Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators

(i) Background and UH Process for Digital Preservation (Fairdata-PAS)
As we all know, research outputs in the form of scientific publications are very well restored and available, since there are tons of professional peer-reviewed journals for academics to publish their results. Further, more and more journals are adhering to open access policy and hence, research results are becoming publicly available. Nowadays there are also several repositories for research data. However, repositories vary in how long storing time they can guarantee for the data. Therefore, the Finnish Ministry of Education and Culture has established Fairdata-PAS service for Finnish research organizations for long-time storing the nationally most significant research data (see more here: https://www.fairdata.fi/en/fairdata-pas/).

Fairdata-PAS is meant for digital preservation of research data for several decades, or even centuries. Each research organization has a limited space for Fairdata-PAS preservation, and preservation also requires preparatory actions. Therefore it must be carefully considered which research data are worth long-time storing.
At the University of Helsinki, the process begins when a researcher proposes a significant data for long-time, Fairdata-PAS preservation. Researchers are first instructed to be in contact with UH Data Support with an online form to evaluate the degree of data documentation. After this proposals go further to Scientific Committees of Faculties, and proposals coming e.g. from researchers of the Faculty of Arts are evaluated by the Scientific Committee of the Faculty of Arts etc. At the Committee meeting, the researcher introduces the data as well as the *value* of the data (10 min presentation), and a Senior Advisor from UH's Research Administration acts as a representing officer. The Committee will make a decision on preservation (yes – no).

(ii) Research Data and Technical Requirements for Digital Preservation (Fairdata-PAS)
It is important to acknowledge that there is no universally accepted definition for 'research data'. Different disciplines have varying conceptions of what counts as data, and the same applies to public funding organizations. For that reason it is also crucial that there is a peer-review mechanism involved in evaluating the need for long-time digital preservation (Fairdata-PAS).

Research data can be qualitative or quantitative, and comes in print, digital and physical formats. Research data may include all of the following:
• Documents (text, Word), spreadsheets
• Laboratory notebooks, field notebooks, diaries
• Questionnaires, transcripts, codebooks
• Audiotapes, videotapes
• Photographs, films
• Test responses
• Slides, artefacts, specimens, samples
• Collection of digital objects acquired and generated during the process of research
• Data files
• Database contents (video, audio, text, images)
• Models, algorithms, scripts
• Contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
• Methodologies and workflows
• Standard operating procedures and protocols

Accordingly, researchers in the humanities often utilize physical questionnaires as data, social scientists have different statistical analyses, natural scientists operate with raw data obtained at an observatory, and medical scientists work with biological samples. To repeat, Fairdata-PAS is meant for *digital* preservation of data.

Before a researcher's proposal for long-time digital preservation is addressed at the meeting of Scientific Committee, Library's Data Support team has ensured that certain technical requirements for data are fulfilled. Briefly, the data must be convertible or processable with reasonable costs for future use. This technical check is also the first check on the overall quality of the data, viz. if the data as a whole is too scattered or uncontextualized as such, then long-time digital preservation can be unjustified, even if the data could have some inherent value for

future use. Pass on the technical check also implies that mandatory anonymization of the data would not render the data useless. Data Support team also considers whether someone of the existing, global data archives would be more suitable storage solution for the data, either as an alternative or parallel solution.

(iii) Decision making

Data preservation is linked to the integrity of science; namely, scientific discoveries must be replicable, and therefore data must be restored. Long-time preservation then enables validation of past discoveries, but restored data can also lead to new discoveries in the future. In a nutshell, long-time data preservation aims to store significant data that is perceived as valuable for scientists' future use. Of course also non-scientists can access the data from long-time preservation, but the assumption here is that the main target group is researchers, since it requires scientific training to know how to use the data. Storing significant data also ensures an unbroken chain of evidence in research projects and more efficient scientific collaboration, since all have access to the same data. The overall reliability of science could be also be improved in the long run, since access to the data diminishes the dependability on researchers' interpretations.

As said earlier, the significance of data is discipline-dependent, and therefore the Scientific Committee of respective Faculty is in the best position to assess whether some particular data is significant enough to warrant the need for long-time digital preservation (Fairdata-PAS). With respect to each case, the Committee have to form a unanimous decision ('YES' / 'NO'). If there is not consensus, then the Committee employs majority-rule voting, and the Chair's vote is decisive in case of even votes. The Committee utilizes the following evidence for the decision:

***Researcher's grounds for long-time preservation**: both written material (the proposal) and reasons brought out orally or with the help of presentation material at the Committee meeting.

***Representing officer's reasoning**: representing officer's viewpoints for the case.

***The Committee's discussion**: since there are various aspects to consider, the Committee's discussion on the case is an invaluable method for finding out all relevant pros and cons. Especially the very first cases are crucial, since there the Committee begins to form its collective view or evaluation policy for long-time preservation.

***The former decisions by the Committee**: the former positive and negative data preservation decisions together form an important pool of yardsticks for the Committee, and similar cases should be treated similarly. This also implies that the minutes of the former meetings should be detailed enough and easily accessible.

The Committee should consider all relevant viewpoints for the case, and form a consensus/majority decision. Value triangle (Fig. 1) may be useful here to figure out the most pressing viewpoints for the relevance of the data. Also researchers are asked to utilize the value triangle when preparing the proposal for long-time digital preservation (Fairdata-PAS). Below is a list of questions the Committee may utilize in the process of making its decision.

EXPECTED USES FOR DATA

*As a whole, is the data comprehensive enough to enable varying utilizations in the future? For example, data of longitudinal studies may be valuable due to temporal comprehensiveness.
*As for the technical requirements of the data, is the data sound enough to enable varying utilizations in the future? Or if the data has to be fixed technically speaking, does the expected value of data outweigh the costs arising from fixing the data?
*Is the data usable as complementing other data?
*Is the data usable as a point of comparison for other data?
*Is the data only partially analyzed?
*Is it reasonable to expect that with future research methods the data can be utilized even more? As an example, currently we have a lots of information about genomes, but yet we don't know how to best utilize or analyze that information.

POTENTIAL FUTURE VALUE
*Is the data crucial for the future progression of the discipline or some areas of it?
*Can further utilization of the data lead to significant scientific discoveries or publications?
*Is the data scientifically or culturally unique?
*Can further utilization of the data lead to commercial applications, business collaboration or patents?
*Does the data have significant educational value, so it could be utilized e.g. in researcher training?

VERIFIED SIGNIFICANCE
*Has the data been utilized in some particularly significant publication or scientific discovery?
*Is the data crucial for national or global research infrastructures?
*Has the data been utilized in significant research collaboration between various organizations?
*Has the production of data in itself required significant investments and resources?
*Has the data already previously been assessed by e.g. Ethics Committee? Committees' reports may be useful for decision-making.

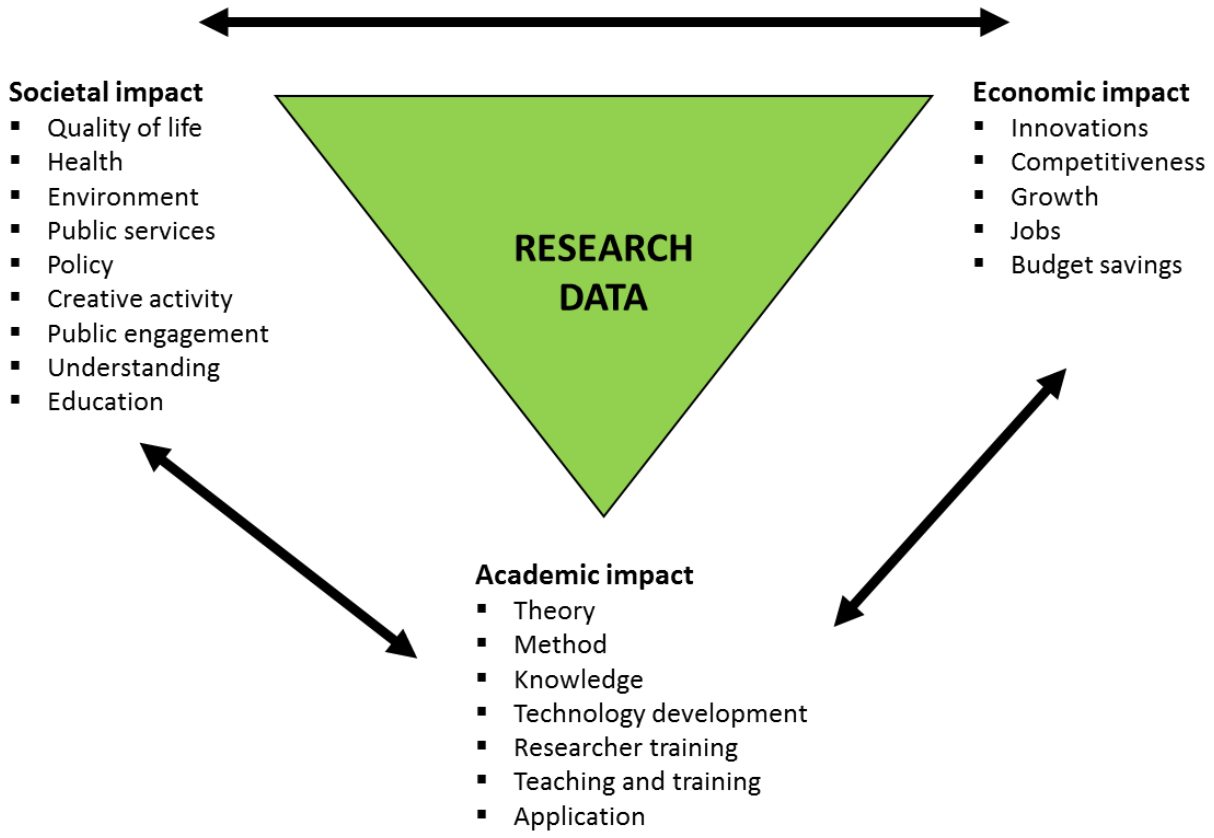

Figure 1. Value triangle for research data

'Impact of research' refers to realized influences or outcomes of research, very often taking place after several years of the actual research work. For example, medical research can led to an invention of a new drug, and its distribution provides income, or economic impact. Clearly, disciplines differ in expected impacts, e.g. historical research can mostly like provide more understanding and cultural impact, whereas life sciences very often contribute for public health and thus have health impact. And yet all disciplines are expected to have academic impact, e.g. advancement of theories and models. Now, research data can potentially lead to various kinds of impacts, and the value triangle is meant to be a visual tool for figuring out the most obvious kinds of impacts.