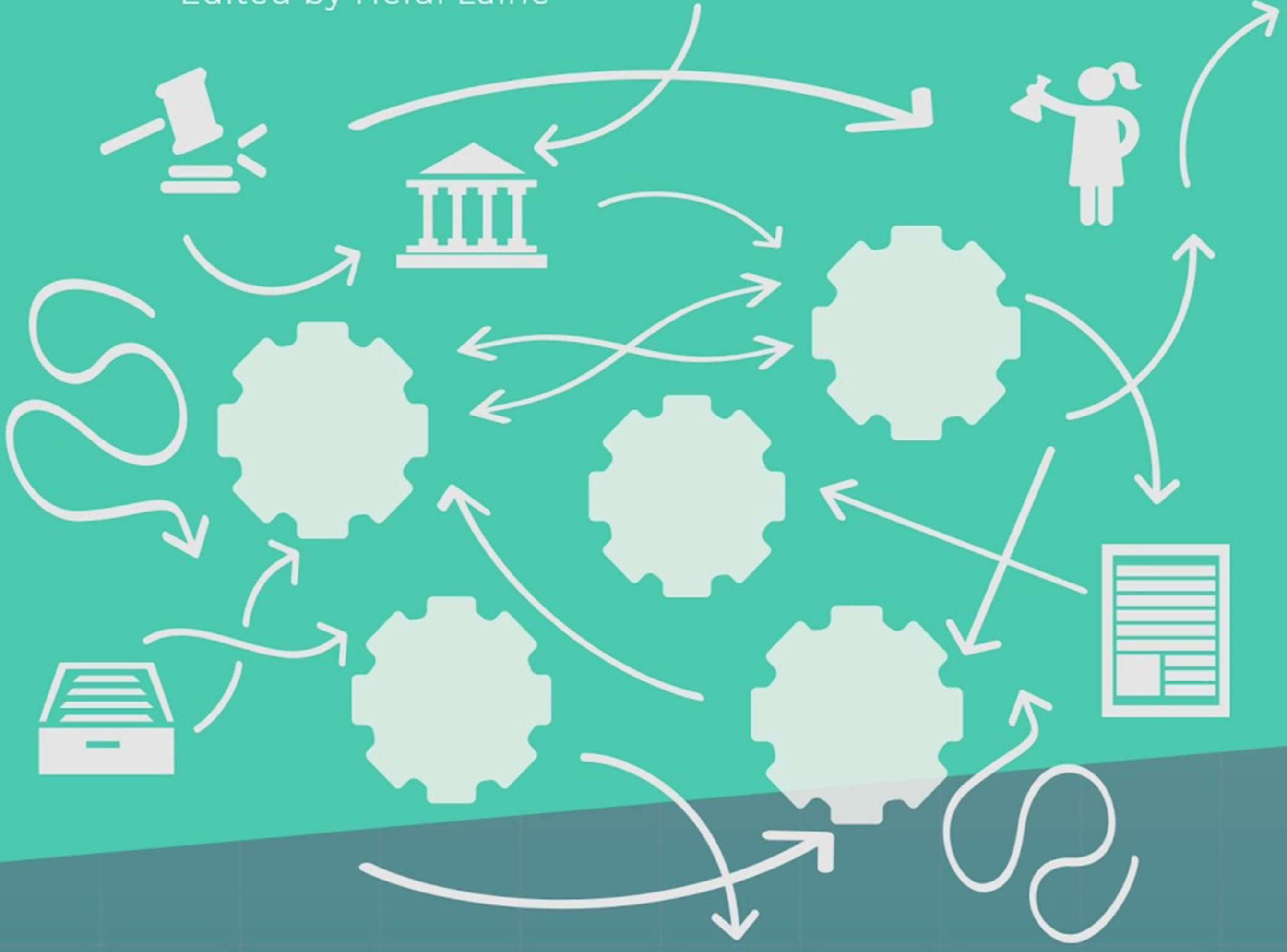
TRACING DATA

Data citation roadmap for Finland

Edited by Heidi Laine



Finnish Committee for Research Data
www.fcrd.fi

URN:NBN:fi-fe201804106446

Published 10 April 2018 in Helsinki, Finland, by Finnish Committee for Research Data

www.fcrd.fi / info@fcrd.fi / @FinCRD

Edited by Heidi Laine, ORCID: 0000-0002-6658-0664

Summary

This document presents a roadmap for the Finnish research community for implementing research data citation practices. The roadmap consists of an evaluation of the current situation, description of the target state and recommendations on measures that would lead from the current situation to the target state. It also presents an information model for data references.

The roadmap has been produced by the Finnish Committee for Research Data (FCRD) in dialogue with other members of the Finnish research community. The Ministry of Education and Culture Open Science and Research Initiative has instigated and funded the work.

Data citation is considered to be one of the core processes of an open scholarly research system. Thus far, data citation practices are poorly implemented worldwide, but once established, they are expected to facilitate the crediting of data work, providing attribution detail, facilitating access, fostering collaboration, and ensuring transparency and reproducibility of science and scholarship. Finland has an opportunity to set an example to other national research systems, thus solidifying our position as a global leader in open science.

To ensure international interoperability the Tracing Data Project has used the FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles (2014) as a key reference as well as a conceptual framework. The full declaration can be found at www.force11.org/datacitation.

The FORCE11 declaration is divided into eight principles, which are:

1. Importance
2. Credit and Attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and Verifiability, and
8. Interoperability and Flexibility.

Recommended measures are presented in the roadmap both by principle and by stakeholder group. Data citation stakeholders are:

- Researchers
- Decision makers
- Institutions
- Data repositories
- Publishers, and
- General public.

Data repositories are institutions that store and curate data. They create the infrastructural foundation for the data citation process. According to the roadmap, repositories are responsible for assigning persistent identifiers (PID's) for data and creating and maintaining data landing pages. Identifiers both identify a data set and provide access to it. When typed to an internet browser address bar, they lead to a data landing page. Research data is accessed only through landing pages, never directly. Data landing page should hold information on the data set, such as metadata and suggested model for referencing the data set in question.

Academic publishers have the role of making sure that publishing authors give due credit to data creators whose work they are utilizing and use the data reference information model. The most important element of the information model is the persistent identifier, as it allows the access to the data source. Data references need to be both machine and human readable, so PID doesn’t suffice alone as a data reference.

Data can be referenced in many contexts; journal articles are only one possible instance. The data reference information model applies also to data references made in blogs, social media, and other data sets. Also, software code, which is increasing its importance as a research output, can be referenced using the information model presented in this roadmap, as long as the code is either published or described openly and given a PID.

Research institutions need to educate both students and researchers about data citation. Institutional data policies are an important instrument for responsible data management. Merely publishing policies is insufficient; they need to be enforceable and enforced. Researchers have the responsibility of following the policies and engaging in dialogue with other stakeholders to make sure that the policies are practical.

Decision makers, funders and policy makers need to recognize data as a research output, give consistent support to data infrastructure maintenance and development and participate in developing and implementing transparent and responsible informetrics.

The general public is not assigned any action in this roadmap but is recognised as the ultimate end-user of all research outputs, as well as a source of legitimization for public spending on scientific and scholarly research. Academia cannot thrive without broad societal support.

Data reference should consist of following elements:

Creator, title, host organisation, publication time and/or date, persistent identifier.

Useful additional elements are:

Version, resource type, license status, ORCID, embargo information.

DATA CITATION RECOMMENDATIONS OVERVIEW

Data repositories

- All datasets intended for citation must have a globally unique persistent identifier that can be expressed as unambiguous HTTP URI.
- Finnish data repositories should use either DOI or URN as their PID of choice, since they are the best managed and most reliable PIDs in the Finnish environment.
- The persistent identifier must resolve to a landing page that supports access to the actual data set.
- Assigning PIDs and creating landing pages is the responsibility of the data repository.
- Landing page should facilitate access to metadata, either by holding metadata or a link to metadata.
- The landing page should include reference model for citation, and ideally also metadata helping with discovery, in human-readable and machine-readable format.
- National data centers, libraries and archives should agree on the required metadata content of a data landing page.
- Data that no longer exists should have a persistent landing page, which may direct the user to a current version of the old data set.
- License all metadata with a CC0 license or equivalent.
- Make metadata freely harvestable through open APIs.
- The persistent identifier must be embedded in the landing page in machine-readable format.
- Pilot the RDA Data Citation model for dynamic data in one or several national data centers.
- Release all data citation related content intended for broad audiences, such as guidelines and standards, in open format, i.e. CC-BY, or equivalent.

Researchers, learned societies

- Include principles of data as evidence and data transparency in next version of Finnish RCR guideline by TENK.
- Update the TENK CV template to increase the visibility and prestige of research data outputs.
- Recognise data creatorship as a separate issue from text authorship, as well as the need for new concepts and guidelines for crediting creators of research data. Create a multi-institutional, multi-disciplinary national working group to write a similar guide for assigning data authorship as the TENK text authorship guideline, coordinated for example by TENK, or assign suitable national representation to a relevant international activity with the same goal.
- Organize multidisciplinary discussion on data management and citation, with the aim of creating interoperable practices.
- Promote the use of data reference model also when referring to authors own primary source data.
- Define field specific level of granularity for data citation.

National scholarly publishers

- Present all authors with a publication specific data reference model based on the recommendations made in this roadmap and require it's use when referencing data in publications.
- Include a hyperlink, preferably the PID, to underlying data description for all original research publications.
- Create discussion about possible national applications of the FORCE 11 Roadmap for Publishers and Transparency and Openness Promotion (TOP) guidelines.

Research institutions

- Include principles of data as evidence and data transparency into enforceable institutional data policies.
- Include principles and examples of data as evidence and data transparency to research ethics MOOC and open science web course.
- Support and explore the development of data metrics in research evaluation. When implementing new metrics, pay special attention to the transparency of data and methods.
- Create and enforce institutional policies on licensing data, recommended licenses (f.e. CC-BY), and templates for data ownership agreements.
- Include addressing data authorship and ownership relevant questions to data management planning.
- Include introduction to persistent identifiers, both as a concept and a practice, into basic researcher training, preferably starting already in the methods courses for undergraduate students.

Funders, policy makers

- Make data management planning required by all research funders, either in the application stage or after funding is granted.
- Explore mechanisms for evaluating the quality of published data sets for the purpose of assessing the impact of research institutions.
- When allocating research funding, take all research outputs into considerations instead of just publications, f.e. in the vein of US National Science Foundation (NSF), that asks a principal investigator applying for funding to list his/her research "products" rather than "publications" in the biographical sketch section.
- When relevant, accept a (good quality, well described) data publication as a sole output of a research project.
- Give consistent, long-term support to data infrastructure necessary for data citation and access.

www.fcrd.fi

Foreword

Data citation is recognized as one of the make-or-break issues in open science discussions. Many international organizations and networks have addressed the topic, with the Committee on Data of the International Council for Science CODATA among the first ones with their *Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data* (Task Group on Data Citation Standards and Practices, 2013) proceeding paper.

The Finnish Committee for Research Data (FCRD) is the Finnish national member of CODATA. 'Tracing Data' is a project commissioned by the Ministry of Education and Culture Open Science and Research Initiative and executed by the FCRD. The project is tasked with *'[..] producing recommendations concerning data citation practices in the Finnish research system, by way of consulting national research community, for example learned societies and national committees of science, and taking into consideration international discussions and developments in the area of data citation (especially in the realms of ICSU and CODATA).'* (the excerpt is from the contract between IT Center for Science CSC, that coordinates the Open Science and Research initiative, and the Federation of Finnish Learned Societies, that at the time of starting the project in 2017 housed the Council of Finnish Academies, which is the umbrella organisation for national committees of international scientific unions of ICSU).

The primary aim of the project has been to define the core elements of a data reference. Broader and more far reaching recommendations have been made with the data reference information model in mind.

This roadmap is part of a broader national discussion about developing the Finnish research ecosystem in a way that makes it responsive, agile and resilient in the face of globalization, digitalization and societal grand challenges. Some call this development open science. It has also been called e-science, science 2.0 and the fourth paradigm, among other things. It could also be called good science, or just science, period.

One thing that the roadmap is consciously missing is a timeframe for implementing its recommendations. Universities Finland UNIFI is with funding from Ministry of Education and Culture in the process of developing an action plan on open science and data, that aims at recognising next steps and coordinative roles among the national field of actors. We trust that the measures put forward in this work will find their way into the conclusions of that project, due to end by May 2018. This, of course, does in no way mean that where feasible, they could not be implemented even earlier.

In addition to taking into consideration the bigger picture of national open science discussions, ensuring international interoperability is something that has been awarded extra attention throughout the Tracing Data project. FORCE11 principles for data citation (Data Citation Synthesis Group, 2014) were recognised as an essential point of reference in the project and it was decided that they would be used as a framework for the national level implementation of data citation. This decision was based both on the quality and scope of the definitions, and the level of engagement of the international research data community behind them, as it is necessary to make the Finnish solutions interoperable with the global landscape.

Other important resources for the work include the report and data from the Open Science and Research initiative open science maturity assessment for national research institutions (Ministry of Education and Culture, Open Science and Research Initiative, 2016), data policies of national research institutions, materials from the CODATA Data Citation Workshop series and outputs from several Research Data Alliance groups, especially the working group on dynamic data citation (Rauber et al., 2016).

The main output of the project is this roadmap document. It consists of two main components:

1) Recommendation for information model for data reference, to be adopted and enforced by relevant national actors, and
2) A national application of the FORCE11 data citation principles in the form of stakeholder specific recommendations derived thereof.

The project was planned, coordinated and executed by FCRD secretary Heidi Laine. The project management board consisted of the following experts:

Management Board Chair, Project Director and Coordinator Ari Asmi, University of Helsinki
Head of Research Strategic Support Ella Bingham, Aalto University
Senior Adviser Juha Hakala, National Library of Finland
Director Helena Laaksonen, Finnish Social Science Data Archive
Director Petri Myllymäki, Helsinki Institute for Information Technology
Chief Information Specialist Susanna Nykyri, Tampere University of Technology Library

In addition, FCRD chair Professor Pekka Orponen has participated in the management group meetings and supervised the work as a FCRD liaison.

Abbreviations, acronyms and key concepts used in this document

Altmetrics
Altmetrics are non-traditional metrics proposed as an alternative to more traditional citation impact metrics, such as impact factor and h-index.

Bibliometrics
Bibliometrics is statistical analysis of written publications, such as books or articles. Bibliometric methods are frequently used in the field of library and information science, including scientometrics

BioCaddie
Biomedical and health care data discovery indec ecosystem BioCaddie is a project that is developing a data discovery index (DDI) prototype which will index data that are stored elsewhere.

CC
Most often used to refer to Creative Commons copyright licenses that are suitable for public free of charge sharing of cultural artefacts. Can also refer to the American non-profit organization that has released the licenses.

CC-BY
A Creative Commons (CC) license is one of several public copyright licenses that enable the free distribution of an otherwise copyrighted work. CC-BY license allows reuse and modifications of the licensed content as long as the source is named.

CODATA
Committee on Data of the International Council for Science ICSU (soon to be International Science Council ISC after a merger with International Social Science Council ISSC)

CSC
CSC – IT Center for Science Ltd. (also known as Finnish IT center for science) is a publicly owned company that provides IT support and modeling, computing and information services for academia, research institutes and companies in Finland.

Data landing page
Recommended Finnish translation is 'kuvailusivu'.

Datametrics
Datametrics is an emerging field of infometrics that focuses on data.

DOI
Digital object identifier. One type of persistent identifier. Managed by the International DOI Foundation.

Etsin
Etsin is an online metadata catalogue that enables discovery of research datasets. It offers access to datasets in various fields via a joint metadata model.

FCRD
Finnish Committee for Research Data. National committee of CODATA, expert body focusing on research data management issues.

FORCE11
Future of Research Communications and E-scholarship. A non-profit, community-based organization, formed in 2011.

H-index
An author level metric that attempts to measure the productivity and citation impact of a scientist or scholar. The h-index can be manually determined using citation databases or using automatic tools. Each database is likely to produce a different h for the same scholar, because of different coverage.

HTTP URI
Uniform resource identifier (URI) is a string of characters used to identify a resource. Hypertext Transfer Protocol (HTTP) is the foundation of data communication for the World Wide Web. Uniform Resource Locator (URL), colloquially called a web address, is an HTTP URI.

Infometrics	Informetrics is the study of quantitative aspects of information. This includes the production, dissemination, and use of all forms of information, regardless of its form or origin. Informetrics encompasses the following fields: scientometrics, webometrics, cybermetrics and bibliometrics.
ISO	International Organization for Standardization
long tail data	The long tail of research data represents a huge number of data sets and a large diversity of data types, but we have little concrete information about the scope and characteristics of this data.
Metadata	Data about data. Descriptive metadata serves the purpose of discovery and identification, structural metadata concerns containers of dara, administrative metadata provides information that help to manage a resource.
MOOC	A massive open online course (MOOC) is an online course aimed at unlimited participation and open access via the web.
Namespace	A set of symbols that are used to organize objects of various kinds, so that these objects may be referred to by name. Namespaces are commonly structured as hierarchies to allow reuse of names in different contexts.
PID	Persistent identifier. A long-lasting reference to a document, file, web page, or other object.
RCR	Responsible conduct of research.
Research data	Data collected, observed, or created for purposes of analysis, to produce original research information and results. The definition excludes physical resources which digital research data is based on, such as physical samples. See also the definition at www.force11.org/node/4770.
SFS	Suomen strandardoimisliitto, Finnish Standards Association in English.
TENK	Tutkimuseettinen neuvottelukunta, Finnish National Board on Research Integrity.
URI	Uniform resource identifier. A string of characters used to identify a resource.
URN	Uniform resource name. One type of persistent identifier. In Finland the National Library assigns URNs.

Table of contents

Tracing Data

Data Citation Roadmap for Finland

1. Introduction

Due to the digitization of scholarly research processes and resources, sometimes referred to as the fourth paradigm of science, e-Science, Science 2.0 and / or Open Science, the research policy discussions have started to focus more and more on research data and its vast, untapped potential.

The capacity to collect and analyze multi-source data is transforming most domains of science and scholarship. However, instead of flowing freely, data is hitting walls, namely of researchers personal hard-drives. The data to answer many of humankind's most wicked challenges is already out there, and so are many technical solutions for sharing it around. Only a bridge between the two is missing. A concerted effort to manage, share, and cite data is needed to ensure that these rich resources are available to the public, to scientists working in the academic sphere and to individuals and communities who can benefit from such data.

Establishing data citation practices is a necessary measure to create a parallel to the bibliographic citation system, thus creating new incentives for data stewardship and data sharing, while also making research data more visible, accessible and exploitable, and enhancing the overall status of data as research outputs. Uniform and interoperable data citation protocols are a prerequisite for the acceptance of research data as a legitimately citable contribution to the scientific record. A functioning data citation ecosystem ensures that research results can be verified and re-purposed for future study.

Data citation metrics can be tracked, similar to publications. They have the potential to counterbalance some of the skewed incentives currently in place due to the heavy reliance on certain narrow bibliometric measures in evaluating institutions, groups and individuals alike.

2. Information model for data reference

The Tracing Data Project has listed elements that a data reference should consist of. The elements have been grouped into two categories: necessary (see Table 1) and optional (see Table 2). The order of the elements can vary according to the requirements of the publishing platform. In-text citations should follow the publishers' guidance.

A data citation is similar to literary citation, with the exception that data can be cited in data, not just text. A reference made in an article or other publication to one's own primary data can also be considered as data citation.

For the purposes of this roadmap, research data is defined as data collected, observed, or created for purposes of analysis, to produce original research information and results. The definition excludes physical resources which digital research data is based on, such as physical samples.

Table 1 Data reference information model: necessary elements

Mandatory elements	
Element	*Description*
Identifier	Persistent identifier of the data set, which provides access information (HTTP URI) to the landing page, from which users can access the relevant data, which may or may not be a part of a dynamic data set. <u>This is the single most important element of the data reference information model.</u>
Creator(s)	The person or persons / entity or entities who / which have produced the data.
Publication date / time	The date or time when the dataset has entered the repository / archive, with as much precision as is customary to the field of research in question.
Title	Name of the data set as it appears in the repository / archive. Intended to be understood foremost by humans (vs. machine readability) so should be informative but concise.
Host institution	The unique identification of the repository / archive hosting the data (e.g. "Finnish Social Science Data Archive", or by their domain "http://www.fsd.uta.fi").

Table 2 Data reference information model: optional elements

Optional elements	
Element	*Description*
Version	If a specific version or subset of the data set has been used, version/subset information should be included in the reference.
Resource type	Information about the data resource that helps human reader (as opposed to machine readability) to understand the nature and possible use constraints of the data, such as file format, computational language etc.
License status	What is the license under which the use of the dataset has been made possible.
Embargo information	Embargo is a request or requirement by a data source that the data cannot be published until a certain date or certain conditions have been met. Just like sensitive data, data that is under an embargo can be cited as long as the metadata is published openly.
ORCID	Open Researcher and Contributor ID. ORCID is a nonproprietary alphanumeric code to uniquely identify scientific authors and contributors. Issued by the nonprofit ORCID organization.

3. Data Citation Principles and Recommended Action

In this chapter we have used FORCE11 data citation principles to create a framework for evaluating the level of maturity of the Finnish research environment in the context of data citation. The evaluations are based on the expertise and experience of the project coordinator and the project management group as well as data provided by the Open Science and Research initiative. Based on this evaluation, a series of stakeholder specific recommendations for action have been made. Some, but not all, recommendations have been inspired by the FORCE 11 data citation roadmap for data repositories (Fenner et al., 2017). The stakeholder categories have been adopted and adapted from Christine Borgmans book 'Big Data, Little Data, No Data' (2014), with the addition of policy makers and general public. There are no recommendations directed at the latter mentioned group, but its inclusion was felt necessary in order to keep in mind the broader societal implications of research data management practices. The stakeholders are presented in no particular order.

Picture 1 Data citation stakeholders

3.1 Data citation stakeholders

3.1.1 Research institutions

Research institutions host researchers conducting academic research. They educate, train and employ researchers. These institutions have a pivotal role to play in making data citation practices a natural and integral part of day-to-day research activities. Additionally, research institutions are the places where research data originates and thus have power to shape data management practices through data policies and data infrastructure choices.

3.1.3 Data repositories

Research data repositories host and manage research data. Data centres, libraries and archives can all act as research data repositories. Research data repositories play a central role in the data citation ecosystem, as they provide stewardship and discovery services to find data, give persistent access to the data being cited, and provide unique identifiers and facilitate metadata creation, all essential ingredients for data citation. Repositories are data citation nodes that need to work closely with a variety of stakeholders including publishers, reference manager providers and researchers.

3.1.4 Academic publishers

In the context of this endeavour we refer to national academic publishers, because the international publishers are largely beyond national reach. National publishers have a big impact in some fields, especially in the humanities. In more internationally oriented disciplines the influence is more limited. Taking a positive and proactive stance towards data citation could enable national publishers to become best practice examples to their equivalents outside Finland and increase their appeal to potential authors.

3.1.5 Researchers

3.1.5.1 Learned societies

Learned societies, such as discipline specific societies and academies of science and letters, represent the civil society level of the research community. They are the representatives and mouthpieces of individual researchers and disciplinary cultures, from early career researchers, to senior level alike, irrespective of the seniority of their membership. They can promote positive cultural change and good practices among researchers and make sure that research policy is developed in a way that benefits the community

3.1.5.2 Individual researchers

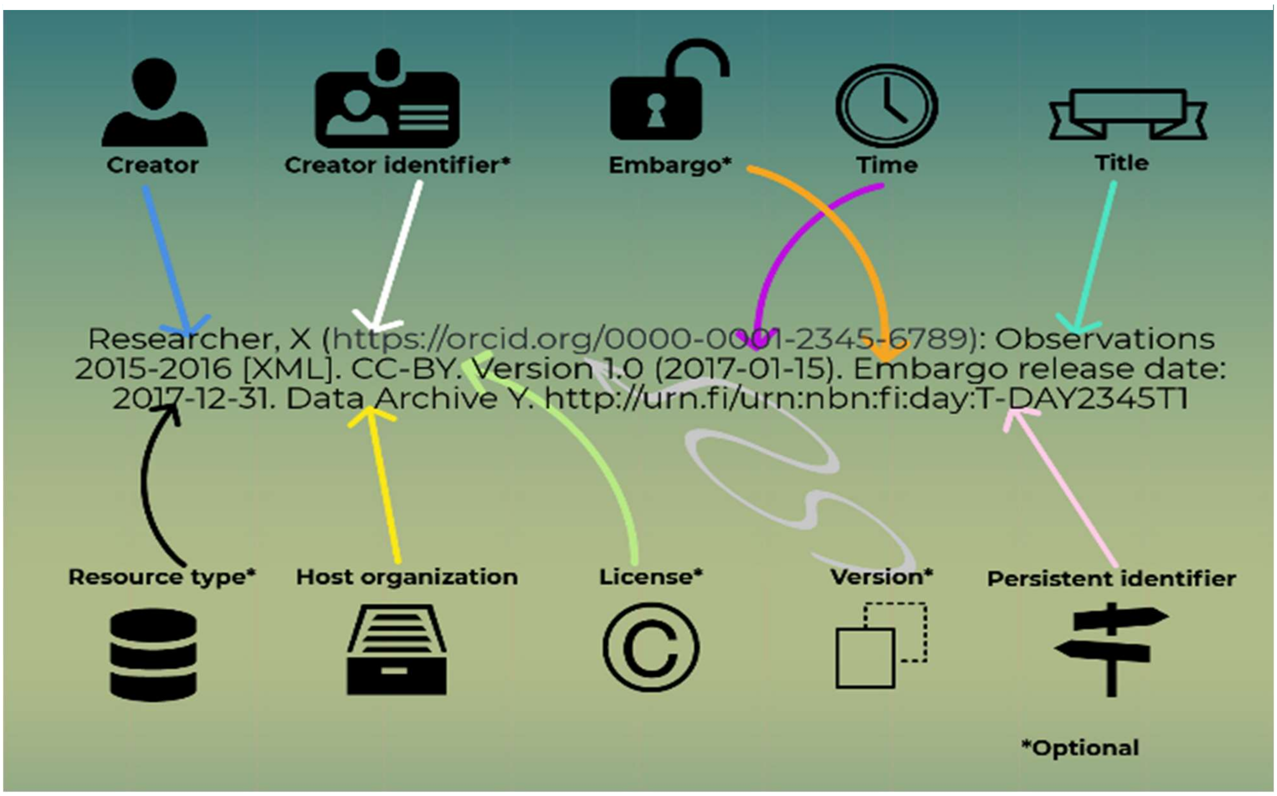
This is the most vital stakeholder group: the individuals conducting research. All of the others are facilitators. Researchers and former researchers are represented in all of the stakeholder groups, but we felt it important to highlight the impact of day-to-day habits, practices and choices of individuals. In order for data citation and related benefits to become reality, using published data and citing data needs to become as routine and mundane a part of researchers work, as creating literary references is today.

For the purposes of this project we have grouped the Finnish Advisory Board on Research Integrity (TENK) to this stakeholder category, since they coordinate a self-regulatory mechanism for promoting good research practice and eliminating misconduct. Their newly established network of research integrity advisors in research institutions is an important resource also for data citation efforts.

3.1.7 Funders and policy makers

Research funders can be private or public entities. They finance academic research, research infrastructure and supporting services. Funders have the power to change research culture and create positive incentives for

Picture 2 Data reference model



Researcher, X (https://orcid.org/0000-0001-2345-6789): Observations 2015-2016 [XML]. CC-BY. Version 1.0 (2017-01-15). Embargo release date: 2017-12-31. Data Archive Y. http://urn.fi/urn:nbn:fi:day:T-DAY2345T1 responsible data management through their funding instruments.

In terms of handling research data, funders and policy makers are the most removed stakeholder group, but at the same time have immense leverage. Funders create incentives, both carrots and sticks, on individual and institutional levels alike by valuing certain things as achievements worthy of being included in research evaluation and disregarding others. Tesearch Policy makers and politicians define on a more general level, through public budgets, what is prioritized and rewarded in the research community.

3.1.8 General public

General public are the ultimate end-users of research results and outputs. If research data is openly available, it may be used by citizens: school children and students, journalists, public officials and policymakers, jobseekers, small business owners, retired researchers, and many more. Indirectly the public benefits from added openness of science in the form of accelerated innovation and other applications of scientific results.

3.2 Target state and how to get there: evaluation and recommendations

3.2.1 Importance

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications."

National target state:

When allocating research funding or making recruitments, the evaluators and reviewers examine all relevant research outputs, not just traditional publications. Evaluators have the necessary competence for assessing the value of research data and looking beyond quantitative metrics when weighing data against publications. Discipline specific differences in levels of data intensity are taken into considerations when comparing fields and individuals alike.

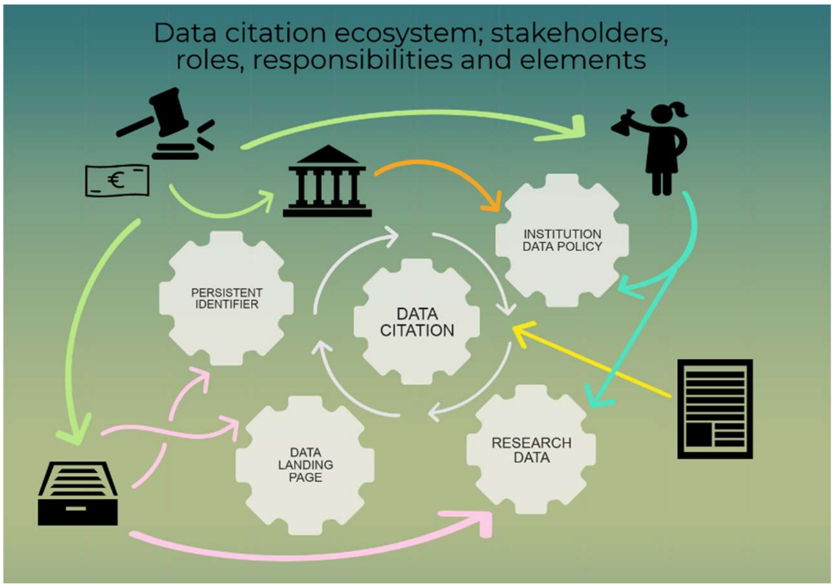
Researchers don't feel the need to 'salami slice' their results into several publications, since funders and recruiters recognise that a well described, reusable and citable data set outweighs mediocre articles in value. A traditional prose publication is no longer necessary at the end of a project, if data publication is deemed more appropriate for the results in question. However, this doesn't take away the responsibility to make the results understandable for a broad audience beyond discipline specific community.

Current national situation and readiness:

Data citations are not accorded the same importance in the scholarly record as bibliographic citations.

For example, the Finnish Publication Forum mechanism, which was created to evaluate research outputs of universities and other institutions, includes in its classification only academic journals, book series, conferences and

Picture 3 Data citation ecosystem



Picture 4 Research hall at National Archives of Finland in 1946, source: NARC

book publishers. There are some twenty to thirty data journals among those classified. None of them have been valued higher than category one (three being the highest).

Individual researchers are evaluated using such documentation as their CV's, publication history and H-index readings. The H-index measures researchers productivity in terms of publications and the citation impact of his/her publications. The Finnish advisory Board on Research Integrity CV template, that has the stated aim to provide guidelines for drafting an appropriate CV from the perspective of research ethics and integrity includes production and distribution of research data as a merit.

Key stakeholders:

Funders, policy makers, research institutions

Recommendations:

- Explore mechanisms for evaluating the quality of published data sets for the purpose of assessing the impact of research institutions.
- Support and explore the development of data metrics in research evaluation. When implementing new metrics, pay special attention to the transparency of data and methods.
- When allocating research funding, take all research outputs into consideration instead of just publications, for example in the same vein as the US National Science Foundation (NSF), that asks a principal investigator applying for funding to list their research "products" rather than "publications" in the biographical sketch section.
- When relevant, accept a (good quality, well described) data publication as a sole output of a research project.
- Update the TENK CV template to increase the visibility and prestige of research data outputs.

3.2.2 Credit and Attribution

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data."

National target situation:

All data have one or more creators or authors. An organisation is assigned the creatorship or authorship of data only in special cases, for example if the data is automatically generated. Creatorship / authorship are understood to be separate categories from data ownership. Creatorship / authorship is an inalienable right and cannot be transferred, unlike ownership.

There can also be other roles that are indicated and credited in connection to a specific data set, such as owner, curator, steward, etc. Organizations have guidelines for assigning the above-mentioned roles. Agreeing on how to assign data related credit among a research group is standard practice at the beginning of a research project.

All published data is licensed in accordance to intellectual and proprietary ownership.

Current national situation and readiness:

According to the Finnish Advisory Board on Research Integrity (TENK) authorship disputes are one of the most rapidly growing categories of causes behind allegations of research misconduct.

There are guidelines on defining authorship for publications, as well as a lively debate on who does not deserve to be named as an author, but currently no guidance on assigning data creator roles. To make things more complicated there is even a lack of understanding of and concepts for different roles related to producing data.

The University of Helsinki data policy states the following about crediting data creators:

> '*6. The University of Helsinki supports the identification and resolution of legal issues related to research data. Principal investigators are responsible for concluding contracts on the ownership and user rights of research data at as early a stage as possible or, where applicable, before the beginning of the research project.*' (University of Helsinki, 2015)

Having a discussion about how to assign credit and ownership in the beginning of a research project is certainly sound advice, but these discussions would benefit from general concepts and principles, however broad. TENK has recently published a guideline for assigning authorship in publications. After interacting with the Tracing Data Project, they are planning to include some guidance on determining data authorship. However, a NEJM opinion piece by Bierer et al. (2017) titled 'Data Authorship as an Incentive to Data Sharing' suggests that data authorship is such a complex issue, that addressing it as a side note does not sufficiently cover all of its aspects.

Key stakeholders:

Data repositories, publishers, researchers

Recommendations:

- Recognise data creatorship as a distinct issue and discussion in the TENK authorship guideline (already in progress).
- Create a multi-institutional, multi-disciplinary working group to define principles for defining data authorship, coordinated for example by TENK, or assign suitable national representation to a relevant international activity with the same goal.
- Create and enforce institutional policies on licensing data, recommended licenses (e.g. CC-BY), and templates for data ownership agreements.
- Include addressing data authorship and ownership relevant questions to data management planning.
- Present all authors with a publication specific data reference model based on the recommendations made in in this roadmap and require its use when referencing data in publications.

3.2.3 Evidence

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited."

National target situation:

All research data that is used as evidence for a published analysis is deposited in a repository for temporary, long-term or permanent preservation, unless the data is destroyed immediately after analysis for a legitimate reason. A suitable repository is chosen in accordance to relevant institutional or funder data policy.

All digitally published research results include a hyperlink to the underlying data source or to a description of the data e.g. in a metadata catalogue. The latter may apply also to data that has been destroyed. Metadata about research data may be preserved longer than the data itself.

Negligence in preserving the data and failure to make it available may be seen as research misconduct. Researchers accept and recognise that data is an essential part of their argumentation. Researchers routinely check data sources behind research results that they plan to make references to and consider results with insufficient data transparency as less reliable.

Current national situation and readiness:

It is standard practice that when a researcher makes an empirical claim they refer to the underlying evidence. However, there is currently no uniform way of making references to research data and when made, they rarely provide clear access information leading to the actual data. Finnish responsible conduct for research (RCR) guideline (TENK, 2012) does not mention data transparency or providing access to underlying evidence when making empirical claims.

The level of readiness in terms of implementing the principle of data as evidence is good. There are national level researcher skill courses, such as a MOOC (massive open online course) on research ethics and an open science web course, which in theory reach entire cohorts of PhD students.

National scholarly publishers do not currently demand data transparency from authors. Because of organization through the Federation of Finnish Learned Societies and the Finnish Association for Scholarly Publishing, platforms for discussing joint policy exist. Initiatives such as Kotilava, Journal.fi and Julkea! blog show that the field is keen on addressing challenges and creating new solutions.

Many Finnish researchers and research projects publish internationally. A number of major international publishers are involved in data citation and transparency efforts, such as the FORCE11 Data Citation Roadmap for Publishers (2017), or the TOP guidelines (Nosek et al. 2015). Some of the guidelines recommend publisher owned data repositories, which can down the road create a situation where important research data becomes proprietary, with paywalled access and restricted use by copyright.

Key stakeholders:

Researchers, publishers, research institutions

Recommendations:

- Include principles of data as evidence and data transparency in enforceable institutional data policies.
- Include principles of data as evidence and data transparency in research ethics MOOC and open science web course.
- Include principles of data as evidence and data transparency in next version of Finnish RCR guideline by TENK.
- Include a hyperlink, preferably the PID, to underlying data description for all original research publications.
- Create discussion about the possible national applications of the FORCE 11 Roadmap for Publishers and Transparency and Openness Promotion (TOP) guidelines.

3.2.4 Unique identification

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles:
"A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community."

National target state:

The persistent identifiers used in data references are actionable and allow access to the data landing page either with a click of a mouse, or by copying them to a web browser address field. Because of this ease of use, researchers routinely check the data behind research results they come across during their reading or other information gathering activities.

Data landing pages facilitate access to the actual data (files, or data which can be retrieved with a database query). Landing pages hold such information on the data, that makes its reuse uncomplicated (if the data is available for reuse), such as licensing information, rich metadata, etc. They may also contain technical metadata about the files (such as file size) and other information regarding e.g. license and ownership and history of the data.

All published data gets a permanent identifier. The process of acquiring an identifier is made simple for the researchers: it happens automatically when depositing data to a data repository. If, as an intermediate measure, the data is temporarily stored elsewhere, the researcher can acquire a PID from elsewhere (e.g. the National library). If same research data is deposited in several repositories, all of the copies get their own identifier. This is not ideal but can occur for example when researching indigenous communities outside Finland and both the researched community and researcher have a legitimate claim to the data. The different copies are named in metadata, to the extent possible.

Some data repositories only accept certain types of data. That means that data from one project can end up in different repositories, each part getting their own identifier. The different pieces are linked together in metadata records and landing page, and with the help of indexing services, such as Etsin, BioCaddie and the like.

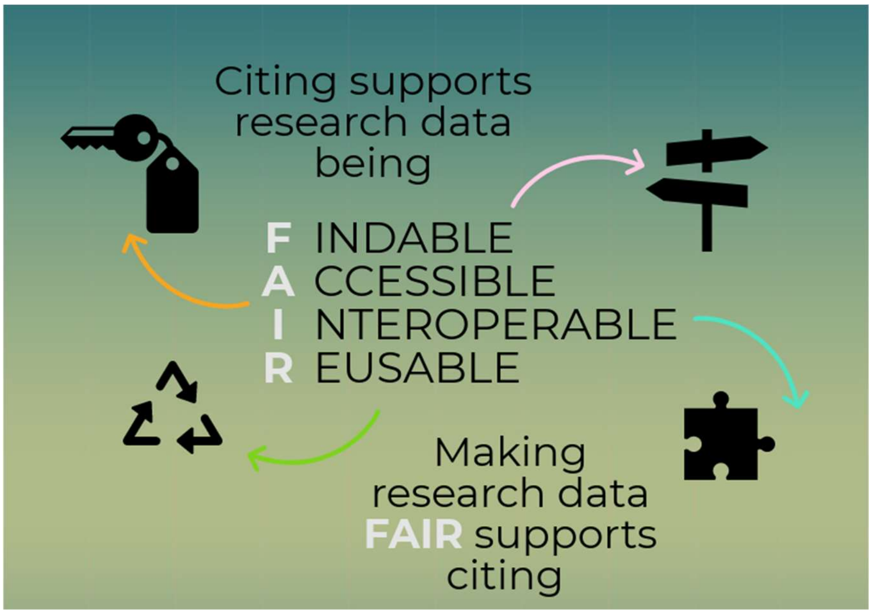
Researchers are educated to understand the importance of unique persistent identifiers. They know that the identifier is the single most important component in a data reference and use them correctly and whenever necessary.

Current national situation and readiness:

Persistent identifiers are making their way to the Finnish research data environment, as is the case internationally. The PID's in use in Finland and by Finnish researchers are uniform resource name (URN) and digital object identifier (DOI). The National Library has been assigning URNs for publications for more than 15 years, and currently they are also used for research datasets. URN system is managed by the National Library; many organizations such as CSC assign them using a namespace (akin to a family name in human names) the National Library has given them. URNs fulfill demands for persistence to the capacity of today's technology. In Finland DOIs are most often used by scholarly journals. For example, the journal management and publishing service Journal.fi uses DOIs.

It is safe to say that most of the research data originated in Finland doesn't currently receive a PID, as most of the data is not deposited in a trustworthy repository.

Picture 5 FAIR data principles


Citing supports research data being
F INDABLE
A CCESSIBLE
I NTEROPERABLE
R EUSABLE
Making research data FAIR supports citing

Readiness to implement the demand of unique identification on a national level is good, to the extent it can be technically achieved, because of the high operational level of Finnish data centers and repositories. Most likely it will be easier to get data repositories to assign PID's than it will be to get researchers to deposit their data.

Key stakeholders:

Data repositories, researchers, research institutions

Recommendations:

- All datasets intended for citation must have a globally unique persistent identifier that can be expressed as unambiguous HTTP URI.
- The persistent identifier (PID) must resolve to a landing page that supports access to the actual data set.
- Finnish data repositories should use either DOI or URN as their PID of choice, since they are the best managed and most reliable PIDs in the Finnish environment.
- Include introduction to persistent identifiers, both as a concept and a practice, into basic researcher training, preferably starting already in the methods courses for undergraduate students.

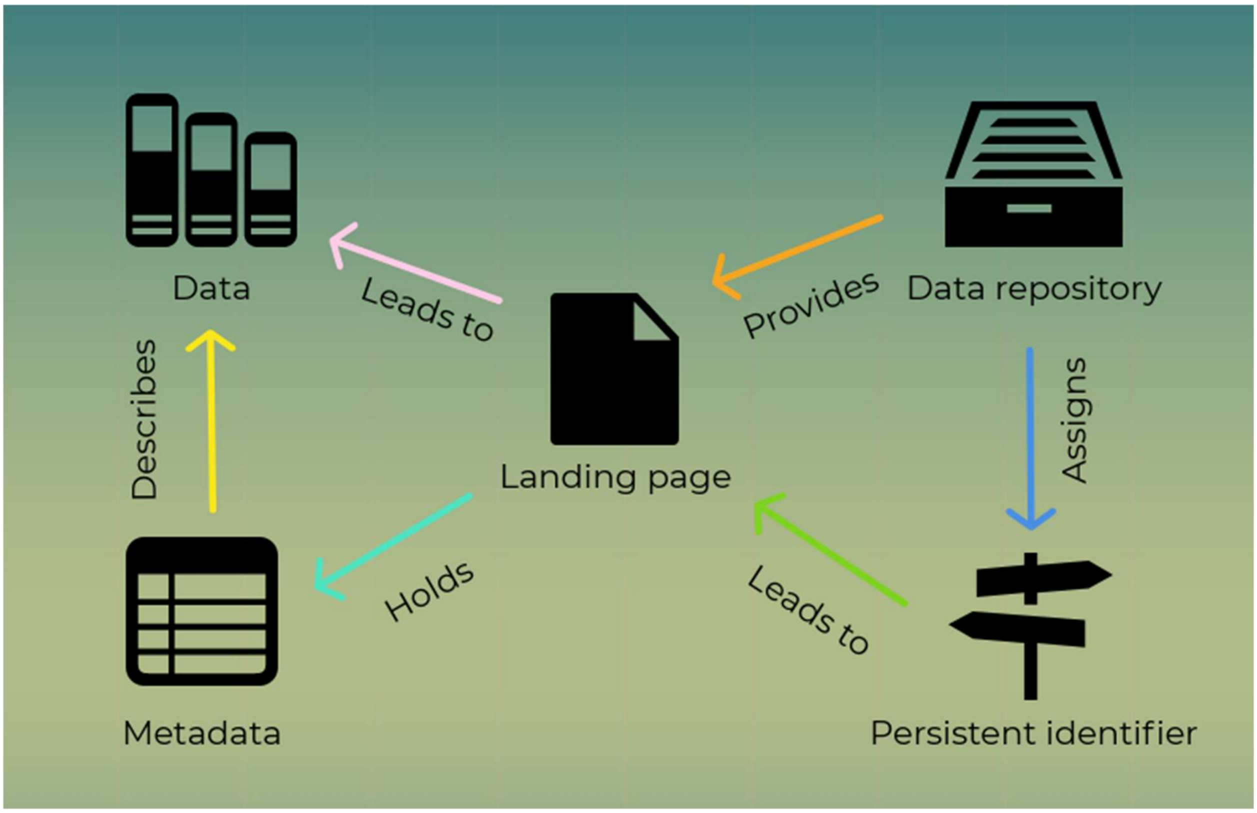
3.2.5 Access

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data."

National target state:

Every data reference includes a persistent actionable identifier. Identifier is broadly recognised as the most important element of a data reference and researchers routinely

Picture 6 Data landing page



double-check the PID's for typing errors and such, before using them in a data reference.

Data centres create a landing page for every data set with a unique PID (landing page isn't necessarily unique, but can relate to several datasets from a research project). As a link, a PID leads always to a landing page, instead of the actual data.

Current national situation and readiness:

The prerequisite to a dataset being discovered is it being described in a public online setting. There are national tools for discovering and accessing data, such as the Etsin metadata catalogue for research data in Finland. It feels safe to assume, that these services are currently underused by researchers. According to stufddies conducted among University of Helsinki researchers, more than half of researchers do not use a repository for their data and lack of sufficient metadata is more rule than exception (Salmi et al., 2016, Ala-Kyyny et al. 2018). This situation will most likely change for the better in the near future, as repositories become more and more accessible and user friendly. However, not all deposited data can be made publicly available. Access should be understood as a spectrum rather than a binary state: accessibility doesn't mean that there are no restrictions to use, such as embargos or confidentiality clauses. Even in most sensitive cases certain metadata can still be made universally accessible. It is worth noting that the FAIR data principles are mainly directed at metadata and not the actual data itself (see picture 4 and Wilkinson et al., 2016).

Key stakeholders:

Data repositories, researchers, funders

Recommendations:

- Make data management planning a requirement by all research funders, either in the application stage or after funding is granted.
- Landing page should facilitate access to metadata, either by holding metadata or a link to metadata.
- License all metadata with a CC0 license or equivalent.
- Make metadata freely harvestable through open APIs.
- The landing page should include reference model for citation and ideally also metadata helping with discovery, in human-readable and machine-readable format.
- The persistent identifier must be embedded in the landing page in machine-readable format.

3.2.6 Persistence

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe."

National target situation:

Actionable persistent identifiers operate as links to the data, taking one first to a landing page with metadata, through which the data can be accessed. The landing page is as persistent as the identifier that leads to it. If data gets relocated or destroyed, the landing page will offer status update information. If the data is deleted, rendered inaccessible or access to it is blocked for legal or other reasons, the landing page will still be available and provide status information.

Current national situation and readiness:

The persistence of a data reference depends on the platform where the reference is made: for example, journal articles have their own solutions and requirements for persistence. Data citation information isn't currently collected in any concerted fashion, so persistence is most likely at a weak level currently. Future data citation indexing mechanisms will have to address questions on persistence.

Finnish scientific and other publications which contain references to data (and other publications) are preserved by the National Library due to legal deposit (legal deposit is a legal requirement that a person or group submit copies of their publications to a repository, in the case of Finland, to the National Library). For the time being there is no legal basis for preserving either research data sets or metadata about them. In the future, legal deposit may be extended to research data as well.

Key stakeholders in Finland:

Policy makers, funders, data repositories, publishers

Recommendations:

- Data that no longer exists should still have a persistent landing page, which may direct the user to a current version of the old data set.
- Give consistent, long-term support to data infrastructure necessary for data citation and access.

3.2.7 Specificity and verifiability

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited."

National target situation:

Data references lead via persistent identifiers (PID's) to landing pages created by the data repository. The landing page facilitates access to relevant provenance information for the data set in question.

A data set gets a PID as soon as it is deposited in a repository, whether it is publicly accessible or not. When a data set becomes public at a later stage of the data life cycle its history can also be traced throughout the unpublished phase.

A uniform data reference model, applicable to a wide range of use cases, supports provenance. Whenever a researcher refers to data, whether their own or someone else's, they use the same set of information elements. This helps tracking data use and evolution throughout the lifecycle.

All national research data repositories have an open API for harvesting metadata on their content.

Research data can change over time if new records are added, errors are corrected, and obsolete records are deleted from a data set. Scholars may not use an entire data set or stream data as it is, but rather select specific subsets tailored to their research questions. In order to keep such experiments reproducible and to share and cite the particular data used in a study, researchers have means of referencing the exact query, view or morsel of a larger data set, even if the data source is continuously evolving. This applies equally to researchers utilizing so called big data and long tail data (a long tail of some distributions of numbers is the portion of the distribution having a large number of occurrences far from the "head" or central part of the distribution) alike.

Current national situation and readiness:

Creating and managing provenance data is a challenge to data repositories. Depending on the context, data provenance can either refer to the ownership history, or to a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation on how and why it got to the present state. Sometimes the latter use is understood to be part of long-term preservation metadata.

Tracking provenance and/or long-term preservation metadata for research data is vital to science and scholarship, providing answers to common questions researchers pose when sharing and exchanging data: Where did it come from? Who modified it? Is this copy the same as the copy I deposited? In what way is it the same? How do I resolve discrepancies or anomalies? Currently collecting this information is up to the researchers. Making the process of collecting provenance data fully automated looks promising, as long as data management through repositories, assigning PID's and using the data reference model is efficiently implemented. In the future all provenance metadata and information that relates to long term preservation of research data sets will be available in machine readable form, and it can be shared and re-used in other environments.

SFS 5989 (Lähde- ja tekstiviitteitä koskevat ohjeet) standard has guidelines for data citations, but they cover only static data sets. Guidelines for citing dynamic data sets have been published recently by Research Data Alliance (RDA).

There is promising international precedent for the application of the RDA data citation recommendation for dynamic data. Finland has a network of reasonably well-funded and in global comparison expertly run data centers, that have full capability to pilot and, if so decided, to implement the RDA recommendation.

For paper (plus microfilm etc.) sources the granularity of data citation is already reality, as the journal numbers (diaarinumero) exist on the level of an individual document and can be considered as persistent identifiers. In a digital environment, journal number loses its uniqueness and an additional PID is needed. The digitized resources do not currently reside in settings that would allow measures required by the RDA recommendation to be implemented.

Key stakeholders:

Data repositories, researchers, learned societies

Recommendations:

- Promote the use of data reference model also when referring to authors own primary source data.
- Assigning PIDs and creating landing pages is the responsibility of the data repository.
- Pilot the RDA Data Citation model for dynamic data in one or several national data centers.
- Define field specific level of granularity for data citation.

3.2.8 Interoperability and Flexibility

From FORCE11 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: "Data citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data citation practices across communities."

National target situation:

Field specific scholarly communities are actively engaged in national and international discussions on data management and citation practices to ensure that their unique needs and demands are recognised. There are also multidisciplinary discussion forums for comparing data practices between fields and locating common ground.

When using data citation-based metrics, different data cultures among scholarly disciplines are respected, and researchers in fields that do not create data or cannot publish it (e.g. due to sensitivity) are not disadvantaged.

Current national situation and readiness:

Current data citation principles vary. There is an international standard on information and documentation (ISO 690:2010) and a national application (SFS 5989), but they have not been effectively implemented. One of the reasons could be that the standard definitions are not open data themselves, but copyrighted content, sold for a high price as DRM protected PDF documents.

Organizations have either no data citation guidelines at all, or the guidelines differ from one organization to the next. Some of this variation is inevitable, since principles for citing are not the same in for example sciences and humanities.

The Tracing Data Project data reference information model will contribute significantly to the interoperability of data citation in Finland and beyond. The most essential element of the information model is the PID, the only machine-readable element of the proposed model. Because of the national efforts on PID administration the level of readiness for this principle is at an adequate level.

The main challenge lies with the historical archives and other paper format sources. One solution could be creating electronic PID's per every existing archival record number (diaarinumero), even if the content in case is not digitized. That would facilitate citing and transparency, if not access.

Key stakeholders in Finland:

Learned societies, scholarly publishers, data repositories

Recommendations:

- Release all data citation related content intended for broad audiences, such as guidelines and standards, in open format, i.e. CC-BY, or equivalent.
- National data centers, libraries and archives should agree on the required metadata content of a data landing page.
- Organize multidisciplinary discussion on data management and citation, with the aim of creating interoperable practices.

4. References

Ala-Kyyny, J., Korhonen, T., & Roinila, M. (2018). Tutkimusdatan avaamisen esteet: haastattelututkimus Helsingin yliopistossa. Signum, 49(4), 25–29. https://doi.org/10.25033/sig.69198

Biere, B E, Crosas M, & Pierce, H H (2017): Data Authorship as an Incentive to Data Sharing. N Engl J Med 2017; 376:1684-1687 April 27, 2017 DOI: http://doi.org/10.1056/NEJMsb1616595

Borgman, C (2015): Big Data, little data, no data. Cambridge, Massachusetts/London, England: The MIT Press.

Cousijn, H, Kenall, A, Ganley, E, Harrison, M, Kernohan, D, Murphy, F, Polischuk, P, Martone, M, Clark, T (2017): A Data Citation Roadmap for Scientific Publishers. bioRxiv 100784; DOI: https://doi.org/10.1101/100784

Data Citation Synthesis Group (2014): Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11 [/datacitation]

Fenner, M, Crosas, M, Grethe, J, Kennedy, D, Hermjakob, H, Rocca-Serra, P, Berjon, R, Karcher, S, Martone, M, Clark, T (2017): A Data Citation Roadmap for Scholarly Data Repositories. bioRxiv 097196; DOI: https://doi.org/10.1101/097196

Finnish Advisory Board on Research Integrity (2013): Responsible conduct of research and procedures for handling allegations of misconduct in Finland. Guidelines of the Finnish Advisory Board on Research Integrity 2012. Helsinki, Finland [Available at http://www.tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf]

Ministry of Education and Culture, Open Science and Research Initiative (2016): Evaluation of Openness in the Activities of Research Organisations and Research Funding Organisations in 2016. Doria 2016-11-22; URN: http://urn.fi/URN:NBN:fi-fe2016111829246

Nosek, B, Alter, G, Banks, G C, Borsboom, D, Bowman, S D, Breckler, S J, Buck, S, Chambers, C D, Chin, G, Christensen, G, Contestabile, M, Dafoe, A, Eich, E, Freese, J, Glennerster, R, Goroff, D, Green, D P, Hesse, B, Humphreys, M, Ishiyama, J, Karlan, D, Kraut, A, Lupia, A, Mabry, P, Madon, T, Malhotra, N, Mayo-Wilson, E, Mcnutt, M, Miguel, E, Levy, E, Paluck, Simonsohn, U, Soderberg, C, Spellman, B A, Turitto, J, Vandenbos, G, Vazire, S, Wagenmakers, E J, Wilson, R, Yarkoni, T: Promoting an open research culture. Science 26 Jun 2015 : 1422-1425. DOI: http://doi.org/10.1126/science.aab2374

Rauber, A, Asmi, A, van Uytvanck, D, Pröll S (2016): Identification of reproducible subsets for data citation, sharing and re-use. Bulletin of the IEEE Technical Committee on Digital Libraries [1937-7266]. 2016 vol:12 nr:1 s:6. URL: https://www.rd-alliance.org/system/files/documents/RDA-Guidelines_TCDL_draft.pdf

Salmi, Anna; Ojanen, Mikko; Kuusniemi, Mari Elisa (2016): Project MILDRED Research Data Repository Survey, University of Helsinki. figshare. https://dx.doi.org/10.6084/m9.figshare.3806394.v4
Retrieved: 18:08, April 5, 2018 (EET)

Task Group on Data Citation Standards and Practices (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 12, pp.CIDCR1–CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043

Wilkinson, M. D. *et al.* (2016): The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*3:160018
DOI: http://doi.org/10.1038/sdata.2016.18