Finnish Social Science Data Archive

Arja Kuula & Sami Borg

# Open Access to and Reuse of Research Data – The State of the Art in Finland

Finnish Social Science Data Archive 7, 2008

Finnish Social Science Data Archive 7, 2008
Arja Kuula and Sami Borg (2008).
Open Access to and Reuse of Research Data – The State of the Art in Finland

Publisher: Finnish Social Science Data Archive (FSD)
University of Tampere
Åkerlundinkatu 2 A, 5th floor
Tampere, Finland
Adress: Finnish Social Science Data Archive – Yhteiskuntatieteellinen tietoarkisto
FIN-33014 University of Tampere
Finland
Tel: (03) 3551 8519
Fax: (03) 3551 8520
E-mail: fsd@uta.fi
WWW: http://www.fsd.uta.fi

Distribution: Bookshop TAJU
P.O. Box 617, FIN-33014 University of Tampere
Finland
Tel: (03) 3551 6055
Fax: (03) 3551 7685
taju@uta.fi
http://granum.uta.fi

ISSN 1459-8906
ISBN 978-951-44-7479-8

Graphic design
Vinjetti Ky

Layout
Marita Alanko

Tampere 2008

# Contents

# Open Access to and Reuse of Research Data – The State of the Art in Finland

Arja Kuula & Sami Borg

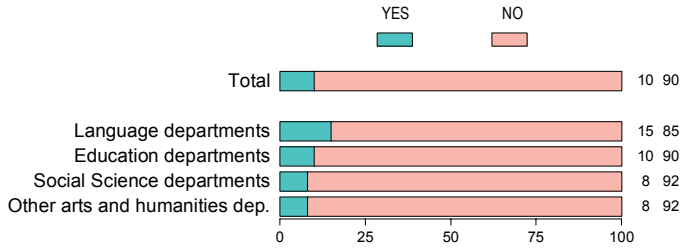
In 2004, Ministers of science and technology of the OECD countries adopted a Declaration on Access to Research Data from Public Funding. In this declaration, they recognised the importance of access to research data and invited the OECD to develop a set of guidelines based on commonly agreed principles to facilitate optimal cost-effective access to digital research data from public funding. This request was taken up by OECD's Committee for Scientific and Technological Policy, which launched a project by asking a group of experts to develop a set of principles and guidelines. Next, the developed principles and guidelines were submitted to an extensive consultation process, after which they were approved by the OECD's Committee for Scientific and Technological Policy in October 2006, attached to an OECD Recommendation, and endorsed by the OECD Council in December 2006.

In 2006, the Ministry of Education in Finland allocated resources to the Finnish Social Science Data Archive (FSD) to chart national and international practices related to open access to research data. Consequently, the FSD carried out an online survey targeting professors of human sciences, social sciences and behavioural sciences in Finnish universities. Some respondents were senior staff at research institutes. The respondents were asked about the state and use of data collected in their department/institute. Almost half of the respondents considered the preservation and use of digital research data to be relevant to their department. The number of respondents (150) is large enough to warrant statistical analysis even though response rate was low at 28%.

# 1 Preservation of research data – current situation

First, the survey charted how research data were used and preserved in Finnish universities. Professors were asked whether their department had any guidelines on the preservation of digital research data. A great majority (90%) said no (figure 1.1):

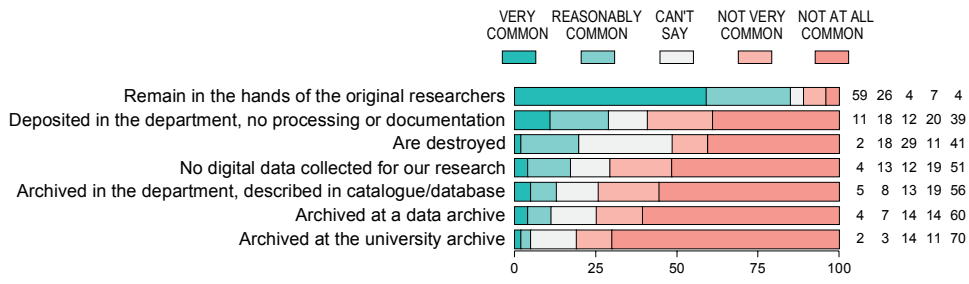
Figure 1.1. DOES YOUR DEPARTMENT HAVE ANY GUIDELINES OR REGULATIONS ON THE PRESERVATION OF DIGITAL RESEARCH DATA? (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

Next, the respondents were asked what happened to research data in their department/institute after the original research had been completed. Seven alternative scenarios were given, some of them very common, some much less so. Figure 1.2 presents the results:

Figure 1.2. HOW COMMON A PRACTICE IN YOUR DEPARTMENT/INSTITUTE AFTER THE ORIGINAL RESEARCH IS COMPLETED: THE DATA COLLECTED... (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

By far the most common practice was that original researchers retain their data themselves (85%). One in four (28%) responded that the data were stored in the department/institute though without any further processing or documentation while 12% said the data were archived in the department/institute and described in a catalogue/database as well. Archiving at a university archive or a data archive was not very common. The FSD's influence could be seen in social sciences, making archiving at a data archive a bit more frequent (16%) than in other sciences. Languages, the arts and humanities also had some national-level archiving solutions.

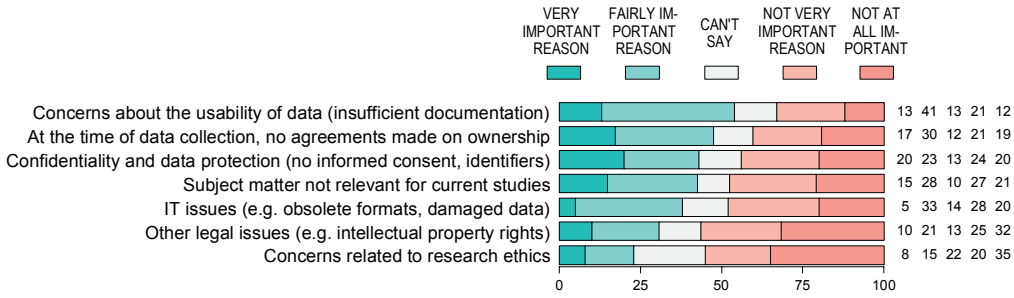
The results show that it is rare for Finnish universities to have an archiving and preservation policy regarding research data. The guidelines that exist are generally issued by departments, not the university. After completing their research projects, researchers generally store the data themselves but without any long-term preservation plan. Only a few store their data in the department/institute.

National Archives Service has not published any national guidelines on the preservation of research data. Any university-level guidelines that exist are typically included in the Records Management Schedule (RMS) which usually focuses on administrative documents. In fact, to comply with the Finnish legislation on contracts and archiving, university archives should archive not only administrative documents related to research projects but also the research data itself, if collected with public funding. However, in the present situation the easiest solution seems to be that the preservation of data is planned and carried out at the departmental level.

# 2 Barriers to and disadvantages of open access

The OECD guidelines take it for granted that digital research data will be reused. This may be true in countries where the infrastructure and culture of data reuse have already been established and the scientific community has accepted open access and reuse. This is not the case in Finland. Therefore the survey conducted by the FSD studied what kind of barriers there were to open access and what were the respondents' perceptions of the potential disadvantages of data reuse. The respondents were asked to estimate why the data collected in their field of research were not used, and how significant a barrier certain concerns were to open access.

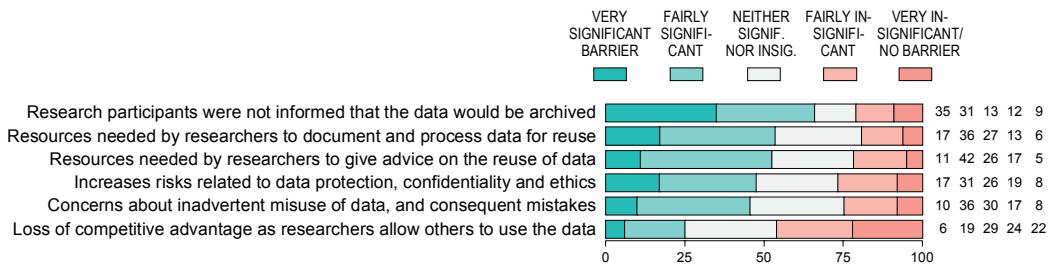
Figure 2.1. HOW MUCH DOES THE FOLLOWING EXPLAIN WHY THE DIGITAL DATA COLLECTED IN YOUR RESEARCH FIELD ARE NOT REUSED (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

The main finding was that there seemed to be several different barriers of roughly equal importance.

Figure 2.2. YOUR OPINION ON HOW SIGNIFICANT A BARRIER THE FOLLOWING IS TO ENHANCING OPEN ACCESS TO DATA (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

Next, the results of figure 2.1 and figure 2.2 will be discussed addressing each specific concern separately. Responses to open-ended questions will also be taken into account.

## 2.1 Concerns about inadvertent misuse of data, and consequent mistakes

Figure 2.2 shows that nearly every second (46%) respondent regarded the concern about inadvertent misuse of data as a very or fairly significant barrier to open access. This may imply that many researchers think only they themselves are capable of using their data correctly. In human sciences this concern is taken to be relevant for both quantitative and qualitative data. Some datasets may be regarded as particularly vulnerable to misinterpretation. One respondent observed:

> "I think open access would provide great opportunities to study new questions. On the other hand, it is equally easy to see the problems. I find it terrifying to think that the data my group has collected would be misinterpreted and used to justify a point of view which I myself would perceive not only as wrong but also as unethical."

The special nature of qualitative research, for example, can be used to justify the view that no-one else except the original researcher can understand the data. Natasha Mauthner et al. (1998) point out that qualitative data are not suitable to be archived because using archived data is incompatible with the interpretative and reflexive nature of the research paradigm. (mt. 743)

It is also true that an interviewer can perceive and partly interpret the emotions, expressions and exclamations of the interviewee. Social interaction may contain elements that are difficult to express verbally. However, principal researchers often employ field or research staff to collect and process the data. At the analysis stage, even those principal researchers who have personally collected the raw data mainly work with material adapted from it. Scientific conventions require that researchers are able to express and justify all interpretations based on a particular dataset even those made in authentic situations.

Nigel Fielding (2000) and Louise Corti (2006a) feel that reusing qualitative data is more of a practical issue than an epistemological one. To ensure that data are reusable for further research, there must be sufficient documentation on the context and on how the data were collected. The fact that so many researchers resist data reuse may be an indication that data are not documented well enough to allow reuse at present. Without detailed documentation, data reuse may result in inaccurate, if not downright erroneous, interpretations.

As seen in figure 2.1, two in five (43%) respondents said that an important reason for the non-use of previously collected data was that the subject matter was not relevant for current studies. Some professors commented that it is not worthwhile to process all digital research data for archiving because sometimes data are collected for very specific purposes.

On the other hand, Markku Leppänen (2006) has stated that there are no unequivocal criteria for deciding which datasets should be stored permanently. He does, however, list a few important criteria: usability and conditions of access, level of uniqueness, social, cultural and scientific value, and cost of preparation for archiving. One potential criterion is whether the data can be used for teaching research methods.

## 2.2 No agreements on ownership

When data are collected, usually no agreement is made on who owns the dataset. One in two (47%) respondents considered the lack of agreement regarding ownership as an important reason for which data were not reused. One third mentioned other legal issues (e.g. intellectual property rights) as a very or fairly important reason.

In addition to legislation, perceptions of data ownership are affected by academic practices and conventions. In fact, their influence may be even stronger than that of legislation. Research process requires creative thinking throughout, from drawing up a research design to making analyses based on the collected data. In empirical research, research design, data collection and data processing are crucial parts of a research project. Before a dataset is ready for analysis, a lot of work has been done at various stages of the process all of which have required a number of important decisions from the researcher(s). No wonder, therefore, that researchers find it so hard to accept that the data they have put so much effort and time in designing and collecting do not remain in their own hands and use.

The data are generally considered to belong to the original investigator or the research team, as part of their intellectual capital. Finnish researchers, for example, often take their data with them when changing jobs. The same perception of data as part of investigator's intellectual capital can be seen in the survey responses. Regardless of discipline, by far the most common practice in Finland seems to be that original researchers store their data themselves.

People often think that depositing their data to an archive would mean giving the copyright away. But that is not the case. The licence given by a researcher to a data archive to distribute and preserve his/her data does not transfer the moral rights under copyright. For example, the moral rights of a dataset archived at the FSD remains with the original researcher(s) even though the data are distributed by the archive and the archive controls access to it. Moral rights under copyright entail that the author of the dataset must be acknowledged in any publication based wholly or in part on the data. Thus the planning and agreements made on data processing, storing and reuse are in practice more important than copyright laws.

However, in cases where the object of analysis is a digital dataset protected by the copyright law, copyright issues may form a barrier to open access. This is particularly relevant for arts and culture research. The Finnish copyright law dictates that digitalised articles, advertisements and the like cannot be archived for research. Some respondents expressed a hope that this problem would be solved since the copyright law excludes some research questions even at present.

According to the Finnish contract law, research data remain the responsibility of and in possession of the body producing the research (usually a university department/institute) unless otherwise agreed. Quite recently, in the autumn of 2008, the Academy of Finland renewed its funding application guidelines by adding a new requirement concerning research data: "The Academy requires that applicants give an account of how the project's research material will be obtained, how it will be used and stored and how its later use will be made possible. The information management plan shall be presented in connection with the research plan. It is recommended that research projects funded by the Academy deliver the social science research data they have gathered to the Finnish Social Science Data Archive (FSD)." (Academy of Finland: Application guidelines to all calls, 2008)

Recommendation to archive social science data to the FSD has been in the Academy guidelines for years. Regardless of that, very few research teams have made any agreements on the copyright, ownership or archiving of data, either before or after the data collection. The contract practices of Finnish universities have not been much help so far either since the contracts have seldom contained any clauses on research data. Hopefully, the fact that a central research funder has now begun to require a data management plan will make researchers more favorable towards data archiving and sharing.

## 2.3 Competition for academic positions and funding

One respondent in four regarded the loss of competitive advantage as a significant barrier to open access to data. A similar proportion of respondents saw it as a fairly insignificant barrier. 22% of respondents thought it formed no barrier at all or only a very insignificant one.

Several publications have mentioned the loss of competitive advantage in the competition for academic credits and awards as a barrier to open access (see Clubb et al. 1985, 57 58; Sieber 1991, 142 143). Scholars are reluctant to release data in which they have invested time, money and energy to just any researcher. Particularly painful is the thought that other researchers may do better in the competition for funding and publications by using the data that the researcher him/herself has collected and later released to others.

The fear that re-users may inappropriately criticise the findings of the original study may form another barrier. The fear is probably emphasised if the researcher has been subjected to or witnessed unfair competition. The other side of the coin is that researchers may also fear that the weaknesses of the original study will be revealed when someone compares the findings with the data itself. This may increase resistance to open access.

It is an established academic practice that original investigators have the right to be the first users of their data. Originality of research has acknowledged value in science. In practice, the person who first publishes a new scientific finding will get the merit (Kiikeri and Ylikoski 2004, 127 129). Original researchers who have designed and collected the data have an indisputable right to publish the most relevant findings before releasing the data to others to use. This principle was mentioned by some respondents. However, releasing the data for reuse after the relevant findings have been published supports the basic principles of science which are objectivity, a critical attitude, autonomy, and progressivity (Niiniluoto 2002).

Objectivity and a critical attitude entail that there is a possibility to verify unclear findings, even though in practice it is almost impossible to replicate a research in an exactly similar manner and circumstances. Replicating is particularly difficult in qualitative research but not without difficulties in quantitative research either (Ray and Valeriano, 2003). Still, open access to data may improve the quality of research. The mere knowledge that findings may be checked against the data will force researchers to be systematic and thorough in their analyses.

Equally relevant is giving other researchers the possibility to ask new scientific questions and make different types of analyses with the same empirical data. This possibility forms part of the progressivity of science, and its importance has been emphasised with the increase of open access to data. For example, the qualitative interview data The Edwardians: Family Life and Work Experience Before 1918 (Thompson and Lummis 2000), collected by Paul Thompson

in the 1970s, has after its digitalisation been used for over 100 studies. The data are still being used for different types of research since it contains rich and unique information on the effects of industrialisation on family life (Corti 2006b). Thompson's data are a prime example of how open access supports the generation of new scientific information. If the data had remained solely in the hands of Thompson, a number of studies and publications might never have been carried out and published.

Some survey respondents commented on the progressivity and collective nature of science. Below some examples:

> "If a researcher collects information only for himself, he is like a piggy bank that never gets emptied. True and reliable information will increase when it is shared and assessed openly and critically."

> "I myself have only positive experiences of giving other [researchers] access to my own datasets, some of them large, even though I continue to use them myself. When you give something you usually get something back."

Making one's data available to others may also offer competitive advantage in terms of citations. Gleditsch, Metelis and Strand (2003) studied citations to all articles in the Journal of Peace Research for the period 1999–2001. They found that an article where the author has made the data available is on average cited twice as frequently as an article with no data but otherwise equivalent credentials. Ray and Valeriano (2003) write that 90% of articles in their field (international politics) are never cited. Therefore, they have a positive attitude towards the possibility that researchers get cited through offering access to their data. They point out that releasing data and having other scholars publish findings based on the data is more likely to help scholars than harm them (mt. 84). According to Sieber (1991), funding opportunities are sometimes improved if the grant proposal shows how the data collected would be useful to other scholars (p. 143).

## 2.4 Usability and IT problems

Many datasets, at present probably the majority, can no longer be used because they were not documented and processed for archiving and reuse from the beginning. This fact is reflected in the survey findings. The respondents regarded concerns about the usability of data (e.g. insufficient documentation) as an important reason why data were not reused in their field (54%) (figure 2.1). No matter how many datasets a department/institute may have they are of little use if not documented and processed properly.

The same problem was brought into focus when the respondents estimated how significant a barrier certain concerns were to enhancing open access. 54% said they regarded the resources needed by researchers to document and process data for reuse as a very or fairly significant barrier (figure 2.2). Almost as many thought that the resources needed by researchers to give advice on the reuse of their data was a significant barrier. The heavy workload most scholars have to cope with was reflected in the responses, as was the fear that open access would further add to that load. A comment on the issue:

> "The greatest problem is that researchers and/or research project leaders would have to invest a lot of time and energy to do this."

IT problems frequently prevent reuse as digital data tends to become obsolete very quickly. More than one in three respondents (38%) considered obsolete formats or damaged data as an important reason for which existing data were not used. Even when attitudes towards reuse are positive, the rapid development of formats and equipment may undo all good intentions. Hence, it is essential to take the possibility of archiving and reuse into account and plan the entire life cycle of the data from the very beginning. When formats change, conversion is needed, but it requires know-how and appropriate equipment. If the data cannot be released without some anonymisation, the degree of anonymisation needs to be decided early on.

When a dataset is preserved for further use, decisions must be made on what information will be available concerning the data and where, and in what format the data will be stored. These decisions are vital regardless of whether the data are archived at a data archive or at a university repository. Valuable datasets are of no use to research if no basic information on the data exists or can be found. If original researchers retain their data, relevant information may be lost when they retire or transfer to other jobs.

Documentation and conversion are key issues for later use. Processing old data for reuse is time-consuming and challenging, and occasionally even impossible without additional resources. The respondents voiced their fears that decision-makers would implement open access principles just by ordering people to adhere to these principles, leaving researchers to regard this as an extra administrative and technical task. They emphasised the need to create guidelines and detailed instructions which include information on the IT equipment and software needed.

> "Increasing open access is good. But we also need instructions on how to store and process data."

> "We are talking about a fundamental change which must affect scholar attitudes. At the moment researchers feel that this is an extra task they must do, in addition to all the other tasks that take time away from the research itself. It will probably take a generation before these principles become a practice, in case we are serious about their implementation."

## 2.5 Lack of informed consent and confidentiality

Research ethics and confidentiality issues are relevant to open access, particularly in cases where the research participants have not been told that the data would be archived for the use of the scientific community. The survey findings reflect the importance of these issues. Two thirds (66%) regarded the lack of informed consent as a significant barrier. The issue also came up in comments.

48% thought that open access increased risks related to confidentiality, research ethics and data protection. If the data contain identifiers, the regulations under the Finnish Personal Data Act must be taken into account. The Act dictates that a dataset from which participants can be identified must be destroyed immediately after the original research has been completed, un-

less the participants have given consent to some other option. Although few scholars actually destroy their own data, the Act certainly prevents reuse unless the data are anonymised to an appropriate level first.

If the data contain identifiers, ethics committees of Finnish hospital/health care districts have occasionally stipulated that information given to research participants must state that the data will only be used for a particular research project. Some respondents commented on this practice, saying it formed a barrier to open access. It seems likely that the ethics committees have applied a very strict interpretation of the Personal Data Act.

When information is given to participants before data collection, they are often told that the data will be used for scientific purposes only. The main thing then is that the data collected must not be released for any other purpose, for example, to authorities or people who make decisions concerning the participants. Researchers must also ensure that the data are not processed or stored carelessly and that the participants and their private affairs are not discussed indiscreetly. Researchers may talk and write about participants only for research purposes and even then in a manner that prevents identification.

At times confidentiality is equated with secrecy. In the respondent comments, some professors explained that participants had been promised that only one named researcher would use the data. Strictly speaking, however, secrecy is not the same as data confidentiality. Confidentiality of data refers to information on particular individuals and the promises given regarding the use of this information. Promises on how the data will be used, who will use it, for how long, and how the data will be processed and stored should be given to participants prior to data collection. When talking about research data, confidentiality means that participants trust that the data are used, processed and stored as agreed. In this sense confidentiality may mean a dataset that is archived in a data archive which imposes strict conditions on reuse and follows good practices in data security.

Research participants tend to have a very positive attitude towards archiving data for scientific purposes. After all, the fact that they participated in the first place is an indication that they are willing to enhance research on the subject. The FSD has contacted the participants of a few qualitative studies to ask whether they would give consent to archiving the data for research and teaching purposes, even though the original researchers had promised that only they themselves would use the data. Nearly everyone has consented. The majority said that they do not object to archiving because they had seen the relevance of studying the issue from the beginning and therefore regarded archiving a positive move that would allow other researchers to have access to the data.

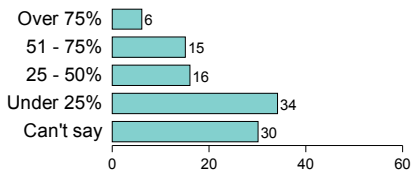
Information given to potential participants is decisive when people decide whether they want to participate or not. If the data contain identifiers, promises given to participants also determine the future of the data that is, whether the data can be reused later or whether it must be destroyed as soon as the findings have been validated. As regards confidentiality, the easiest solution in Finland is to state that the data will be archived for scientific purposes after the original research has been completed. If researchers took the entire life cycle of data into account before giving information to potential participants, confidentiality issues would not prevent scientific reuse.

# 3 Present situation in Finland

Finnish universities, government institutes, research institutes, and municipalities presumably retain tens of thousands of digital research datasets. Only a minor part is archived and reused. Without proper documentation and processing, data become unusable within 5–10 years from data collection. Even though there is no need to preserve all research datasets from public funding, archiving a significantly greater number than at present would be sensible.

The survey respondents were asked to estimate how large a proportion of the data collected during the past 10 years in their department/institute was reusable.

Figure 3.1. WHAT PROPORTION OF DIGITAL RESEARCH DATA THAT HAVE BEEN COLLECTED IN YOUR DEPARTMENT/INSTITUTE ARE REUSABLE (%).

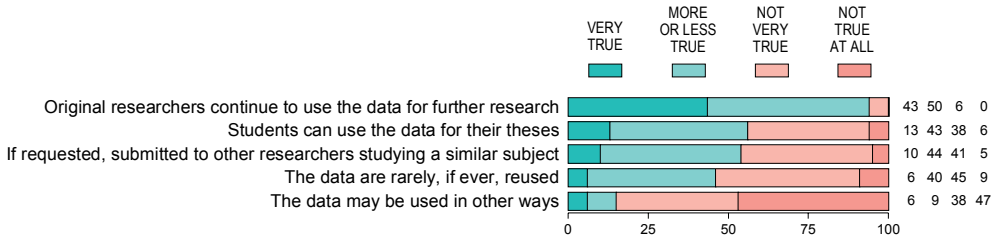



Life Cycle of and Open Access to Research Data from Public Funding (2007)

Only a fifth (21%) thought that half of the data collected were reusable. Many could not say what the situation was. The results indicate that the majority of data collected is no longer reusable. From the results of other survey questions one can deduce that the main reasons were insufficient documentation, obsolete formats or damaged data.

However, in nearly all departments some digital datasets were being reused.

Figure 3.2. IS THE FOLLOWING TRUE IN YOUR DEPARTMENT/INSTITUTE: AFTER THE ORIGINAL RESEARCH IS COMPLETED... (%).



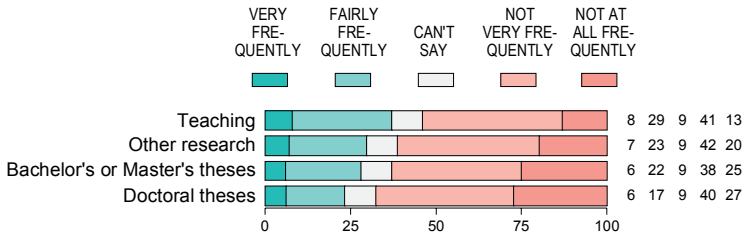

Life Cycle of and Open Access to Research Data from Public Funding (2007)

The original researchers seem to be the most frequent re-users, making use of the data they had collected earlier. 43% of the respondents thought this was very true in their department and 50% that this was more or less true.

The other specified ways to use existing data were less popular. Only one in ten thought it was very true that other researchers studying a similar subject could get access to the data. However, over 40% thought it was more or less true.

What about students? Did the respondents think students would get access to the department's data for their theses? 56% said yes and 44% no. Similarly indicative is the result that nearly every second respondent thought that the existing data were rarely, if ever, used. These findings were confirmed by the results of the question asking how frequently the data collected were being reused (see figure 3.3):

Figure 3.3. AFTER THE ORIGINAL RESEARCH IS COMPLETED, HOW FREQUENTLY ARE THE DIGITAL DATA COLLECTED IN THE DEPARTMENT/INSTITUTE REUSED FOR... (%)




Life Cycle of and Open Access to Research Data from Public Funding (2007)

Data were most frequently reused for teaching purposes, but even here only less than one in ten estimated that it was very frequent. Other uses are less common. Roughly one in four thought that existing data were being used for other research or for Bachelor's, Master's or Doctoral theses more or less frequently.

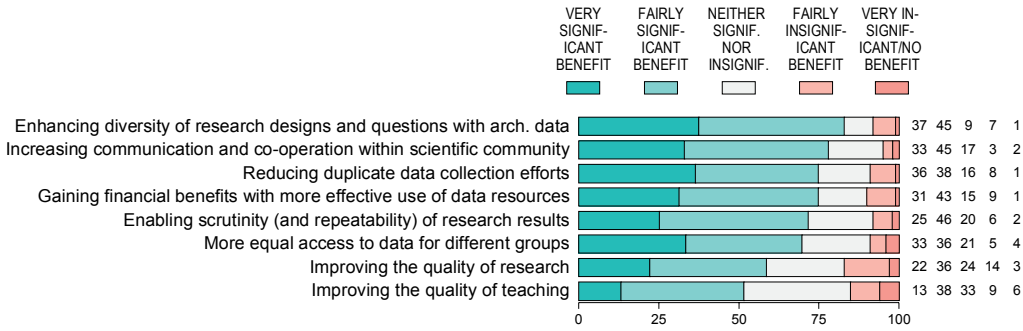
When interpreting the results one must keep in mind that the departments and researchers who reuse data were probably overrepresented among the respondents. Thus the real situation may be even bleaker than the one described here. Data are being reused for research and teaching but there is no established culture of promoting reuse. Reuse in Finland constitutes principally of the original researchers reusing the data they themselves have collected.

# 4 Opinions on the benefits of open access

How to increase open access to data and what benefits might open access bring? We have seen that the professors acknowledged many reasons for not increasing open access. Some of the reasons were connected to resources required from the original researcher, some to concerns about inappropriate use of data, and others to confidentiality.

On the other hand, increasing open access to research data collected with public funding may promote efficient and cost-effective use of data in a situation where meagre research resources force researchers to think twice before collecting their own data. Figure 4.1 below displays the respondents' opinions on the potential benefits of enhancing open access:

Figure 4.1. YOUR OPINION ON HOW SIGNIFICANT A BENEFIT ENHANCING OPEN ACCESS TO DATA WOULD BE FOR... (%).



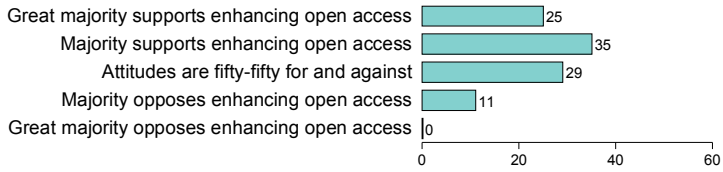

Life Cycle of and Open Access to Research Data from Public Funding (2007)

The respondents estimated all specified benefits as very or fairly significant. Only 10% chose the answer alternative 'fairly insignificant' or 'insignificant benefit' for any potential benefit. Thus, the overall result is that the benefits of open access were estimated to be more significant than the barriers (see figure 2.2 above).

The most significant benefit was estimated to be enhancing the diversity of research designs and questions with the use of archived data. The same argument is nowadays often raised also in medicine, biosciences and technical sciences. Social sciences, arts and humanities professors also recognised other benefits. One in three estimated the following benefits as significant: reducing duplicate data collection efforts, gaining financial benefits with more effective use of data resources, increasing communication and cooperation within the scientific community, and providing more equal access to data for different groups. Improving the quality of research and the quality of teaching were considered to be less significant benefits but still significant enough.

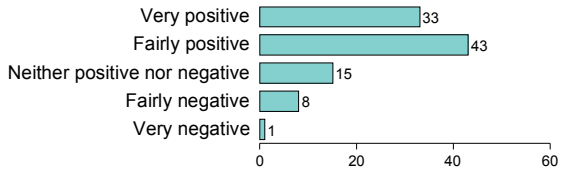
Opinions on the benefits and barriers of open access are not contradictory since it is easy for everyone to see both benefits and potential barriers. A researcher may have a positive attitude towards open access in general but a less-than-enthusiastic one to open access to his/her own data. To get a more comprehensive picture of existing attitudes, the respondents were also asked to estimate what was the general attitude of researchers in their own field to open access and what was their own attitude to open access to digital research data collected by themselves.

Figure 4.2. YOUR ESTIMATE ON THE GENERAL ATTITUDE OF RESEARCHERS IN YOUR RESEARCH FIELD TOWARDS ENHANCING OPEN ACCESS TO DIGITAL RESEARCH DATA GENERATED WITH PUBLIC FUNDING (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

Figure 4.3. WHAT IS YOUR ATTITUDE TO OPEN ACCESS TO DIGITAL RESEARCH DATA COLLECTED IN YOUR OWN RESEARCH (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

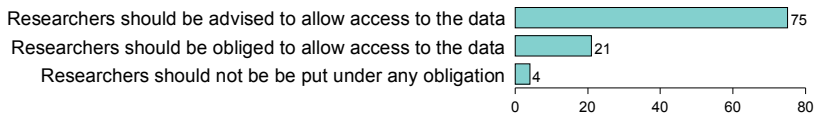
One in four (25%) thought that a great majority of researchers in their field supported enhancing open access, and 35% that the majority supports it (figure 4.2). Less than one in three (29%) thought that the attitudes were fifty-fifty for and against. Only one in ten estimated that the majority resisted enhancing open access. As for access to the data they themselves had col-

lected, one in three was very positive and 43% fairly positive. Only 15% chose a neutral attitude and less than one in ten a negative one.

There were few disciplinary differences. Education professors opposed open access a bit more than other respondents, but even among them the attitude towards open access to their own data was very or fairly positive (56%).

The third question covering attitudes towards open access was somewhat different. It cited an example case and asked how binding the guidelines on open access should be in that case.

Figure 4.4. HOW BINDING SHOULD THE GUIDELINES ON OPEN ACCESS TO DATA BE? [ Example: a dataset collected with public funding, no confidentiality or copyright problems, not actively used by the research group 5 years from the collection ] (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

The result was positive to open access even though the question did not specify which body would be giving the guidelines. Three quarters said that researchers should be advised to allow access to their data. One in five would have liked to have binding guidelines. Only a very small proportion (4%) said that researchers should not be put under any obligation.

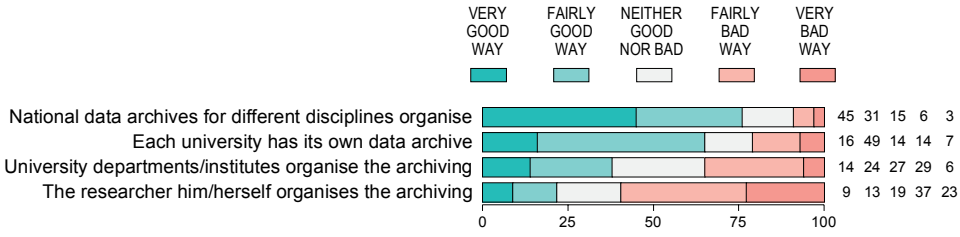
# 5 Organising archiving and reuse

There are various ways to organise archiving and access to research data. The original researcher(s) or the department/institute can retain the data. The advantage of this option is that researchers will be able to give advice to re-users and, if needed, to control the reuse. At the same time they will be aware of what kind of research their data are being used for. To guarantee open access in any real sense, metadata should be easily available, for example, on the department/institute or the research project website.

However, when potential re-users want to search for data systematically, they often find it easiest to do so through digital repositories. Digital databases of libraries and data archives are one example. The FSD, for example, describes in its data catalogue not only data archived at the FSD but also datasets that are stored by the original researchers or the research team. The problem with this practice is that researchers do not necessarily have the time to advise re-users or to get the dataset together and transfer it. Another problem is that researchers do not have the know-how or time to convert data into a format that ensures longevity. Surprisingly often the same applies to research organisations, even though they may have internal guidelines on the preservation of data.

The respondents seemed to recognise these problems since they did not support the practice of primary researchers organising the archiving and dissemination of research data themselves. The majority (60%) said that it was a fairly or very bad way to organise the archiving and dissemination.

Figure 5.1. WHAT IS YOUR OPINION ON THE FOLLOWING WAYS TO ORGANISE THE ARCHIVING AND DISSEMINATION OF DIGITAL RESEARCH DATA? (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

As regards the option where university departments/institutes would provide long-term data preservation and dissemination for reuse, opinions were divided. 38% thought it was a very good or fairly good solution while 35% considered it a bad alternative. While university departments/institutes can provide data descriptions and data catalogues in a more centralised

fashion than the researchers themselves, departments/institutes do not usually have enough staff for developing and maintaining the technical processes needed for long-term preservation, nor for administering the dissemination and reuse of data.

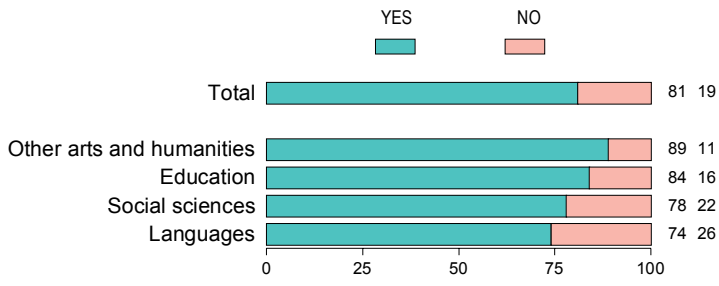
The respondents considered data archives to be the best way to organise the archiving and dissemination of digital research data. Nearly two thirds thought that data archives for each university would be a very good or fairly good way of organising the matter. However, the most popular solution seemed to be national disciplinary data archives. Almost half (45%) saw national data archives as a very good way and almost a third (31%) a fairly good way. It has to be remembered, however, that establishing national data archives for different disciplines also requires resources, planning, guidelines, and implementation of best practices.

The advantage of discipline-specific data archives is that they would have the necessary know-how. They would thus lessen the workload at the individual and department levels, because, without data archives, open access would mean increasing the workload of researchers. This is a point well worth considering at a time when researchers often feel they have far too little time for research.

## 5.1 Views on the OECD Recommendation

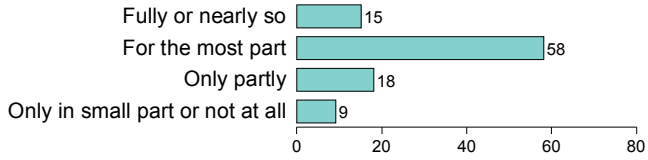
The respondents were asked whether the survey in question was the first time they had received information on the OECD guidelines. Four fifths (81%) answered it was. The differences between disciplines were quite small (figure 5.2). Another question charted the extent to which the respondents thought the OECD guidelines could be implemented in their own research field. The responses are presented in figure 5.3 below:

Figure 5.2. IS THIS SURVEY THE FIRST TIME YOU RECEIVE INFORMATION ON THE OECD GUIDELINES ON OPEN ACCESS TO RESEARCH DATA? (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

Figure 5.3. IN YOUR OPINION, TO WHAT EXTENT CAN THE OECD GUIDELINES BE IMPLEMENTED IN YOUR OWN RESEARCH FIELD? (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

15% thought that the OECD guidelines could be implemented fully or for the most part. The majority thought they could be implemented for the most part. Roughly one respondent in five selected the alternative "only partly" (18%) and a mere nine per cent thought that the principles could be implemented only in small part or not at all.

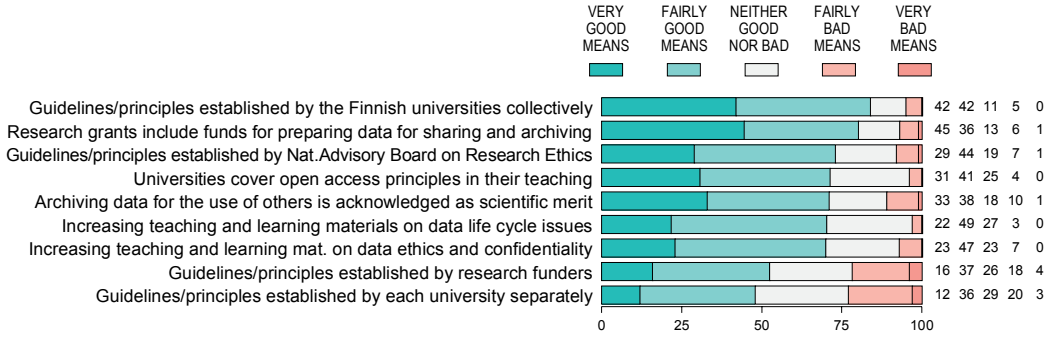
Professors of social sciences and linguistics had the most positive attitude towards the possibility of implementing the OECD guidelines. Three in four thought that the guidelines could be realised for the most part. In addition, over a fifth (22%) of the professors of social sciences believed that the principles could be implemented fully or nearly so. Among the professors of linguistics, the corresponding proportion was somewhat smaller (14%), but on the other hand, they had the largest proportion of respondents who thought that the guidelines could be realised for the most part (65%).

Compared to professors of education, the representatives of the two disciplines mentioned above were much more positive: in fact, only five percent of the professors of education selected the most favourable alternative. If we look at the other end of the response scale, it appears that the representatives of other humanities disciplines were the most reserved as regards the possibility of implementing the guidelines. There may be several reasons for this. The positive attitude of the professors of social sciences can perhaps be explained by the establishment of the Finnish Social Science Data Archive in 1999. In addition, open access to research data has been widely discussed in the field in recent years. Linguistic datasets usually contain less confidentiality issues than, for example, qualitative datasets that are popular in the fields of social sciences and education. The relatively negative attitude of the professors of education can perhaps be explained by the fact that datasets containing pupil information are typically considered sensitive. Respondents in those datasets can be identified more easily than in datasets based on random samples representing the whole population.

## 5.2 Implementation of the OECD Recommendation: opinions on the means and the responsible body

The survey also studied opinions on the ways to implement open access principles, as well as opinions on which bodies should take part in drawing up the guidelines on open access to research data (figures 5.4 5.6).

Figure 5.4. HOW GOOD OR BAD DO YOU REGARD THE FOLLOWING MEANS TO BE FOR IMPLEMENTING OPEN ACCESS TO RESEARCH DATA IN YOUR FIELD? (%).



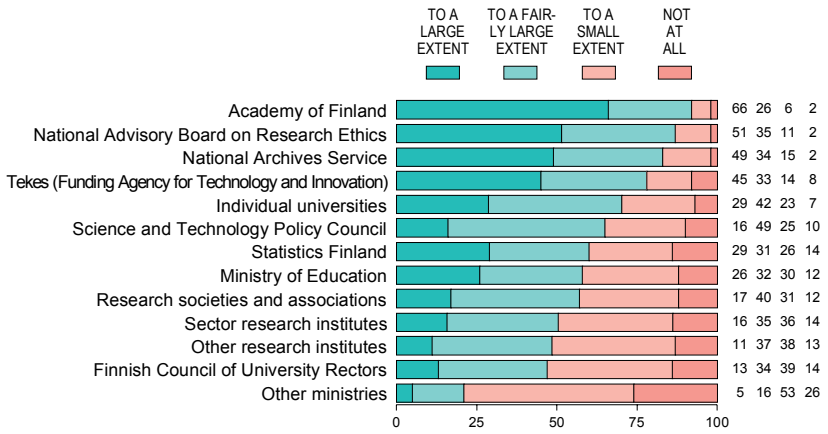

Life Cycle of and Open Access to Research Data from Public Funding (2007)

The means listed here do not exclude each other – on the contrary, it is possible and perhaps even desirable to use more than one at the same time. The respondents favoured two options: guidelines/principles established by the Finnish universities collectively, and research grants which contain funds for preparing data for sharing and archiving. These two means were the only ones that over 40% saw as very good. Over a third considered them to be fairly good, which means that altogether over 80% regarded these means as fairly or very good.

Five other options were also supported by the majority of the respondents. The order of preference was as follows: guidelines/principles established by the National Advisory Board on Research Ethics, universities cover open access principles in their teaching, archiving data for future reuse is counted as scientific merit, increasing teaching and learning materials on data life cycle issues, and increasing teaching and learning materials on data ethics and confidentiality. Guidelines/principles established by research funders or by each university separately were regarded as less significant.

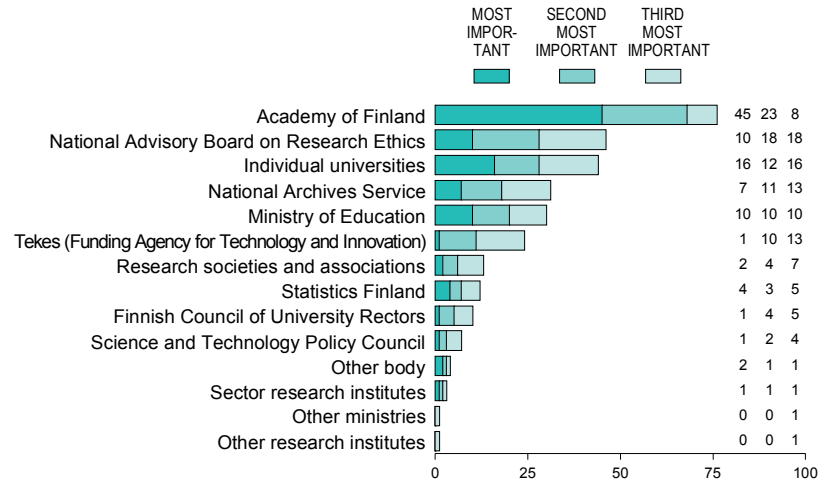
Finally, the respondents were asked to evaluate to what extent various bodies should participate in drawing up the guidelines on open access to digital research data generated with public funding. The respondents estimated the role of each body by using a four-point scale ranging from "to a large extent" to "not at all". They were also asked to specify what they thought were the three most important bodies in formulating the guidelines out of the listed 13.

Figure 5.5. TO WHAT EXTENT SHOULD THE FOLLOWING BODIES TAKE PART IN DRAWING UP THE GUIDELINES ON ACCESS TO DIGITAL RESEARCH DATA GENERATED WITH PUBLIC FUNDING (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

Figure 5.6. SPECIFY THREE BODIES WHICH YOU WOULD CONSIDER AS THE MOST IMPORTANT FOR DRAWING UP THE GUIDELINES (%).




Life Cycle of and Open Access to Research Data from Public Funding (2007)

The Academy of Finland was overwhelmingly the most popular choice: three in four respondents included it among the three most important bodies (see figure 5.6 above). At the same time, two thirds agreed that the Academy should participate in formulating the guidelines to a large extent. According to both figures, the National Advisory Board on Research Ethics was regarded as the second most important body.

# 6 Views on implementing the OECD Recommendation

It is a definite advantage to society as well as to the national and international scientific communities that research data can be used after the original research has been completed. This ensures the efficiency of public research investments and maximises their impact, in addition to increasing the productivity of research.

There are a number of scientific disciplines and research areas, all of which operate in different environments, studying a great variety of subjects. There are differences in study goals, degree of confidentiality, degree to which the results are commercially exploitable, financing structures, and reuse potential of data. Laws and regulations on research differ from country to country. Thus, it is understandable that the OECD guidelines concentrate on publicly funded research and seek to increase open access to data by providing general principles.

In the following, the results and operational models presented in the preceding sections are summarised and compared to the principles stated in the OECD guidelines. This is one way to analyse the guidelines and make them more concrete. Hopefully this will help to bring the national debate in Finland from the discussion of principles to the discussion of what concrete actions should be carried out at the practical level.

## 6.1 Summary of key factors in data life cycle management

### Research plan and ethical evaluation

Careful planning and preparation before data collection will enhance the openness and long life cycle of research data. Each research project needs to have a meticulously written research plan which clarifies the following issues, if the nature of the research so requires:

- if the research concentrates on individuals, specification on what kind of information will be given to research subjects, as well as templates for the consent form and other permission and agreement forms;
- the ownership and copyright of the dataset, and who has control over it, especially if not accordant with the standard practices of the research organisation;
- a plan on how the data will be processed, used, and stored during the original research;
- a plan on how access to data will be ensured after the original research has been completed.

### Research funder support for open access to data

Research funders can support open access to data by recommending or requiring that the data collected with their grants will be made available for the use of the scientific community after the original research has been completed, and by supporting this recommendation/request financially. Potential support measures include:

- formulating and presenting general ethical principles for research,
- establishing own data policy which is binding on grant holders,
- providing funding as part of the research grant for preparing the data for archiving and sharing.

### Securing the reuse potential during data collection

Data collection forms a crucial stage in data life cycle management. Decisions made at this stage cannot be changed afterwards, and they determine the reuse potential of the data. Particularly the information and promises given to research participants on the future use of the data have a direct impact on the reuse potential. Also of relevance are how well the data collection succeeds and how well the results and different aspects of the collection are documented.

### Primary use

The primary use of research data means using the data according to the purposes stated in the original research plan. It is often difficult to determine the duration of the primary use stage, especially at the beginning of a research project. However, this should not result in leaving the duration totally open. During the primary use, attention should be paid to sufficient documentation at all stages and proper preservation of the data. From the viewpoint of open access to research data, it is essential to know who are the persons who will decide whether the data will be released for secondary use, and what kind of reuse is possible and when. Unclear delineation of responsibilities and prolonged or extended primary use may result in the data being underused.

### Archiving and publishing data

Without systematic preservation procedures, usability of digital research data may diminish drastically within a couple of years. In the long run, data may even be completely destroyed if proper preservation measures are not undertaken. Managing the life cycle of digital research data requires decisions on which datasets will be preserved and on long-term preservation measures (i.e. archiving). The latter also entail meeting the standards set for documentation and long-term preservation of data.

In some cases, it seems best that the organisation which has collected the data will be responsible for its preservation and reuse. In this scenario secondary users have an opportunity to get more information about the data directly from the collector.

From the viewpoint of long-term preservation and reuse, it is definitely less recommendable to leave the responsibility for the preservation and dissemination of data to individual researchers. Changing this practice that still prevails in Finnish universities and other Finnish research organisations constitutes one of the key goals in the national implementation of the OECD Recommendation.

The safest solution is to let experts take care of data preservation. This scenario ensures open access to data, guarantees long life cycle of data and ensures efficient use of research investments. It requires experts who are working either within research or data collection organisations, or within service units which offer centralised archiving solutions and to which long-term preservation and dissemination of the data can be outsourced.

Centralised archiving is the best way to ensure that research data are documented according to national or international standards and that access to data is provided not only to Finnish researchers but also the international scientific community.

It is also essential to note that data archives publish study descriptions of research data, bibliographical citations to the data and other additional information that will facilitate reuse, and compile databases of this information. The databases also contain bibliographic citations to publications based on the archived data, thus enabling researchers to get a deeper understanding of the earlier research in their field both through existing publications and through existing data.

### Supporting reuse of data

Open access to research data requires that documentation on data and the overall quality of data are sufficient. Nowadays it is possible to handle confidential information securely enough through various online applications provided that the data systems and interfaces offer a reliable user registration, the definition and identification of usage rights, and the management of confidentiality issues as regards the data content.

Access to digital data containing very sensitive information can also be arranged through on-site solutions, in which the data are used on site at the depositing organisation responsible for securing the confidentiality of the data. The depositing organisation can also offer guidance to secondary users.

## 6.2 Benefits of implementing open access

It is not always necessary to collect new research data. If more extensive information were provided on the existing data resources, the scientific community would find it easier to identify new information needs. Open access reduces unnecessary data collection.

There are more advantages than disadvantages in the OECD Recommendation. Table 6.1 below specifies the expected benefits. Openness means much more than just saving money.

## 6.3 Open Access: research publications vs. research data

The digitisation of research environments is a substantial argument for open access to research data. The ongoing digitisation process has already had a profound impact on scientific research practices. The principle of openness promoted in the OECD Declaration can to a large extent be implemented with the help of Internet services. Therefore, metadata on research projects

| Type of benefit | Benefit / label |
|---|---|
| Financial | Using research data more efficiently and maximising impact. |
| Scientific | Increasing the openness of science, providing up-to-date data, repeatability and controllability of research, increased interaction through the joint use of data. |
| Research ethical | For example, taking research subjects into account when collecting and using data. |
| Judicial | Taking legislation into account when collecting, using and preserving data. |
| Democratic | Defining and specifying the position and responsibilities (rights and obligations) of research subjects, researchers, research organisations, and funders. |
| Equal | Promoting more equal access to research data. |
| Social | Improving the quality and cost-effectiveness of the information resources of society. |
| Political, specific | The OECD Recommendation obliges the member countries. |

Table 6.1
Benefits of implementing the OECD Recommendation

must be more readily accessible on the web. In addition, there must be open access to the research data itself securely through the web for informed use. This is a key issue in the development of attractive and competitive research environments.

The Open Access objectives, which concentrate on digital publications, cannot be wholly adapted to digital research data as such. The openness of research data is usually restricted to the scientific community. In addition to intellectual property rights, research data also raises the issues of data protection and confidentiality. However, the key objectives behind open access to publications and research data are the same: open access to and usability of scientific information promote the development of science and equal access to that information. Cultural and attitudinal barriers to increasing open access are similar, at least partly (see for example Björk 2004).

Table 6.2 below summarises some differences between digital research publications and digital research data. As regards different types of research data, there are several distinct issues which can be solved only through discipline-specific and situation-specific considerations.

Some scientific publications have a data policy which they apply to raw data on which the publication is based on. Journals of natural science often require that the data analysed in the published article is available to other researchers. They recommend that the data should be archived or directly downloadable from the web. This data policy is often applied in the fields of chemistry, astronomy, biology, and physics.

| Research publication | Research data |
|---|---|
| Information transformed into results | Information not transformed into results |
| Use requires basic software and instruments and their command | Use often requires special software and instruments and their command |
| Self-explanatory | Requires additional information and documentation if not archived |
| Should not include sensitive information | May include sensitive and confidential information |
| Use does not require permission | Use often requires permission |
| Ownership and copyright often clear | Ownership and copyright often unclear |
| Openly accessed by the scientific community for a fee or for free | Several degrees of openness (from completely open to closed) |
| Understood as scientific output (mentioned in the CV) | At the moment not understood as scientific output/merit even if the data were published (usually not mentioned in the CV) |
| Ready to be used by others as such | Use requires processing |

Table 6.2
Research publication, research data and Open Access: a simplified difference chart

Humanities journals tend to have less strict guidelines. Sometimes they require that the data analysed in the published article should be available to the editorial staff for validation if necessary. In addition, they may require that the research data is archived, or if it is desirable from the viewpoint of data protection oblige the researcher or research organisation to preserve the data, for example, for five years commencing from the date of publication. This data policy, often implemented by psychology journals, ensures that the analyses performed on the data can be verified later, if necessary.

Ray and Valeriano (2003) give numerous examples on the data policies of scientific journals. For example, International Interactions (published by Taylor & Francis) requests that the researchers submitting articles for publication archive their data or demonstrate the availability of the data for other researchers by other means. If the data do not include information that enables identification of individuals, the data may be required to be published in its entirety. Journal of Peace Research (PRIO International Peace Research Institute) and Political Analyses (Oxford Journals) request the authors of articles to provide their readers with access to the original data in order to let them verify the analyses at will. (Mt. 77–78)

When researchers choose to retain their data themselves, dissemination for further use requires that they compile a user guide and ensure that the data are in a format compatible with contemporary technology environments. If researchers have not documented and transformed the data into a preservation format during the research or immediately after completing it,

dissemination for reuse will come more and more labour-intensive as time goes on. Researchers rarely possess sufficient knowledge of the fundamental principles of digital archiving, and therefore King (1995, 446–447), for example, recommends depositing data in a data archive.

## 6.4 Extensive cooperation needed to support the implementation of the OECD Recommendation

Promoting open access to digital research data in Finland will require extensive and long-lasting cooperation between various authorities and the Ministry of Education. The cooperation should produce national recommendations and operational models to promote the use of and open access to research data.

The existing international models and operational strategies should be taken into account when developing national guidelines, recommendations and data policies. The cooperation should involve publicly funded research organisations collecting data, key research funders, and scientific organisations from various disciplines. From the point of view of the Ministry of Education, the key actors include the Academy of Finland, Finnish universities, the Finnish Council of University Rectors, the National Advisory Board on Research Ethics, and the Committee for Public Information. Promoting open access to digital data can take the form of extensive discussion forum, for instance. The agenda of the forum could include at least the following issues:

1. National-level discussion and conceptualising of the general operational models of implementing the OECD Recommendation.
2. Developing research design and agreement practices that support the long life cycle of research data.
3. Clarifying the rights and responsibilities of actors connected to research data.
4. Improving data life cycle management through training and education.
5. Encouraging research funders and data collectors to create data policies.
6. Discussion and proposals for recommendations and research funder policies that would promote the reuse of data.

Establishing and asserting rules, recommendations and operational practices supporting open access to research data would be a big leap forward in Finland in terms of science and research policies, and would significantly improve the quality of our national research environment. Defining the principal practices and setting the strategic objectives is a task in which the whole scientific community and its partner organisations should take part. Therefore, the national implementation of the OECD Recommendation requires long-term strategic planning and cooperation across different disciplines and between all parties involved. Practical solutions connected to operational environments, data preservation, technical questions, and legislative issues require interdisciplinary discussion and clarification.

Some issues can be solved in a reasonably short time. The means for changing the prevailing operational and cultural practices are mostly in the hands of research funding bodies. The Academy of Finland has reacted to the prevailing problems by starting to require this autumn

(2008) that a long-term data management plan must be submitted with funding applications, which should lead to better planned data collection, processing, and preservation measures. Hopefully this will ensure that in the future valuable research data will no longer be in danger of becoming obsolete and outdated.

# Literature


Björk, Bo-Christer (2004) Open access to scientific publications – an analysis of the barriers to change? Information Research 9(2). Available: http://informationr.net/ir/9-2/paper170.html [Accessed 10.5.2007]

Clubb, Jerome & Austin, Erik & Geda, Carolyn & Traugott, Michael (1985) Sharing Research Data in the Social Sciences. In Fienberg, Stephen E. & Martin, Margaret E. & Straf, Miron L. (eds.) Sharing Research Data. Washington, D.C.: National Academy Press, 30–88.

Corti, Louise (2006a) Editorial. Methodological Innovations Online 1(2). Available: http://sirius.soc.plymouth.ac.uk/~andyp/viewarticle.php?id=33&layout=html [Accessed 11.12.2006]

Corti, Louise (2006b) ESDS Qualidata: accessing, exploring and using data. Research methods festival. Oxford, July 2006. Available: http://www.ccsr.ac.uk/methods/festival/programme/rwa/ [Accessed 31.1.2007]

Fielding, Nigel (2000) The Shared Fate of Two Innovations in Qualitative Methodology: The Relationship of Qualitative Software and Secondary Analysis of Archived Qualitative Data [43 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 1(3). Available: http://www.qualitative-research.net/fqs-texte/3-00/3-00fielding-e.htm [Accessed 30.11.2006]

Gleditsch, Nils & Metelits, Claire & Strand, Håvard (2003) Posting Your Data: Will You Be Scooped or Will You Be Famous? International Studies Perspectives 4(1), 89 97.

Kiikeri, Mika & Ylikoski, Petri (2004) Tiede tutkimuskohteena: filosofinen johdatus tieteentutkimukseen. Helsinki: Gaudeamus.

King, Gary (1995) Replication, Replication. PS. Political Science and Politics 28(3), 443–499.

Kuula, Arja (2006) Tutkimusetiikka aineistojen hankinta, käyttö ja säilytys. Jyväskylä: Vastapaino.

Leppänen, Markku (2006) Miten tutkimusaineistojen säilytysarvo tulisi määritellä? Esitelmä Arkistoyhdistyksen syysseminaarissa 3.11.2006 Tieteiden talo, Helsinki.

Mauthner, Natasha & Parry, Odette & Backett-Milburn, Kathryn (1998) The data are out there, or are they? Implications for archiving and revisiting qualitative data. Sociology, 32(4), 733–745.

Niiniluoto, Ilkka (2002) Jaettu vastuu kannattelee hyvää tiedettä. Tieteessä tapahtuu 20(4). Available: http://www.tsv.fi/ttapaht/024/niiniluoto402.pdf [Accessed 11.5.2007]

Open Access to and Reuse of Research Data 2006 [computer file]. FSD2268, version 1.0 (2007–07–30). Borg, Sami & Kuula, Arja (Finnish Social Science Data Archive) [authors]. Tampere : Finnish Social Science Data Archive [distributor], 2007.

Sieber, Joan (1991) Social Scientists' Concerns About Sharing Data. In Sieber, Joan (ed.) Sharing Social Science Data: Advantages and Challenges. London: Sage.

Ray, James & Valeriano, Brandon (2003) Barriers to Replication in Systematic Empirical Research on World Politics. Julkaisussa deMesquita, Bruce Bueno & Gleditsch, Nils Petter & James, Patrick & King, Gary & Metelits, Claire & Ray, James Lee & Russett, Bruce & Strand, Håvard & Valeriano, Brandon (2003) Symposium on Replication in International Studies Research. International Studies Perspectives 4(1), 72 107. Available: http://gking.harvard.edu/files/replvdc.pdf [Accessed 24.1.2007]

Thompson, P. and Lummis, T., Family Life and Work Experience Before 1918, 1870–1973 [computer file]. 6th Edition. Colchester, Essex: UK Data Archive [distributor], February 2008. SN: 2000.

Finnish Social Science
Data Archive


Åkerlundinkatu 2 A, 5th floor, Tampere

| | |
|---|---|
| Adress: | Finnish Social Science Data Archive |
| | FIN-33014 University of Tampere |
| | Finland |
| Tel: | (03) 3551 8519 |
| Fax: | (03) 3551 8520 |
| E-mail: | fsd@uta.fi |
| WWW: | http://www.fsd.uta.fi |