are organized and prepared for research, but we also need environments for storing large amounts of insufficiently annotated raw data, which in principle can always be enriched into better described new research material.*

Secondly, the **long-term digital preservation** of audio-visual data, including the curation of data formats, should be arranged:

> *There is currently no operational [body] in Finland which is responsible for the systematic archiving or long-term preservation of audio-visual material collected in Finland... . In my view, most of the audio-visual material collected at the university is currently not properly archived anywhere. On the other hand, this has apparently been the case for decades. Your research infrastructure should provide clear mechanisms and guidelines for data collection, storage and archiving / long-term storage. This could be linked, for example, to the infrastructure now provided by CSC, because we have a lot of good examples of long-term storage systems (PAS), operating environments (IDA) and authentication (e.g. LBR), but I don't really know who is pulling the strings here.*

Thirdly, the storage, reuse and long-term digital preservation of audio-visual data would require **shared practices and standards for organizing and describing** the data. According to one respondent's vision of the future:

> *The data collected in interaction research would be available to researchers in the field as widely as possible. They would be stored so that they were preserved for future generations of researchers, which, in turn, would open up new opportunities for e.g. comparative research. They would be organized and described in a uniform manner, and their distribution would be possible and also secure in terms of data protection. Researchers would have solid expertise*

in data protection and research ethics related to the collection and use of this type of material.

Another concern of respondents is the **tightening ethical regulation** of the collection of video material of human subjects. The collection of authentic video material is subject to licences from research subjects, and in many cases also from their background organizations, such as schools, day-care centres, municipalities or companies. Several respondents report having increasing trouble in acquiring these licences in the first place, or getting permission to reuse already collected materials in another project.

Increasing legal regulation may make it more difficult to obtain and process data to a point where even ethically sound, important basic research may not be realized.

For example, the video and eye tracking data we collect is raw data, the analytical potential of which may probably never be fully utilized. From the perspective of research ethics, it is an interesting question how data collected for a specific purpose can be utilized later.

Examples of good research infrastructures in this domain, given by the respondents, include institutions offering storages and management of speech material, such as *Institut für Deutsche Sprache*, and *NINJAL* in Japan, which collects longitudinal research material. *Hamburger Zentrum für Sprachkorpora* offers an example of an easy distribution of old research material, while *PARADISEC* in Australia is mentioned as a good example of a digital archive of small languages in the world. IT services at the *Max Planck Institute for Psycholinguistics* is given as an example of a provider of much-needed support in managing audiovisual material (specifically in language research experiments and speech recordings in natural settings).

Good examples of infrastructures in the national context include the *Finnish Centre of Excellence in Intersubjectivity in Interaction*, which is, according to one respondent, an example of a well-established framework supporting researchers with data collection, processing and other practical issues. The *IDA Research Data Storage Service* is an example of a Finnish service for saving, organizing and sharing data. According to one respondent, *The Institute for the Languages of Finland (KOTUS)* has licensed its own material sensibly.

6.4 EXPERIMENTAL RESEARCH

14% of the respondents report using primarily experimental data in their research (see section 3.1), and the same percentage consider that their research represents experimental social and behavioural sciences (see section 3.5). These respondents were quite evenly distributed across different faculties and units, and represent e.g. phonetic, didactic, cognitive, linguistic, behavioural, social psychological, learning and political research.

One trend that is especially reflected in experimental research is the **proliferation of ways to capture and use empirical data**. Measuring physical changes in humans in

different experimental research settings and the strengthening alignments of "experimental philosophy" and "metascience" are just a few examples of this trend.

Respondents highlight the benefits of **maintaining subject pools and registers of participants** in trials and experiments across various disciplines. Such registers would allow researchers to seek out suitable participants in a cost- and time-effective manner. An example of an ideal register would include the registers would include a wide variety of participants, such as citizens, politicians, journalists and public administrators.

> *The state-of-the-art is currently survey experimental research based on pre-recruited online panels with citizens and other actors in society. The use of these panels build on a time-share logic and are able to serve many projects and individual researchers simultaneously.*

One important infrastructural need concerns **special facilities** for conducting experiments (see also section 4.2.1). For example, one respondent argued that laboratory facilities with cubicles are needed for experimental research methods. According to this respondent's description:

> *Cubicles should be in the room so that the presence of research assistants does not bother participants' experience of privacy… With the current university research infrastructure one cannot sensibly compete with international research groups. It would be good to have about 16–20 cubicles to allow for continuously running experiments of several research groups (one experiment would require about 4 cubicles). If there are several experiments continuously running in the lab, participants can perform experiments and help several research groups at the same time.*

Some concepts of experimental research, such as the "operations room" concept, are argued to have a rather broad applicability, thus justifying the creation of specific facilities that would enable the observation of (e.g. social, cognitive, policy, etc.) situations and exercises in an experimental setting.

As with audio-visual research, respondents highlight the need for **technical expertise** to take care of equipment used in experimental research, and to develop methods and software for running experiments:

> *Many methods require special technical expertise and equipment. Our eye tracking data is only one example. Long-term development of technical expertise is challenging, as funding is sporadic. Funders are also not very keen on financing the longer-term recruitment of technical staff.*

In relation to the increasing demand for laboratory and special facilities, researchers envisage that HSSH will provide **outsourcing services** to researchers conducting experimental research. The availability of centralized support could free their time and resources so that they can concentrate on intellectual substance. As one researcher formulates it:

> *HSSH could provide "laboratory services" to projects and units in such a way that, for example, a project utilizing sensor data could purchase a service containing data protection, ethics review, data storage, etc. Thus, small research money would not need to be spent on training data and technology experts before moving forward. There could also be a consulting service for this. In addition, basic equipment maintenance, lending and instruction.*

One future vision within the area of experimental research is better availability of **data collections and analyses conducted with exactly the same tests or methods**. Progress in this area could, among other things, increase the interest in and citations of domestic articles.

An increase in the use of web-based experiments is also anticipated, while support for "old-school" experimental research is decreasing. One respondent considered that:

> *[Future visions] are hard to predict, as methods are in upheaval due to the replication crisis. Demands are changing a lot when [experimental research] struggles out of old practices. It is probably the case that old-school experimental research will hardly be supported anymore.*

More generally, respondents suggested that HSSH could **actively bolster experimental research** in the social sciences and humanities:

> *[HSSH] could strengthen Finnish research in behavioural economics, decision making, experimental philosophy, cognitive science, social psychology and game theory, in ways that would open avenues to journals that have rarely published research conducted in Finland. This would also clearly boost the rigour of scientific thinking in Finland. In Finland, the understanding of experimental research in the social sciences and humanities has been quite weak far too long. It is automatically considered as evil, without even knowing what is being condemned.*

Good examples of research infrastructures within the domain of web-based experimental research and the utilization of citizen panels, given by the respondents, include the *Digital Social Science Core Facility (DIGSSCORE)* at the University of Bergen, which integrates and combines survey studies and laboratory research through the Norwegian Citizen Panel and the Citizen Lab, and the *Laboratory of Opinion Research (LORE)*, which is an organization within the University of Gothenburg conducting data collection through web questionnaires. Another example in this domain is *Time-sharing Experiments for the Social Sciences (TESS)* in the U.S., which uses a national panel to provide representative samples for survey research. A domestic example of a state-of-the-art infrastructure for experimental research is a neurophysiological measuring infrastructure collectively built by UH and Aalto University, which also has international benchmarks. Yet another specific example given by the respondents is a research unit at the Department of English and Linguistics at the Johannes Gutenberg University Mainz, which has taken a multimodal and interdisciplinary approach to examining linguistic complexity. The unit investigates the empirical validity of the postulated rules for Easy Language, combined with evidence from linguistic complexity research and evidence-based development of these rules.

7 CONCLUSIONS

The results of the survey show that there is an evident need for developing the local research infrastructure in the social sciences and humanities at the University of Helsinki. The proliferation of digital resources and tools, combined with the movement towards open science, are profoundly changing the environment in which these fields operate. Research is increasingly based on multiple empirical source materials and new computational and mixed methods, and data are collected and used in heterogeneous research constellations. However, these developments have not permeated the practices of SSH fields as much as they could, and many researchers hope for better-equipped research environments to effectively harness the new opportunities. This concerns not only the availability of modern technologies, tools, methods and data, but also shared rules and protocols for using them for different scholarly purposes.

The need for centrally managed solutions for collecting and managing research data is heightened by the new demands from legislation and regulation. One of the biggest challenges is the implementation of GDPR, which concerns a major part of research in SSH fields. Both the juridical interpretation of GDPR with regard to specific data sets, and its demands on the technologies for handling and transferring sensitive data, remain outside the scope of researchers' expertise. While national SSH research infrastructures have formulated guidelines and instruction, local services and solutions are urgently needed.

In regard to developing the university's research infrastructure and services for SSH research, what role should HSSH take among other relevant actors? Drawing on the survey results, and some preliminary discussions of them with both researchers and service providers, HSSH could have two main functions. First, it could provide researchers with support for accessing and using services, data and tools that are distributed across different service providers locally and nationally. Second, it could coordinate and contribute to the integrated development of UH and national resources for data-intensive SSH research.

Concerning the first function, important service providers at the local level, relevant for all SSH fields, include the Centre for Information Technology, Helsinki University Library, the National Library of Finland, and several teams of the University Research Services (e.g. research lawyers, project coordinators, laboratory services). HSSH could act as an intermediary between its founding academic units and these service providers, helping the units to keep abreast of recent developments in the local service infrastructure, on the one hand, and urging service providers to design services that better match what researchers need, on the other. At the national level, the closest research infrastructures are FIN-CLARIN, FSD and CSC. Other relevant actors include e.g. the National Archives of Finland, Statistics Finland, and many government research institutes and universities. With regard to these actors, HSSH should ensure that researchers are aware of their services and equipped enough to utilize them.

The second function proposed for HSSH implies active development of data-intensive SSH research in concert with relevant national actors. This requires strategic choices from the HSSH founding units or UH to allocate resources for infrastructural investments in some key areas of common interest. In addition, HSSH could help researchers with similar interests to self-organize and apply external funding for infrastructural development.

From the perspective of these roles, some of the most important actions that HSSH could take are the following:

- Mapping the current reserve of relevant research instruments, equipment, databases, materials, software and facilities as well as related services available for researchers either in UH or e.g. via national research infrastructures, and **establishing a portal or "one-shot-shop service"** for finding or accessing this material. A related action is collectively defining the concept of research infrastructure and identifying its relevant components in SSH fields.
- **Gathering together the existing (mobile) research equipment** and related software that is currently diffused across the campus, arranging its shared use and upkeep, and coordinating or centralizing the purchase of new equipment and licences.
- Strengthening the juridical, ethical and technical **guidance and services for research data management**. The existing support from the UH Research Services, the Centre for Information Technology and Helsinki University Library is not sufficient, but requires additional resources for addressing these problems from the perspective of SSH research. This involves e.g. centrally managed solutions for collecting, storing and sharing different types of data with personal identifiers, as well as the curation and anonymization of valuable data sets. Some of these needs can be covered by offering training to researchers, perhaps in collaboration with relevant national research infrastructures.
- Participating in the development of **relevant research data infrastructures**, i.e. digital infrastructures promoting data sharing and consumption, in collaboration with organizations that collect and keep data. Datasets of particular interest among SSH researchers are public collections (e.g. those of the National Library of Finland, archives, museums), registers and statistics produced by the public sector (e.g. Statistics Finland, government research institutes, cities) and media and social media material. Workable data infrastructure includes the technology (e.g. centralized access), processes, organization and social networks related to data usage.
- Strengthening the **UH research environment for data-intensive SSH research**. This includes the arrangement of up-to-date equipment, tools and services, as well as maintenance of the technical and methodological expertise needed for the extensive use and development of the infrastructure for specified research purposes. This can mean, for example, permanent "staff scientist" positions, intensive collaboration with data scientists in e.g. the Faculty of Science, and a training programme for interdisciplinary mixed methodologies. Based on the survey results, data-intensive SSH research fields include the following:

    * **Digital humanities**, including e.g. computational linguistics
    * **Social data science**, including e.g. survey- and register-based research and computational social sciences
    * **Audio-visual research**, including e.g. linguistic, anthropological, ethnographic and interaction research based on video, audio, image or multimodal data
    * **Experimental research**, including e.g. social, behavioural, cognitive and educational research based on laboratory or field experiments

APPENDICES

APPENDIX 1: THE QUESTIONNAIRE

The purpose of this survey is to establish an overview of the current status and development needs of research infrastructures for social sciences and humanities at the University of Helsinki. By completing this survey, you can help us plan and implement the research infrastructure mission of the forthcoming Helsinki Institute for Social Sciences and Humanities (SSH). We kindly request that you complete the survey by 7 October 2019.

All responses will be kept confidential. The survey results will be made available to the University community in the form of statistics, summaries and figures from which individual respondents cannot be identified.

We aim to make the anonymised survey data available to the University community – and possibly the wider academic community – at a later date, following the principles for the responsible conduct of research as well as the Data Management Guidelines of the Finnish Social Science Data Archive.

This survey is conducted by the planning group preparing the launch of SSH operations. We are happy to answer any questions you may have. Further information on the survey can be obtained from Pekka Mäkelä, research coordinator (pekka.a.makela@helsinki.fi, phone 02941 29271) and Katri Huutoniemi, senior advisor in research administration (katri.huutoniemi@helsinki.fi, phone 02941 22552).

The survey involves processing of personal data, based on participants' consent. The GDPR privacy notice is attached.

Section A: Background information

A1. Name of respondent

A2. Faculty / unit

- [ ] Faculty of Arts
- [ ] Faculty of Social Sciences
- [ ] Faculty of Educational Sciences
- [ ] Faculty of Law
- [ ] Faculty of Theology
- [ ] Swedish School of Social Science
- [ ] Helsinki Collegium for Advanced Studies
- [ ] Other, please specify

Other, please specify

A3. Which discipline(s) does your research primarily represent?

- [ ] HUMANITIES

611 Philosophy ☐

6121 Languages ☐

6122 Literature studies ☐

6131 Theatre, dance, music, other performing arts ☐

6132 Visual arts and design ☐

614 Theology ☐

615 History and Archaeology ☐

6160 Other humanities ☐

6161 Phonetics ☐

6162 Cognitive science ☐

6163 Logopedics ☐

6164 Speech communication ☐

SOCIAL SCIENCES ☐

511 Economics ☐

512 Business and Management ☐

513 Law ☐

5141 Sociology ☐

5142 Social policy ☐

5143 Social anthropology ☐

5144 Social psychology ☐

5145 Social work ☐

515 Psychology ☐

516 Educational sciences ☐

517 Political science ☐

518 Media and communications ☐

519 Social and economic geography ☐

5200 Other social sciences ☐

5201 Political History ☐

5202 Economic and Social History ☐

- [ ] 5203 Development Studies
- [ ] NATURAL SCIENCES
- [ ] 111 Mathematics
- [ ] 112 Statistics and probability
- [ ] 113 Computer and information sciences
- [ ] 114 Physical sciences
- [ ] 115 Astronomy, Space science
- [ ] 116 Chemical sciences
- [ ] 1171 Geosciences
- [ ] 1172 Environmental sciences
- [ ] 1181 Ecology, evolutionary biology
- [ ] 1182 Biochemistry, cell and molecular biology
- [ ] 1183 Plant biology, microbiology, virology
- [ ] 1184 Genetics, developmental biology, physiology
- [ ] 119 Other natural sciences
- [ ] ENGINEERING AND TECHNOLOGY
- [ ] 211 Architecture
- [ ] 212 Civil and Construction engineering
- [ ] 213 Electronic, automation and communications
- [ ] 214 Mechanical engineering
- [ ] 215 Chemical engineering
- [ ] 216 Materials engineering
- [ ] 217 Medical engineering
- [ ] 218 Environmental engineering
- [ ] 219 Environmental biotechnology
- [ ] 220 Industrial biotechnology
- [ ] 221 Nano-technology
- [ ] 222 Other engineering and technologies
- [ ] MEDICINE AND HEALTH SCIENCES

3111 Biomedicine ☐

3112 Neurosciences ☐

3121 Internal medicine ☐

3122 Cancers ☐

3123 Gynaecology and paediatrics ☐

3124 Neurology and psychiatry ☐

3125 Otorhinolaryngology, ophthalmology ☐

3126 Surgery, anesthesiology, intensive care, radiology ☐

313 Dentistry ☐

3141 Health care science ☐

3142 Public health care science, environmental and occupational health ☐

3143 Nutrition ☐

315 Sport and fitness sciences ☐

316 Nursing ☐

317 Pharmacy ☐

318 Medical biotechnology ☐

319 Forensic science and other medical sciences ☐

AGRICULTURE AND FORESTRY ☐

4111 Agronomy ☐

4112 Forestry ☐

412 Animal science, dairy science ☐

413 Veterinary science ☐

414 Agricultural biotechnology ☐

415 Other agricultural sciences ☐

416 Food Science ☐

A4. Position of respondent

- ☐ Leadership of a faculty or unit (dean, vice-dean, department director)
- ☐ Coordinator of a discipline, research centre or other sub-unit
- ☐ (Other) researcher/teacher
- ☐ Other, please specify

Other, please specify

A5. From whose perspective / on whose behalf are you responding to this survey?

If you are responding on behalf of a larger group, please decide this within the group.

- ☐ Individual researcher/teacher
- ☐ Department, discipline or other administrative unit; please specify
- ☐ (Other) researcher/teacher
- ☐ Research group; please specify

Section B: Characterisation of your research

B1. What kind of research material or data do you primarily use in your research?

Feel free to add comments in the open field.

- ☐ No actual empirical material

Comment

- ☐ Literature as empirical material

Comment

Linguistic corpora etc.

Comment

Images, objects, artefacts, etc.

Comment

Writings or other products of research subjects

Comment

Documents, archives

Comment

'Big data'

Comment

Experimental data

Comment

Survey data

Comment

Interview data

Comment

Observation data

Comment

Register data, register-based statistics ☐

Comment

Media material, social media data ☐

Comment

Other type of data, please specify ☐

Comment

B2. Who primarily generates or collects your research data?

	The researcher	The research group	The researcher or research group in collaboration with others	They are generated or collected outside the research group	Other answer
No actual empirical material	☐	☐	☐	☐	☐
Literature as empirical material	☐	☐	☐	☐	☐
Linguistic corpora etc.	☐	☐	☐	☐	☐
Images, objects, artefacts, etc.	☐	☐	☐	☐	☐
Writings or other products of research subjects	☐	☐	☐	☐	☐
Documents, archives	☐	☐	☐	☐	☐
'Big data'	☐	☐	☐	☐	☐
Experimental data	☐	☐	☐	☐	☐
Survey data	☐	☐	☐	☐	☐
Interview data	☐	☐	☐	☐	☐
Observation data	☐	☐	☐	☐	☐
Register data, register-based statistics	☐	☐	☐	☐	☐
Media material, social media data	☐	☐	☐	☐	☐
Other type of data, please specify	☐	☐	☐	☐	☐

B3. **If you acquire research data (or material) from outside your research group or generate data in cooperation with others, from where or with whom?**

Feel free to add comments in the open field.

Other units in University of Helsinki ☐

Comment

Other Finnish universities or research institutes ☐

Comment

Foreign universities or research institutes ☐

Comment

Research infrastructures (e.g. CLARIN, ESS) ☐

Comment

Memory organisations ☐

Comment

Public authorities ☐

Comment

Open data repositories ☐

Comment

Companies ☐

Comment

Others, please specify ☐

Comment

B4. What kind of methods do you primarily use in analysing your research data?

	Statistical methods	Computational methods	Other quantitative methods	Qualitative methods	Mixed methods	Other methods
No actual empirical material	☐	☐	☐	☐	☐	☐
Literature as empirical material	☐	☐	☐	☐	☐	☐
Linguistic corpora etc.	☐	☐	☐	☐	☐	☐
Images, objects, artefacts, etc.	☐	☐	☐	☐	☐	☐
Writings or other products of research subjects	☐	☐	☐	☐	☐	☐
Documents, archives	☐	☐	☐	☐	☐	☐
'Big data'	☐	☐	☐	☐	☐	☐
Experimental data	☐	☐	☐	☐	☐	☐
Survey data	☐	☐	☐	☐	☐	☐
Interview data	☐	☐	☐	☐	☐	☐
Observation data	☐	☐	☐	☐	☐	☐
Register data, register-based statistics	☐	☐	☐	☐	☐	☐
Media material, social media data	☐	☐	☐	☐	☐	☐
Other type of data, please specify	☐	☐	☐	☐	☐	☐

B5. Can your research or research interests be considered as part of one or more of the following "research infrastructure-intensive" areas?

Feel free to add comments in the open field.

Computational linguistics / Language technology ☐

Comment

(Other) digital humanities ☐

Comment

Archaeology ☐

Comment

Experimental social and behavioral sciences ☐

Comment

Register-based social sciences ☐

Comment

Regular large-scale social surveys ☐

Comment

Computational social sciences / Social data science ☐

Comment

Other; please specify ☐

Comment

Section C: Research equipment

C1. What kinds of research equipment do you need to produce, acquire or process research data?

	Currently use	Currently need, but there are problems accessing it	Would need in the future	Do not need / Cannot say
Laboratory equipment or other "heavy" infrastructure	☐	☐	☐	☐
Mobile research equipment	☐	☐	☐	☐
Special facilities that enable, for example, experimental research designs, measurements, or video interviews	☐	☐	☐	☐
Software designed for research purposes	☐	☐	☐	☐
Large computational capacity	☐	☐	☐	☐
Large data storage capacity	☐	☐	☐	☐
Registers of respondents, subjects, organisations, etc.	☐	☐	☐	☐
Other equipment	☐	☐	☐	☐

C2. Please tell more about the equipment you use or need. If you currently use the equipment in question, please also specify where it is located or who owns it (if you know that).

Laboratory equipment and other "heavy infrastructure"	
Mobile research equipment	
Special facilities that enable, for example, experimental research designs, measurements and video interviews	
Software designed for research purposes	
Large computational capacity	
Large data storage capacity	
Registers of respondents, subjects, organisations, etc.	
Other equipment	

C3. What services do you need to acquire or use research equipment?

	Currently use	Currently need, but there are problems getting it	Would need in the future	Do not need / Cannot say
Services related to the acquisition, maintenance, upkeep and storage of research equipment (incl. facilities)				
Guidance and support for the use of research equipment				
Coordination of procurement and shared used of research equipment				
Surveying of equipment available from partners or through them				
Other services				

C4. Please tell more about the services you use or need. If you currently use the services in question, please also mention the service provider (if you know that).

Services related to the acquisition, maintenance, upkeep and storage of research equipment (incl. facilities)	
Guidance and support for the use of research equipment	
Coordination of procurement and shared use of research equipment	
Surveying of equipment available from partners and/or through them	
Other services	

Section D: Data infrastructure

D1. What services do you need to acquire or use data?

	Currently use	Currently need, but there are problems getting it	Would need in the future	Do not need / Cannot say
IT services for data sharing or remote use				

	Currently use	Currently need, but there are problems getting it	Would need in the future	Do not need / Cannot say
Surveying of data available from partners or through them	☐	☐	☐	☐
Surveying of openly available data	☐	☐	☐	☐
Resolution of issues related to data ownership and user rights (incl. licences, agreements)	☐	☐	☐	☐
Resolution of issues of research ethics and data protection related to the collection and use of data	☐	☐	☐	☐
Support for organising, cleaning, classifying, describing, etc. data	☐	☐	☐	☐
Support for storing, preserving and administering data	☐	☐	☐	☐
Support for making data openly available (incl. anonymisation)	☐	☐	☐	☐
Other services	☐	☐	☐	☐

D2. Please tell more about the services you use or need. If you currently use the services in question, please also mention the service provider (if you know that).

IT services for data sharing or remote use	
Surveying of data available from partners and/or through them	
Surveying of openly available data	
Resolution of issues related to data ownership and user rights (incl. licences, agreements)	
Resolution of issues of research ethics and data protection related to the collection and use of data	
Support for organising, cleaning, classifying, describing, etc. data	
Support for storing, preserving and administering data	
Support for making data openly available (incl. anonymisation)	
Other; please specify	

D3. Should some of the service needs you mentioned above be addressed by providing training for researchers? If so, what type of training do you think should be offered?

IT services for data sharing or remote use ☐

Comment

Surveying of data available from partners or through them ☐

Comment

Surveying of openly available data ☐

Comment

Resolution of issues related to data ownership and user rights (incl. licences, agreements) ☐

Comment

Resolution of issues of research ethics and data protection related to the collection and use of data ☐

Comment

Support for organising, cleaning, classifying, describing, etc. data ☐

Comment

Support for storing, preserving and administering data ☐

Comment

Support for making data openly available (incl. anonymisation) ☐

Comment

Other services ☐

Comment

D4. What (other) deficiencies or development needs related to the research infrastructures for the social sciences and humanities could be addressed by the future Helsinki Institute for Social Sciences and Humanities, and the research infrastructure activities pursued under its auspices? You may freely comment!

Section E: Attitude toward data sharing

E1. If you produce research data (as a researcher or in your research group), what is your view on sharing data with others?

Feel free to add comments in the open field.

We have shared data with others or have agreed on the shared use of data ☐

Comment

We have not yet shared data, but are, in principle, willing to do so ☐

Comment

We are not willing to share data with others ☐

Comment

Other view; please specify ☐

Comment

E2. If you have not shared research data with others, why not?

Feel free to add comments in the open field.

Data sharing involves legal or ethical problems ☐

Comment

The reuse of data involves risks; please specify ☐

Comment

We have not sufficiently familiarised ourselves with the legal, ethical, practical and other conditions of sharing data ☐

Comment

Transforming data into a format that others can use is not worthwhile ☐

Comment

It requires additional work for which we do not have sufficient resources ☐

Comment

We wish to keep our data for ourselves ☐

Comment

Other reason; please specify ☐

Comment

E3. What problems or issues do you think would arise if an agreement provided the University with rights of ownership, manage, or use to the research data you have collected?

Section F: Future visions

F1. What future visions are there in your field for the production, acquisition, use or analysis of research data?

F2. How could the future Helsinki Institute for Social Sciences and Humanities and the research infrastructure activities pursued under its auspices promote the above visions?

F3. **What international or national examples of good research infrastructures would you like to highlight in your research field?**

F4. **Who would you nominate for a working group developing infrastructures in your research field?**

Cordial thanks for your participation. The results of this survey will be made available to you in Flamma and in the PI email-lists.

APPENDIX 2: RESPONDENTS' DEPARTMENT ETC., FREE-FORM ANSWERS

From whose perspective / on whose behalf are you responding to this survey? Free-form answers presented according to the respondents' faculty

Faculty Department, discipline or other administrative unit / Research group / Other
Faculty of Arts
Medieval Publishing (ERC project), Authorial Publishing (Academy project)
Helsinki Computational History Group
Papyrus research group
African Studies (discipline)
Helsinki Centre for Digital Humanities HELDIG, Semantic Computing
Ancient Near Eastern Empires (Centre of Excellence), Ancient Near Eastern (discipline)
Research on spoken interaction
Plain language research
Onomastics
FIN-CLARIN
Department of Digital Humanities; Phonetics, Cognitive science
European area and cultural studies (discipline)
European area and cultural studies (discipline)
Human Sciences – Computing Interaction
The Helsinki Term Bank for the Arts and Sciences
Traffic Research Unit TRU, operating in Cognitive Science, Department of Digital Humanities
Faculty of Social Sciences
Centre of Excellence in Law, Identity and the European Narratives (Centre of Excellence)
Population Research Unit
Criminology (discipline); the use of register data in particular
University of Helsinki Centre for Research on Addiction, Control and Governance CEACG
Environmental Policy Research Group EPRG
Everyday thinking and arguing

Centre for Consumer Society Research CCSR

Faculty of Educational Sciences

Cicero Learning

Didactics of biology and biology education (research team), part of Maker@STEAM community

APPENDIX 3: SOURCES OF ACQUIRED RESEARCH MATERIAL, FREE-FORM ANSWERS

If you acquire research data (or material) from outside your research group or generate data in cooperation with others, from where or with whom? A summary of free-form answers.

Other units in University of Helsinki (n=20)

Helsinki University Library (n=3)

National Library

Faculty of Social Sciences, incl. Sociology

Faculty of Arts, incl. Digital Humanities and Finnish language

Faculty of Medicine, incl. Department of Psychology and Logopedics (n=3), BioMag Laboratory, Cognitive Brain Research Unit, Department of Public Health

Faculty of Educational Sciences, incl. Teacher Education;
research projects MathTrack, CELLS

Faculty of Biological and Environmental Sciences

Faculty of Science, incl. Geosciences, Physics, Computer Science

Faculty of Biosciences and Environment, incl. Biosciences, Fisheries and Environmental Management Group

HiLIFE, incl. Institute for Molecular Medicine Finland (FIMM)

Other comments:

- video-recorded interactions
- other researchers and their output and archives
- I collaborate with several groups, mostly in the Medical Faculty, to analyse their experimental data

Other Finnish universities or research institutes (n=40)

Aalto University (n=5)

University of Jyväskylä (n=4) (incl. Faculty of Sport and Health Sciences)

University of Oulu (n=3)

University of Turku (n=3)

University of Tampere (n=2) (incl. Faculty of Education and Culture)

Finnish Institute for Health and Welfare (THL) (n=3)

Institute for the Languages of Finland (Kotus) (n=3)

Statistics Finland (n=2)

Language Bank of Finland (n=2)

University of Lapland (Faculty of Social Sciences)

Uniarts Helsinki

South-Eastern Finland University of Applied Sciences (XAMK)

VTT Technical Research Centre of Finland

Social Insurance Institution of Finland (KELA)

Geological Survey of Finland (GTK)

Research projects, incl. ARTSEQUAL (multidisciplinary research project), LUMA SUOMI, EU-projects

Sami Research Centre

Other comments:

- other research groups
- other researchers and their output and archives
- project partners
- Comparative linguistics draws on all existing results in the field of lexicography, regardless of where they are produced
- The interviews were collected through crowdsourcing: in addition to Helsinki, a data collection course was organized at four other universities

Foreign universities or research institutes (n=42)

European universities

- University of Cambridge (n=2)
- Edinburgh (n=2)
- Oxford
- St. Andrews
- Liverpool
- Leeds
- Stockholm University
- Uppsala University
- Umeå
- Karolinska Institutet, Stockholm
- Bremen
- Ghent University
- University of Bari
- Universitat Pompeu Fabra Spain

- European University

Universities in the USA and Canada

- University of California, Berkeley (n=3)
- University of Pennsylvania
- Princeton University
- Virginia Tech University
- Montreal
- Toronto

Universities in Australia and New Zealand

- Western Sydney
- University of Melbourne
- University of Auckland

Chinese universities

- Educational University of Hong Kong
- Beijing Normal University

Russian universities

- Higher School of Economics
- Nizhny Novgorod

Japanese universities and research institutes

Baltic universities

Research institutes etc.

- Russian Academy of Sciences (n=2)
- Herzog August Bibliothek
- Eesti Keele Instituut
- NorQuest College Edmonton Alberta

Research projects

- iPAL
- Nordic Centre of Excellence: QUINT

Other parties and other comments:

- Mainly French databases
- E.g. Joint European corpus: Germany, Switzerland, Poland, USA

The data collection has always been done in cooperation with individual international colleagues, not larger organizations

I have a partnership with e.g. Japanese researchers, and my publications deal with their corpus

In practice, the production and collection of research material takes place both within the University of Helsinki and in collaboration with or with the use of material produced by foreign universities

Universities and research groups producing extensive electronic corpuses made from primary sources in the ancient Middle East. Most importantly UC Berkeley

In publication collaboration I sometimes also use material collected by others

Research infrastructures (e.g. CLARIN, ESS) (n=16)

CLARIN (n=10)

Language Bank of Finland (n=5)

European Social Survey (n=2)

DARIAH

Eurostat

World Value Survey

Thesaurus Linguae Graecae

METANET

CSC

Memory organizations (n=26)

Finnish memory organizations

The Finnish Literature Society (SKS) (n=7)

The National Archives of Finland (n=5)

National Library (n=4)

Svenska litteratursällskapet SLS (n=3)

The Finnish Heritage Agency (Museovirasto) (n=3)

Turun museokeskus

Foreign/international memory organizations

Herzog August Bibliothek (library)

The Labour Archives (Työväen arkisto) (n=2)

The People's Archives (Kansan arkisto) (n=2)

Fiskars Museum archive

Brages Pressarkiv (Finnish newspaper archive)

The Consortium of European Research Libraries (CERL)

British Library

National Library of Swede (Kungliga Biblioteket)

archive.org

arXiv

Other comments

Private archives

Public archives

Nordic and Baltic archives and libraries

County museums

All archives that keep archaeological material

Other museums and also congregations

Archaeological material: national and international museums

Museums, national libraries

Public authorities (n=31)

Ministry of Education and Culture (n=2)

Finnish National Agency for Education (EDUFI) (n=3)

Digital and Population Data Services Agency (VRK) (n=3)

Metsähallitus

Social Insurance Institution of Finland (KELA) (n=2)

Finnish Institute for Health and Welfare (THL)

The Bank of Finland

Statistics Finland (n=3)

Matriculation Examination Board (YTL) (n=2)

Institute for the Languages of Finland (Kotus) (n=2)

VATT Institute for Economic Research

Finnish Education Evaluation Centre (Karvi)

National Audiovisual Institute (KAVI)

Finnish Film Foundation (SES)

Finnish Innovation Fund Sitra

Library of Parliament

Regional State Administrative Agency of Southern Finland (AVI)

HUS Helsinki University Hospital

Arbis (Adult education centre)

City of Helsinki

Federal State Statistics Service Rosstat (Russia)

Other comments

- Hospitals, palliative care homes
- Statistics offices, geographical surveys
- Medical centres and other social and healthcare units
- Ministries
- Health care districts

Open data repositories (n=19)

Finnish data repositories

- Statistics Finland (n=4)
- Finnish Social Science Data Archive (FSD) (n=3)
- Hoitoilmoitusjärjestelmä Hilmo (THL)
- Digital and Population Data Services Agency (VRK)
- Archive of Parliament

Foreign/international data repositories

- Data Archive for the Social Sciences (DAS)
- www.wittgensteinsource.org
- Bayerische Staatsbibliothek
- ORACC Open Richly Annotated Cuneiform Corpus
- GitHub
- Archivie.org
- flickr
- GAS Pravosudie

Other comments

- Big data on game players from open online repositories - the specific repositories change from time to time
- International data archives

German libraries' digitized data

Archives and corpora open to researchers, mainly in Japan

Platforms for language analysis

Companies (n=12)

YLE (n=2)

Taloustutkimus Oy

The Research Institute of the Finnish Economy (ETLA)

Council for Aid to Education CAE (USA)

Microsoft Research

Facebook Research

Google Research

Other comments

Publishers (n=2)

Companies' own data

Private hospitals and palliative care homes

Social media dataset vendors

Media companies

Video-recorded interactions

Companies sometimes help us with experiment arrangements

Others, please specify (n=28)

Education Division of City of Helsinki

BIOS research unit

Sami Siida

Finnish Centre for Easy Language (Selkokeskus)

For example, the data from the Universal Dependencies project is extremely important, and the value of such larger open source projects will surely increase in the future.

Financial market information providers such as Bloomberg, Reuters, Compustat, Macrobond

Jointly collected international and quality managed (e.g. ESS and PISA)

academia.edu

archive.org

Deepmind

Cities, municipalities providing education

City/municipality

Libraries

Churches, religious communities, various organizations

Third sector organizations

Research subjects

Schools

Relevant stakeholders for each research target, e.g. third sector organizations

schools - when field tests are conducted in the field, the field is of course actively involved in the production of the material

citizens (by methods of citizen science)

A co-operation project between a Finnish NGO and a foreign university

I sub-contract outside researchers to collect interview data

My “partners” are not institutions but other individual researchers from the same or other disciplines in Finland and abroad

Individual researchers, organizations and other informants from around the world

Researchers from all over the world

Part of the works of art, in particular, are in private collections, whereby access to the materials depends on the cooperation of their holders

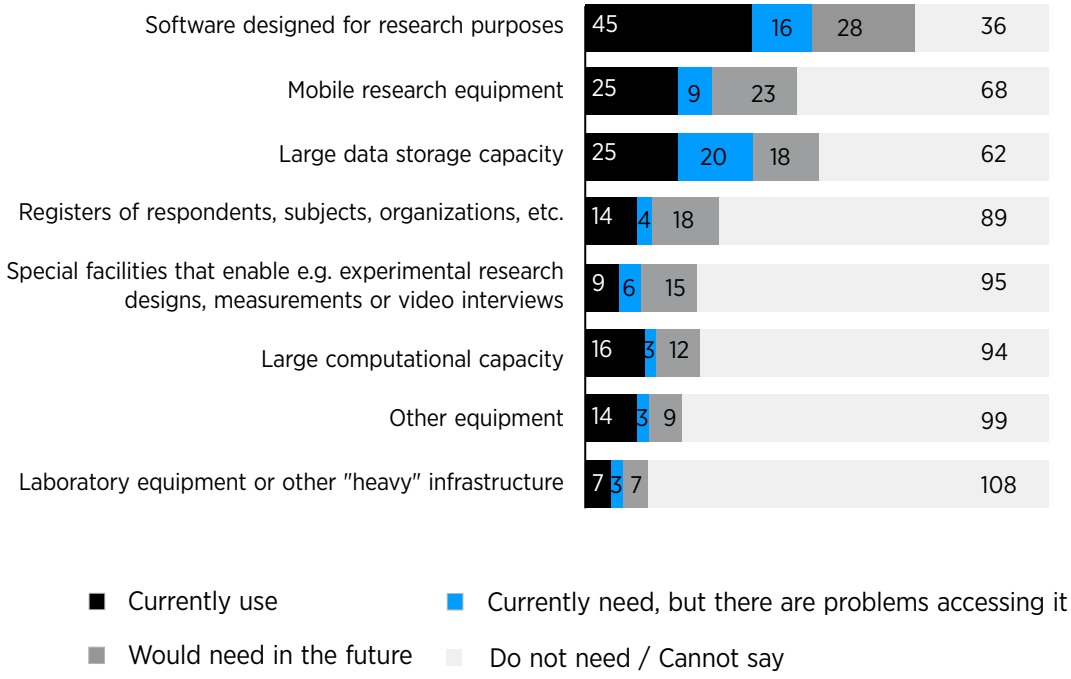
International partners

Local research assistants

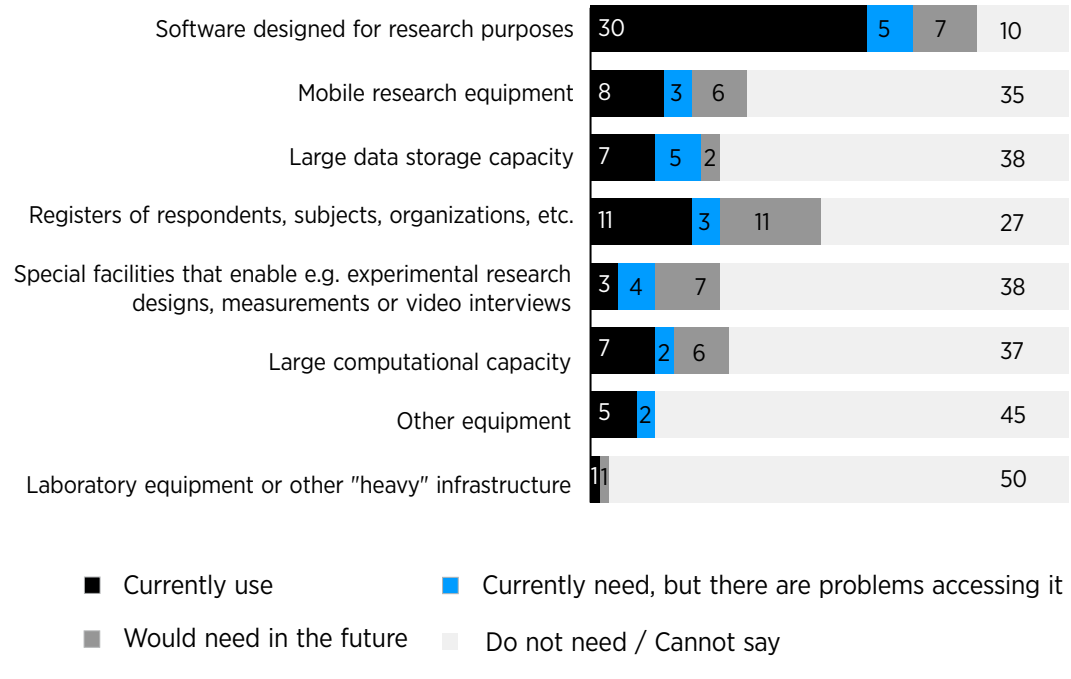
APPENDIX 4: RESEARCH EQUIPMENT, FACULTY-LEVEL RESULTS

What kinds of research equipment do you need to produce, acquire or process research data? Response distribution in the Faculties of Arts, Social Sciences and Educational Sciences.

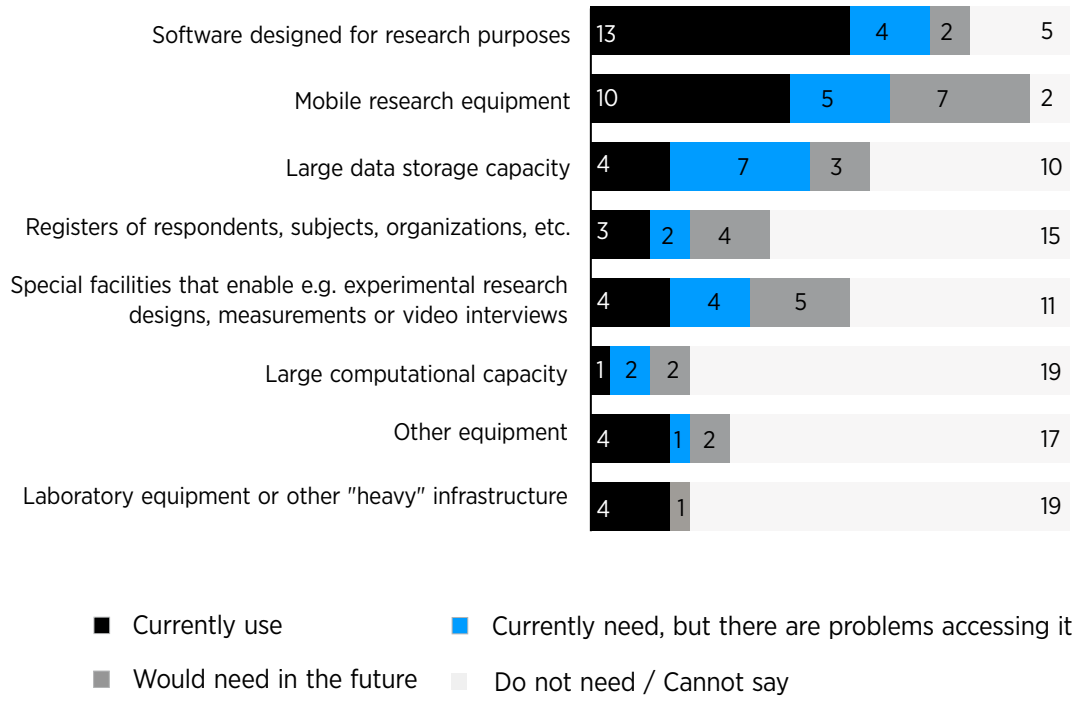
Faculty of Arts (N=125)



Faculty of Social Sciences (N=52)



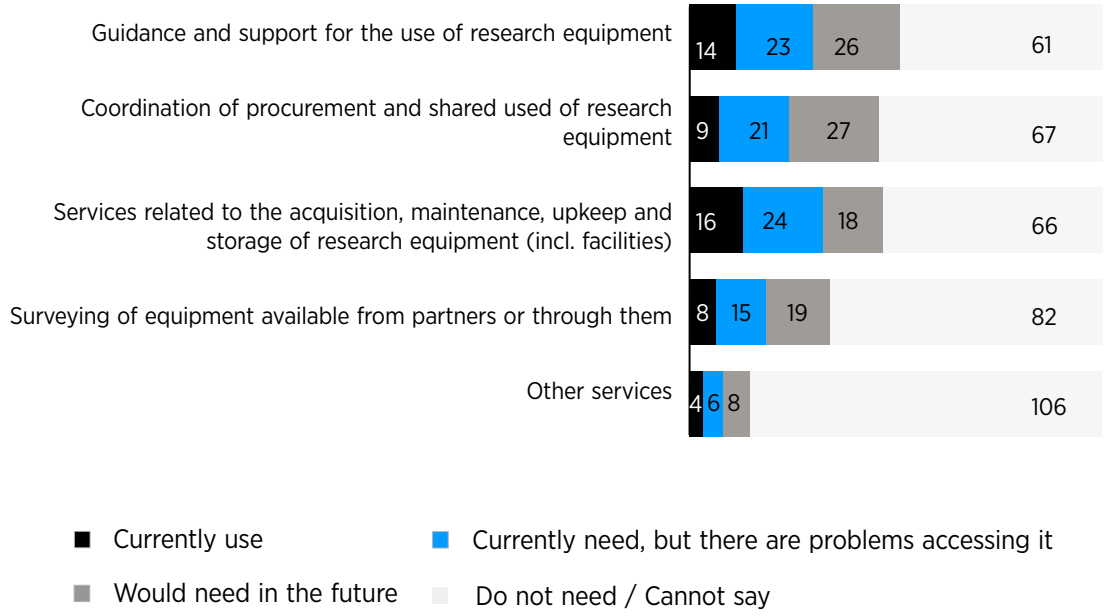
Faculty of Educational Sciences (N=24)



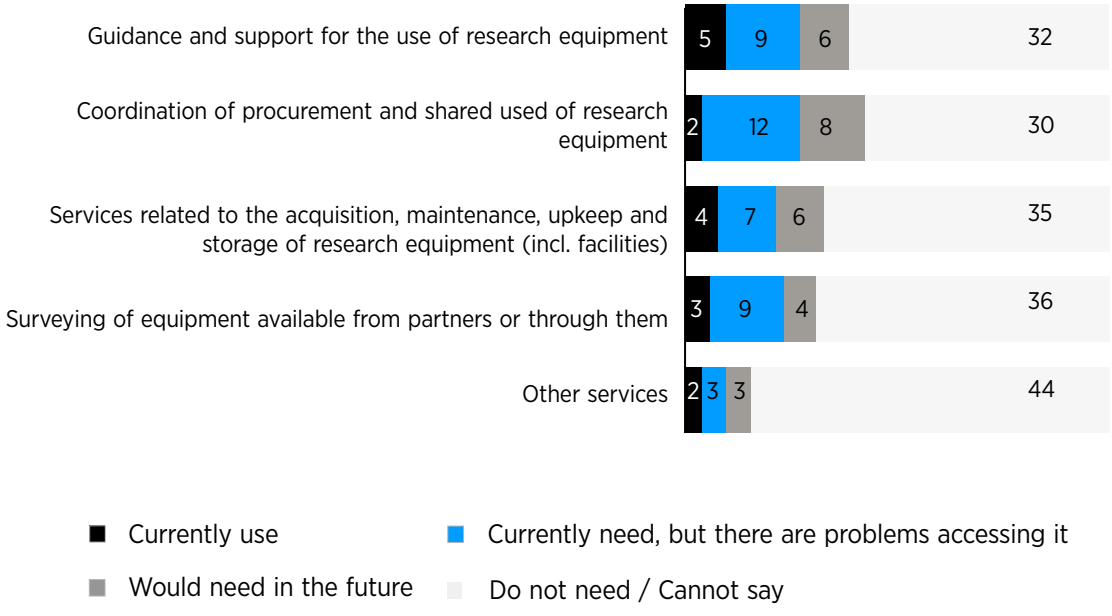
APPENDIX 5: SERVICES FOR ACQUIRING AND USING RESEARCH EQUIPMENT, FACULTY-LEVEL RESULTS

What services do you need to acquire or use for research equipment? Response distribution in the Faculties of Arts, Social Sciences and Educational Sciences.

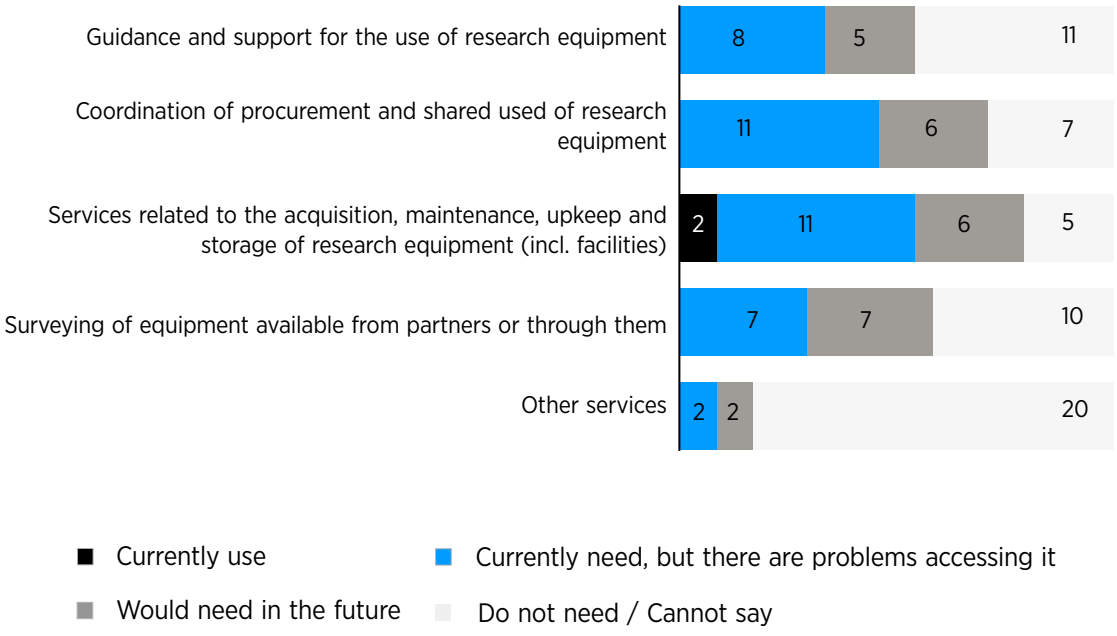
Faculty of Arts (N=124)



Faculty of Social Sciences (N=52)



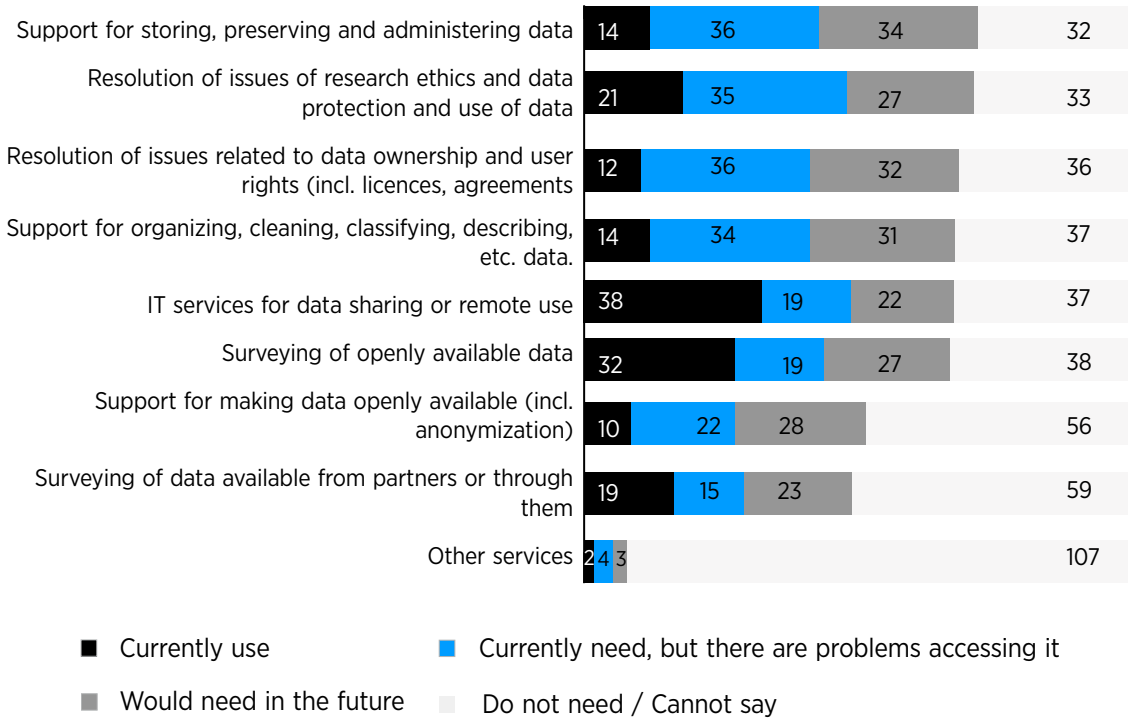
Faculty of Educational Sciences (N=24)



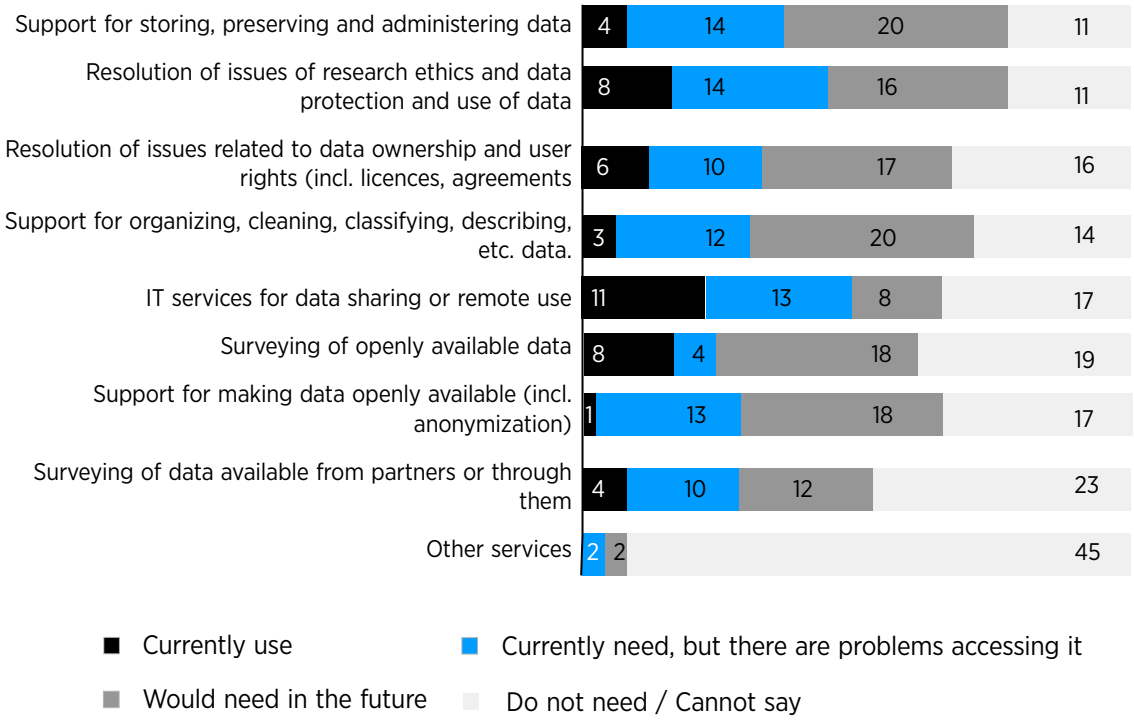
APPENDIX 6: SERVICES FOR ACQUIRING AND USING RESEARCH DATA, FACULTY-LEVEL RESULTS

What services do you need to acquire or use for research data? Response distribution in the Faculties of Arts, Social Sciences and Educational Sciences.

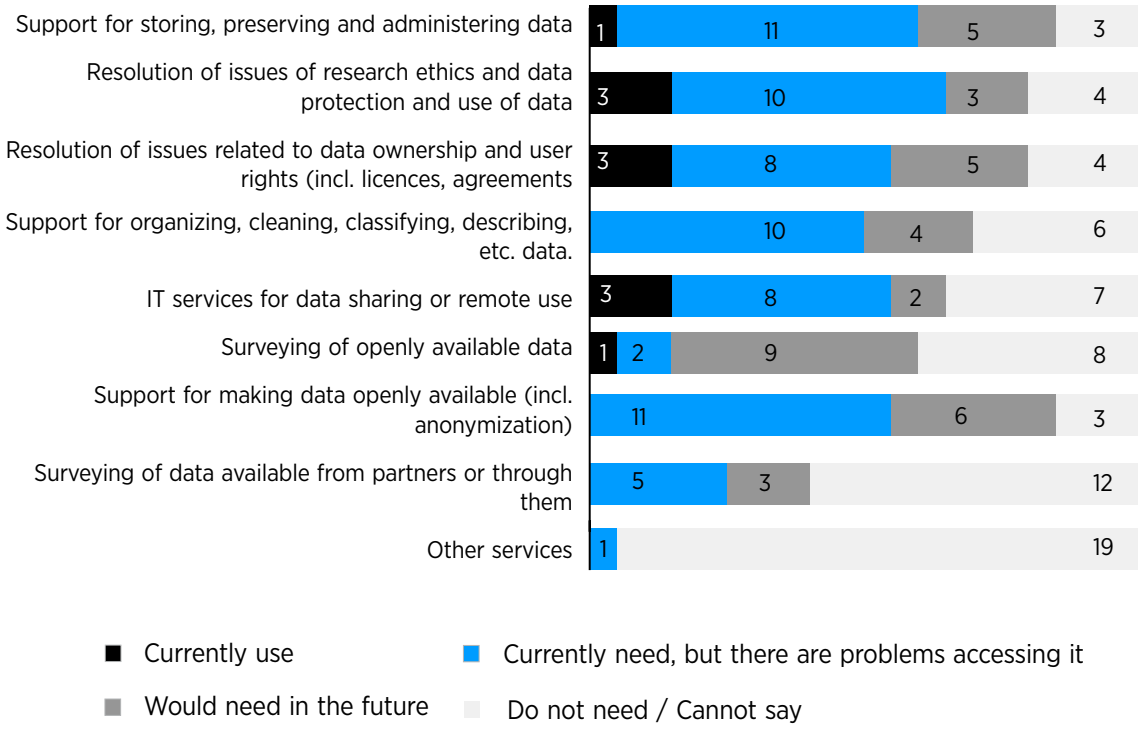
Faculty of Arts (N=116)



Faculty of Social Sciences (N=49)

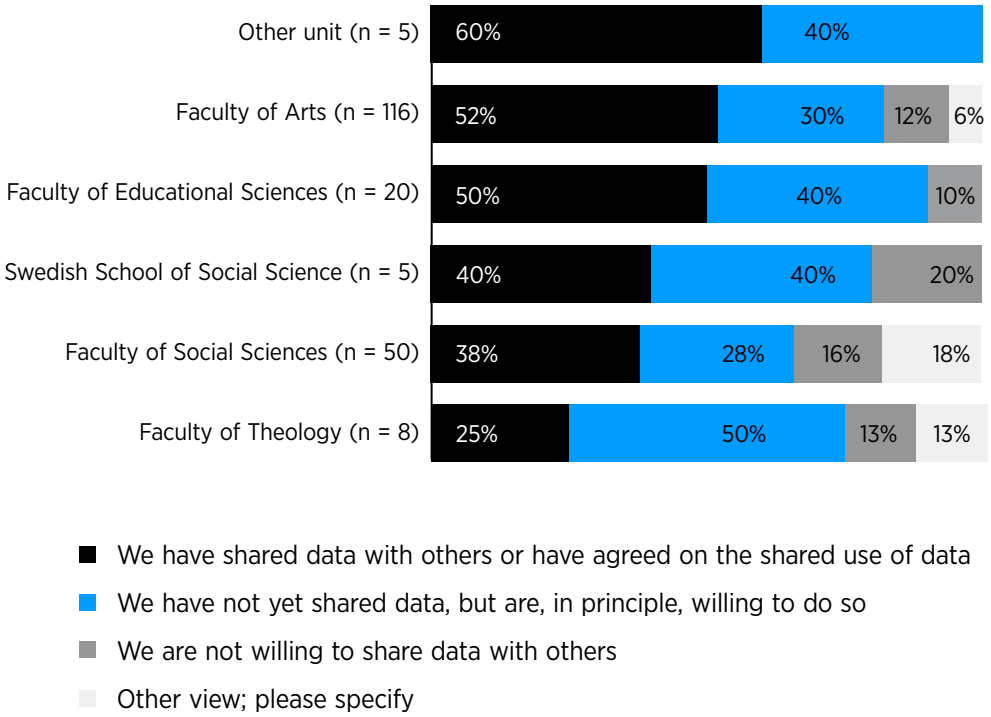


Faculty of Educational Sciences (N=20)

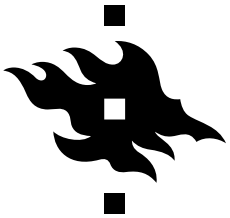


APPENDIX 7: ATTITUDE TOWARDS DATA SHARING, FACULTY-LEVEL RESULTS

If you produce research data (as a researcher or in your research group), what is your view on sharing data with others? Response distribution within faculties and units with at least five respondents.



HELSINKI INSTITUTE FOR
SOCIAL SCIENCES
AND HUMANITIES



UNIVERSITY OF HELSINKI