TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

FAIR to Care

Repository landscaping and support for FAIR in Europe

Tuomas J. Alaterä, FSD

IASSIST 2023, June 1, 2023

Presentation outline

1. Project descriptions: SSHOC and EOSC-Nordic
2. Desk research on repositories: setting and results
3. FAIR support and evaluation
4. Data on minorities
5. Conclusions

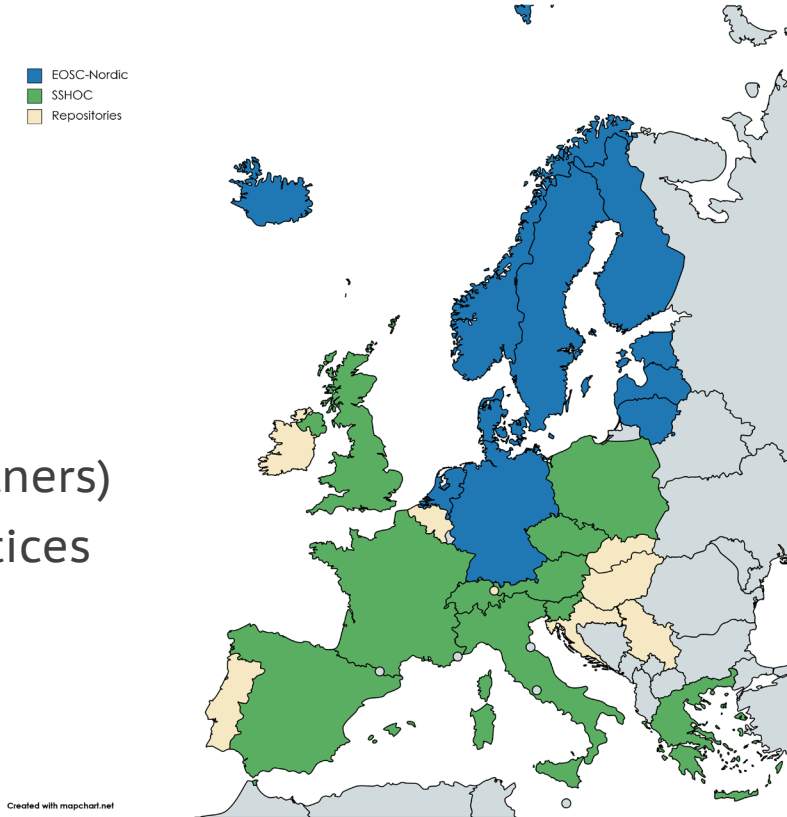
TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Project Descriptions: SSHOC and EOSC-Nordic

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

SSHOC and EOSC-Nordic

- EU funded* infrastructure projects with research / research infrastructure support functions and European Open Science Cloud connection
- A subtask in both to support repository certification and examine the landscape
- SSHOC
  - Discipline specific, Social Sciences and Humanities data
  - Targeted the whole EU/ERA (20 partners and 27 associates)
    - Research infrastructures and individual organisations as participants
  - Cloud based solutions for researchers and curators, training
- EOSC-Nordic
  - Regional, all disciplines (participants from 10 countries, 25 partners)
  - Focus on openness of research data, synergies in policies, practices
  - Strong focus on FAIR uptake and measuring FAIR maturity



*Horizon 2020: the EU Framework Programme for Research and Innovation

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Desk Research on Repositories: Setting and Results

The study and the objectives set

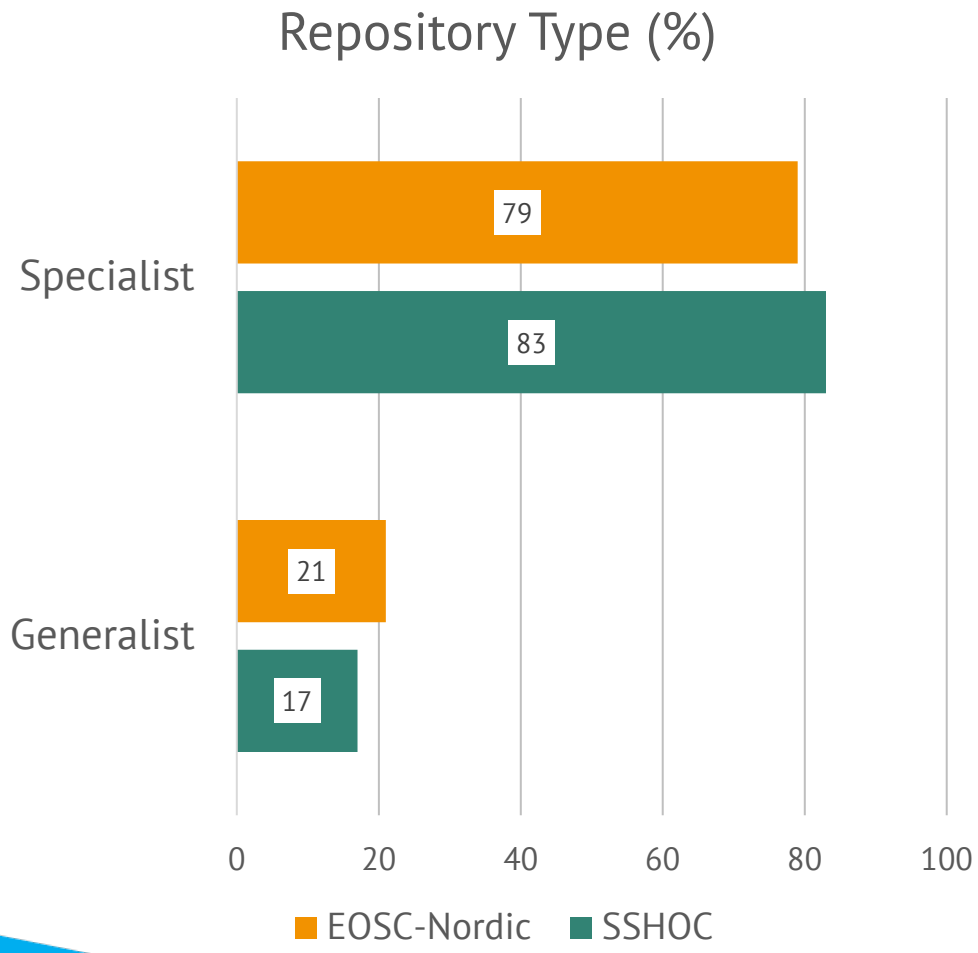
- Gain insight of the basic characteristics of data repositories in Europe and the Nordic and Baltic countries
- Gain insight into the differences and commonalities to better support repositories in FAIRification and certification
- Conducted by accessing the publicly available information on the repository websites in 2021-2022 = *"curious and persistent user"* approach
- EOSC-Nordic sample: 86 repositories studied
  - Further 12 excluded due to website being under construction, access to metadata and other information behind a login or the entity not being a data repository
- SSHOC sample: 93 repositories studied
  - Further 41 excluded as not "organisations that preserve, manage, and provide access to digital research data in a variety of formats"
- Included repositories from 27 countries, some overlap in samples

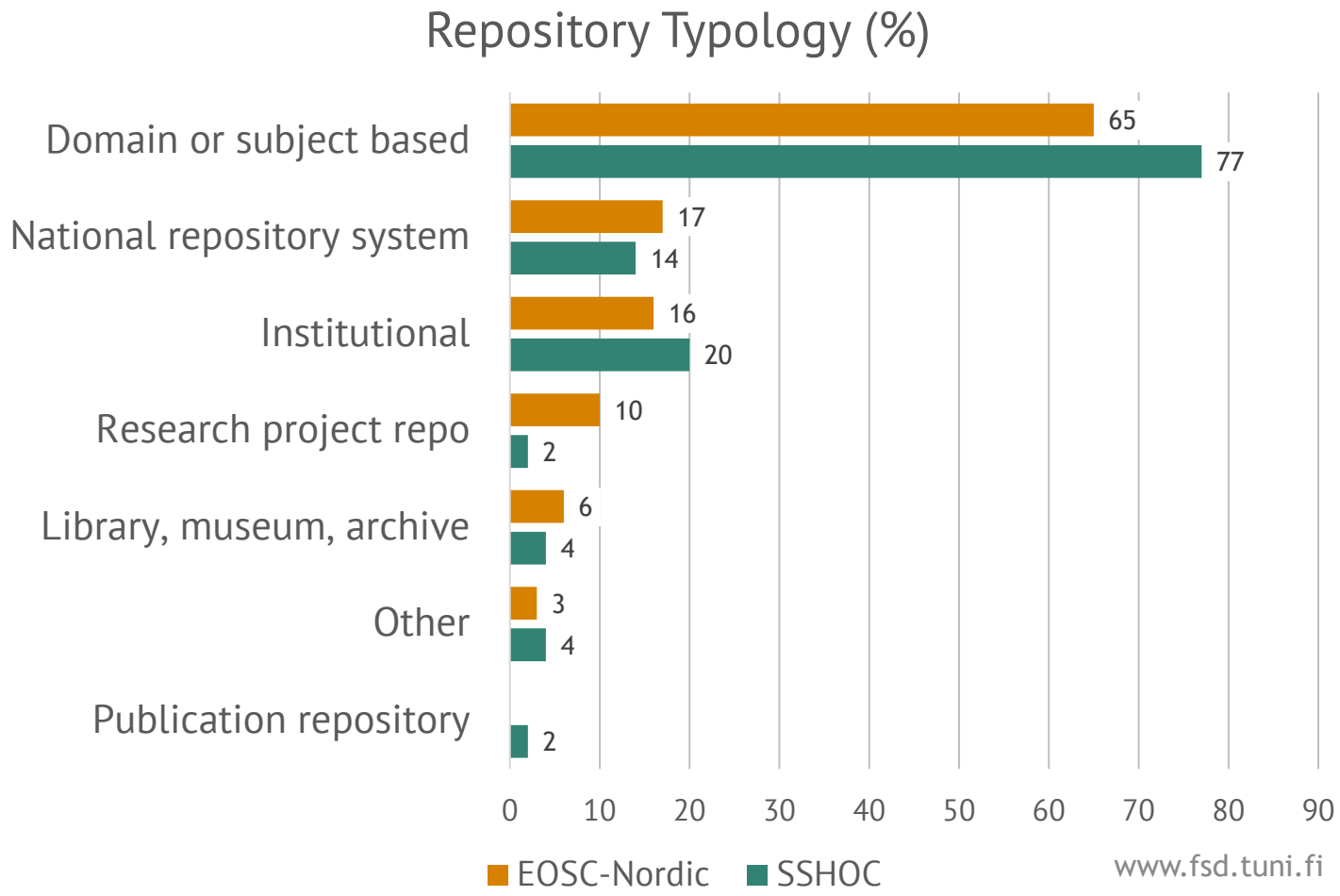
Why? No FAIR in a vacuum

- Data users need to know what the data are about, who the providers are, and how the data can be accessed and used
- To make and keep data FAIR (=repository function), certain basic details must be available to data users / stakeholders
- Same basic needs present in repository certification
- ¾ of repositories in the combined sample were hosted by a larger organisation
  - Usually a university: 43 % in SSHOC, 52 % EOSC-Nordic
  - Dependency on the host is common, and can cause confusion for the user looking for information
- First step: categorise the repositories…

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Repository types and typologies





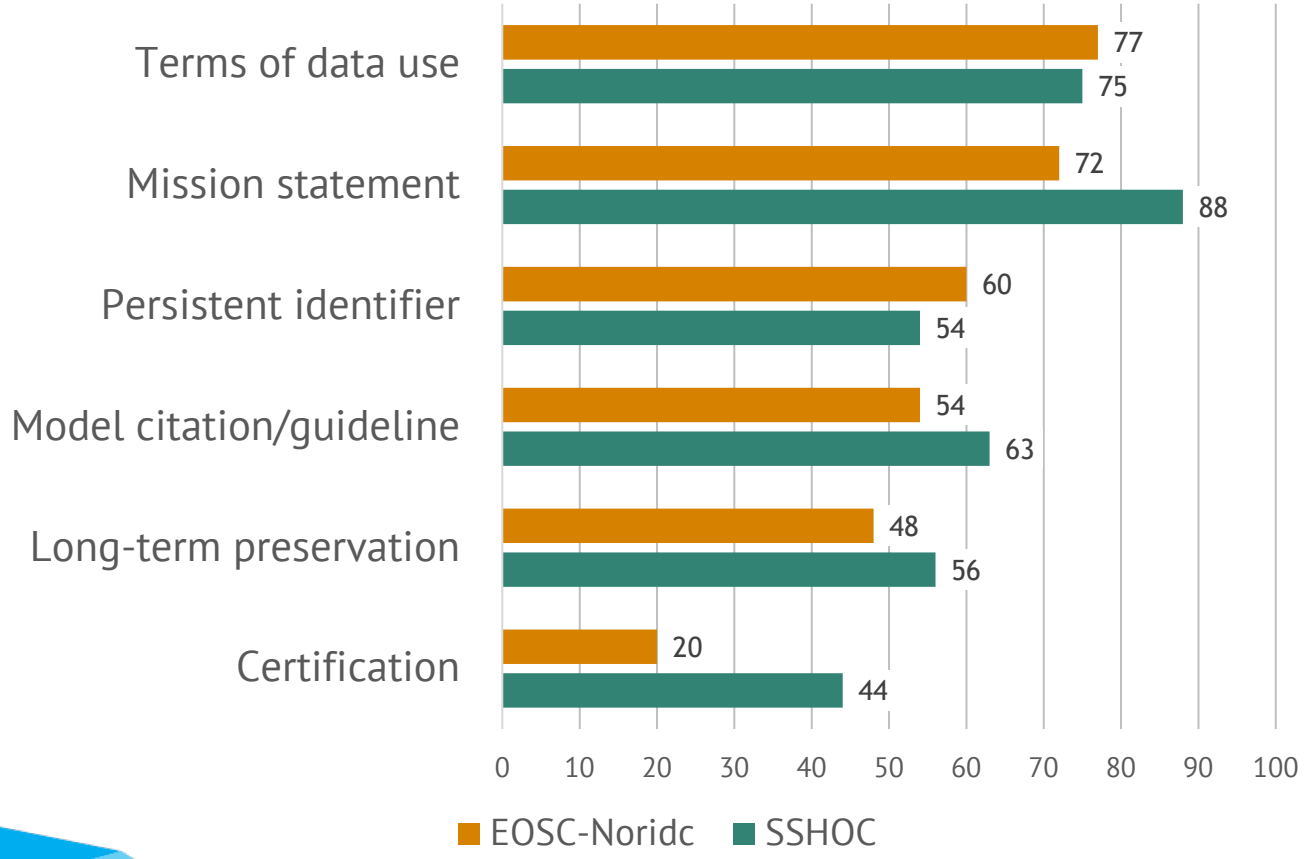
Basic information looked for

- **Mission statement** – does the repository have a mission in data curation?
- **Long-term preservation promise** – data will remain (re)usable in the future too
- **Terms and conditions** for data use – how to access and use
- **Model citation** – how to cite – for men and machines
- **Persistent identifiers** - data will remain findable / accessible
- **Certification** – audited against a defined criteria = quality assurance
- **Organisational identifier** – findability and organisational persistence
- **Designated user community**
  - Almost always findable, but often only in mission statements or organizational descriptions. In EOSC-Nordic sample only about 45% had a clear definition.

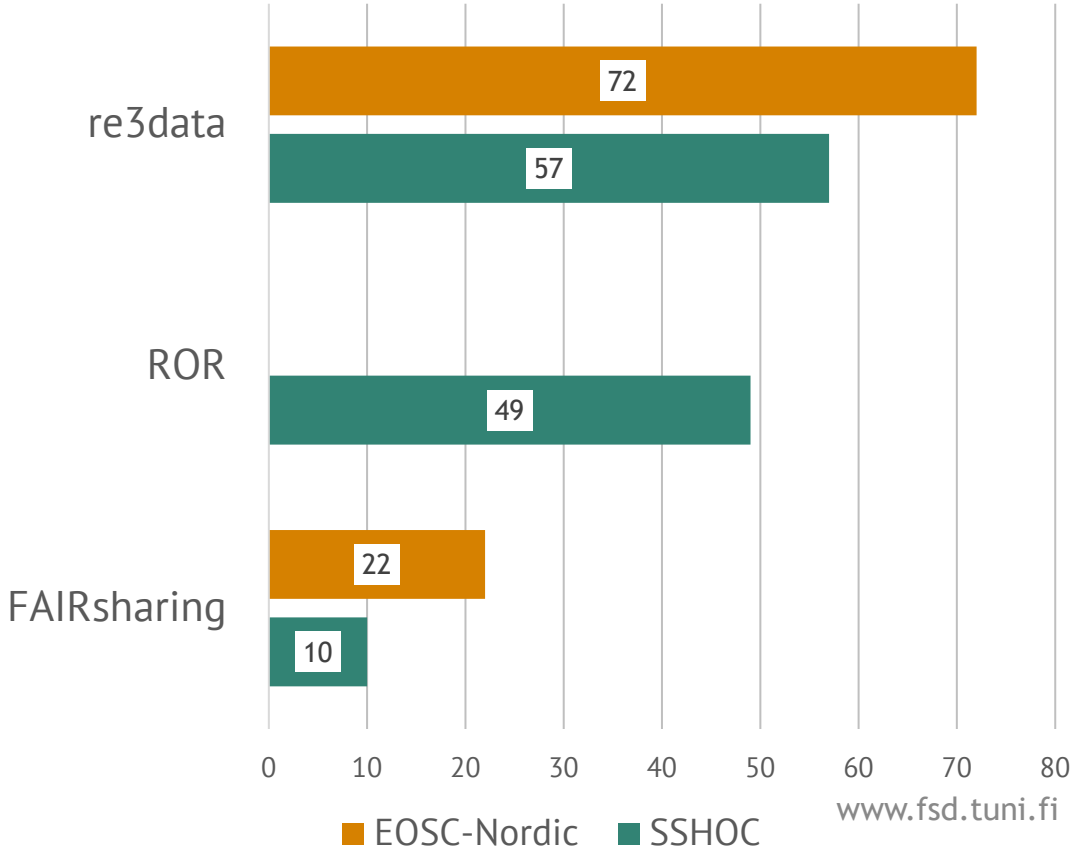
TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Basic information provided by repositories

Information provided by the repositories (%)



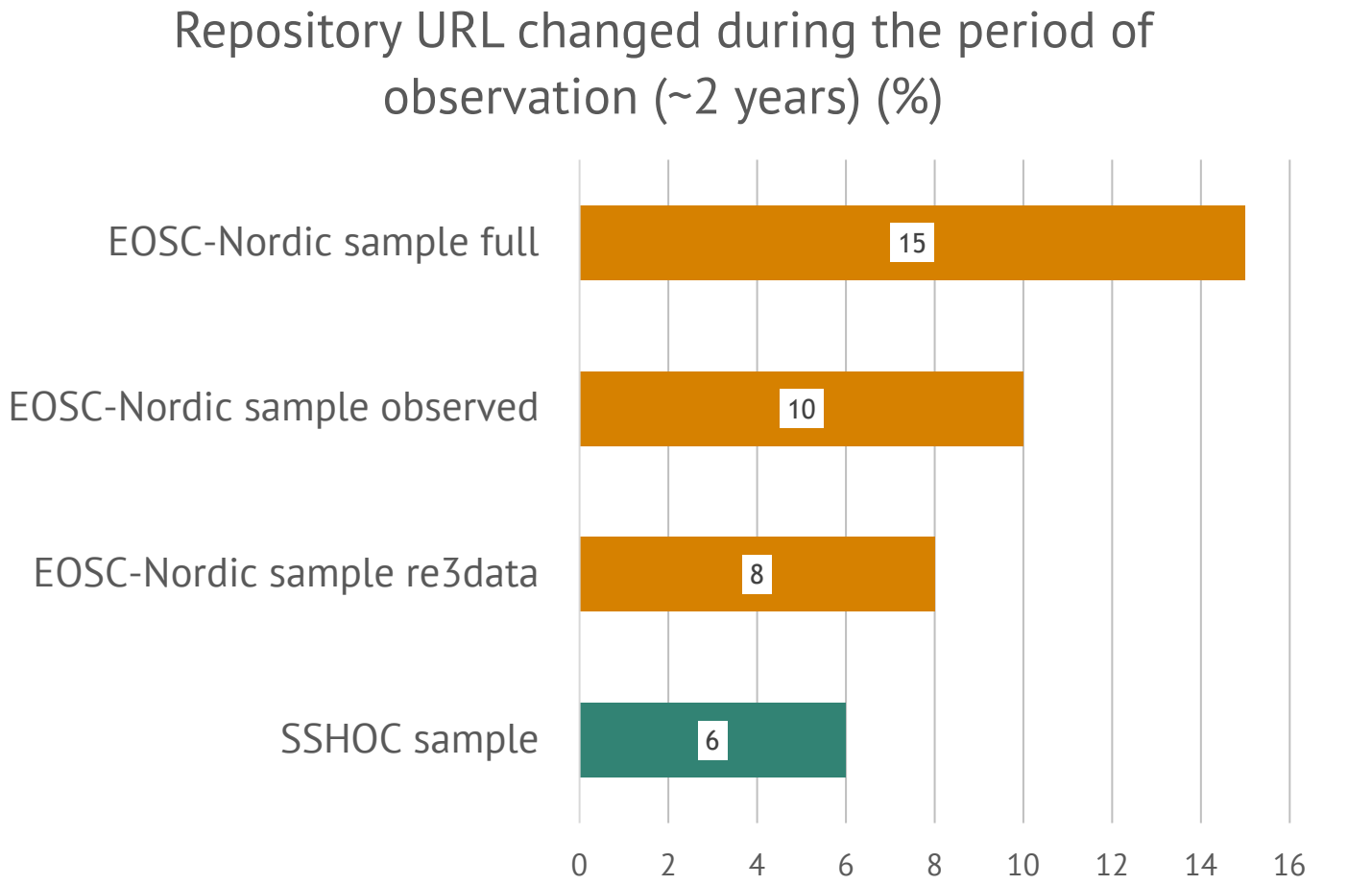
Record in a repository registry (%)



TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Repository persistence

- Tested slightly differently in SSHOC and EOSC-Nordic
  - SSHOC looked at in URL persistence, and name changes
  - EOSC-Nordic used an automated test of URLs
- Not all changes were reflected in research registries



TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

FAIR Support and Evaluation

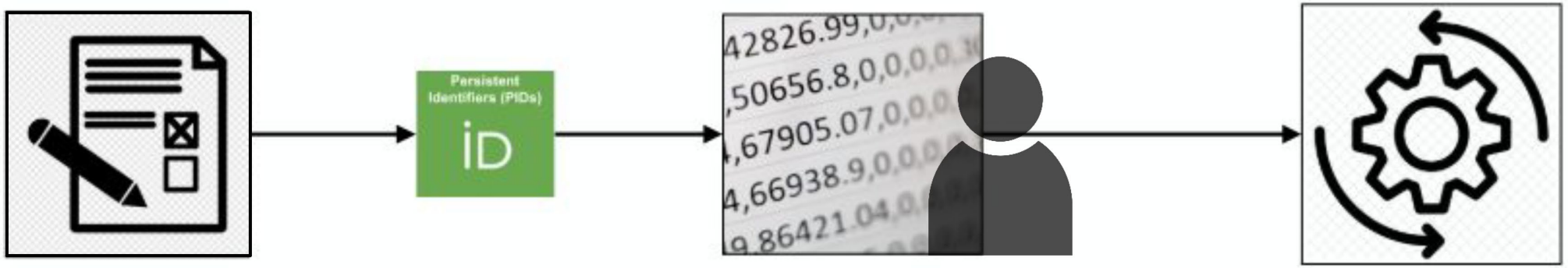
TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Automated measurement of FAIR (EOSC-Nordic)

- Measuring FAIR defined:
  - an ability to process machine-actionable metadata from data catalogues
  - and pass tests created for a FAIR evaluator
- Extremely practical and direct approach
- Instead of running hundreds of the tests manually, an automated test approach developed
- FAIR score is defined as an average of test results for Findability, Accessibility, Interoperability and Reusability

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Automated FAIR maturity assessment



Sample of 98 repositories in the Nordic countries and the Baltics

Almost 25% cannot be evaluated because they lack a GUID

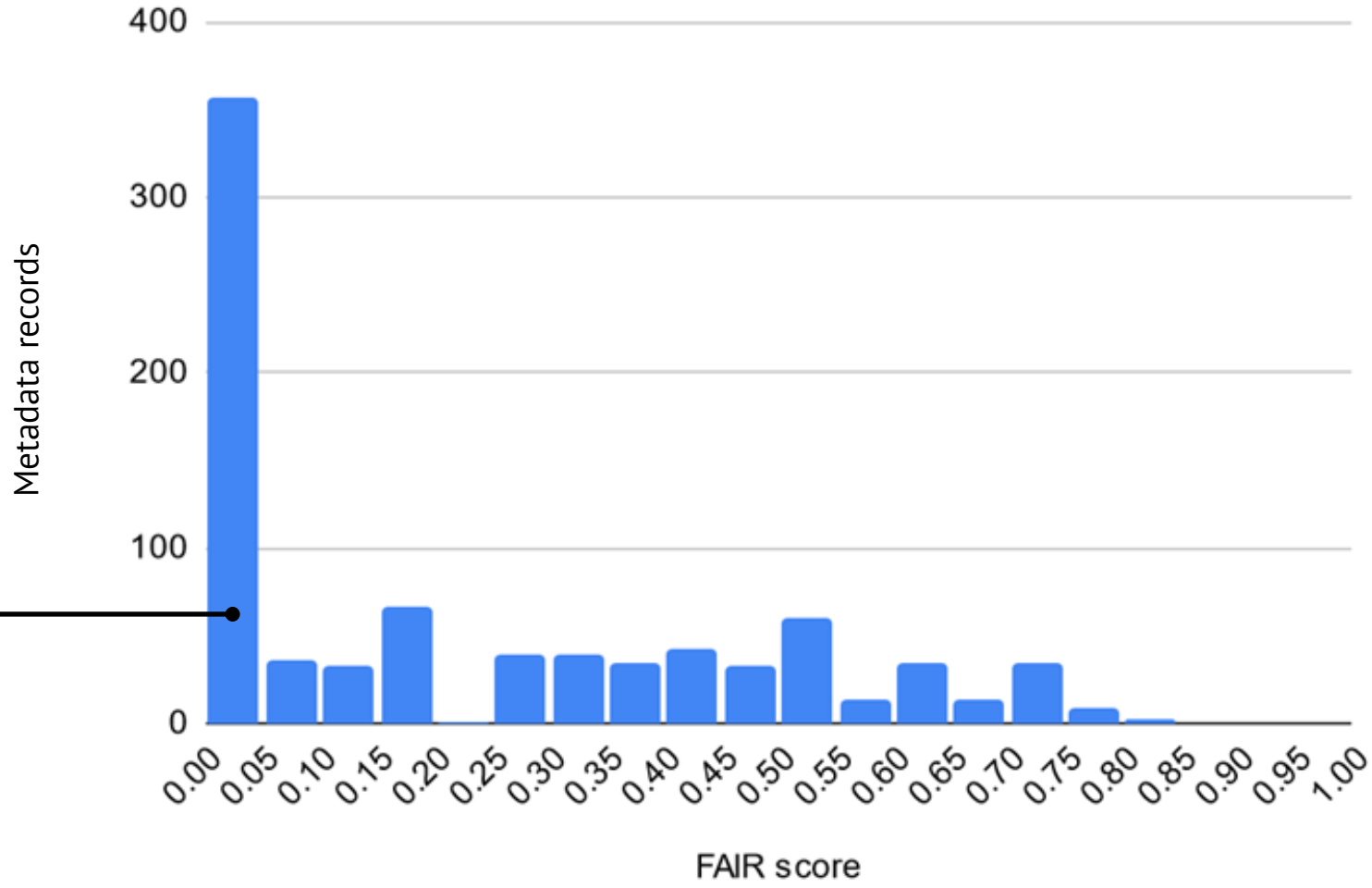
Expert selection of 10 different datasets (metadata records) from each repository

Iterative evaluations during the project period using a standalone F-UJI tool instance

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

FAIR Scores

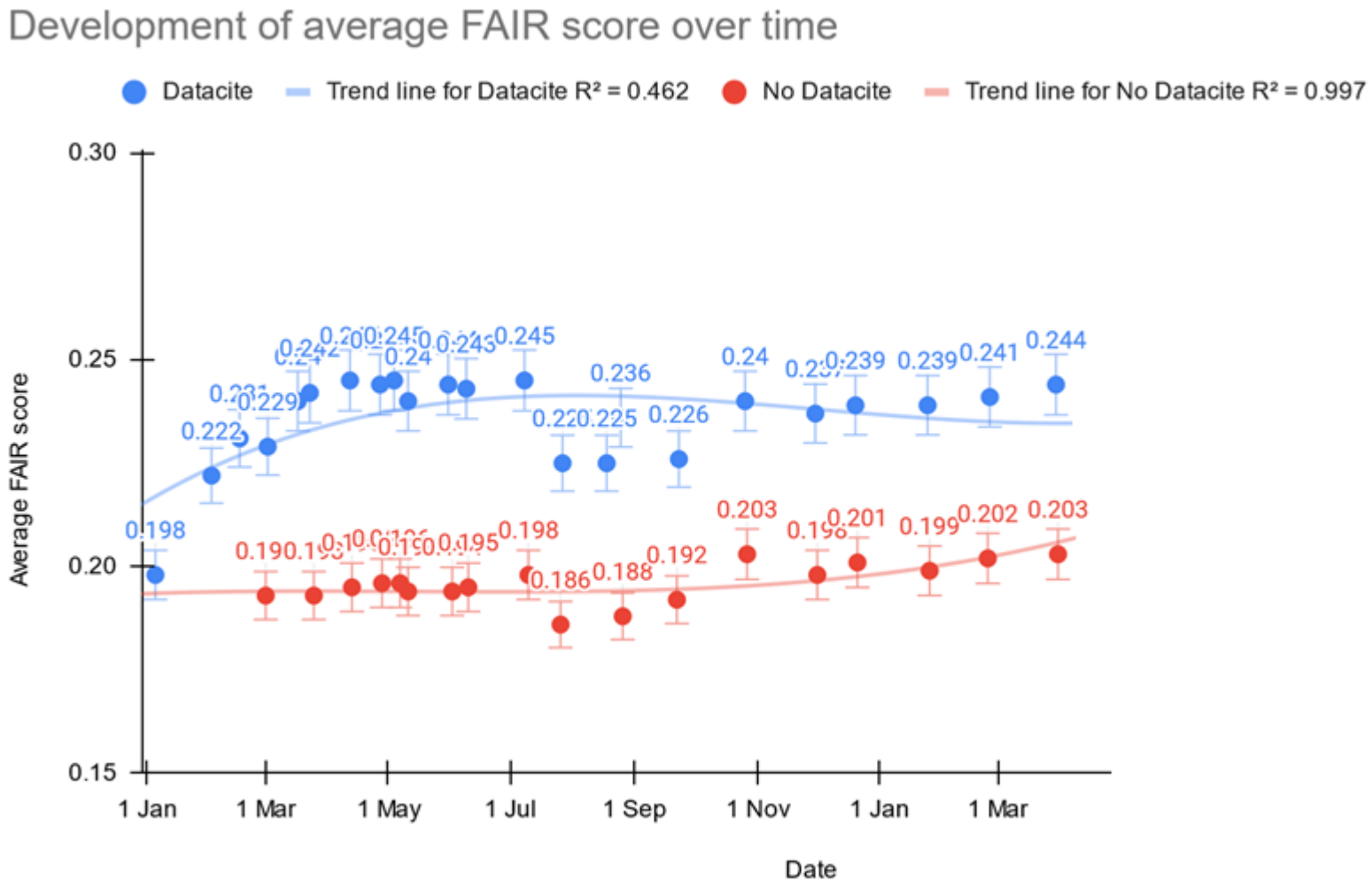
- Overall, FAIR maturity level still relatively low
- Lots of room for improvement, especially in readiness to be evaluated at all
- 0,00 result only possible if evaluation fails



TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

FAIR Scores over time

- Slight increase in averages over time
- Within the sample remarkable single improvements (e.g. ~.2 → .7)
- Test failures excluded from the averages
- DataCite metadata through DOIs has a positive effect on the score



TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Study results explaining the FAIR scores?

FAIR score*	PIDs not in use	PIDs in use
No score	20,0	3,5
Low	8,2	9,4
Medium	9,4	32,9
High	1,2	15,3

FAIR score*	Not certified %	Certified %
No score	20,9	2,3
Low	14,0	3,5
Medium	31,4	11,6
High	9,3	7,0

FAIR score*	No long-term preservation	Long-term preservation mission
No score	15,1	8,1
Low	10,5	7,0
Medium	18,6	24,4
High	8,1	8,1

*Using 2022/3 FAIR scores, low is below .2 and high above .5

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Data on Minorities

Minority data and study results

- Information whether data on minorities were provided by a repository does not really show up in aggregate results
- It is largely spread throughout the (SSH) repositories or in a very few specific repositories
  - Latter are often small and struggle with similar issues as any other small repository
  - Some data are restricted (explored only onsite), some are very open because they are seen as cultural heritage data
  - Any repository quality stamps applied to minority data too – but is there expertise specific to those data?
- Not satisfied with this result, should be looked in more detail

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Conclusions

Conclusions 1/2

- Repositories are not alike. (And what is a repository..?)
  - Results should not be generalised too heavily (purposive sampling)
- Information is relatively well available for all stakeholders, but there could/should be even more
- Use of PIDs is common but not as high as it should be for FAIR
- FAIR measurement can be automated, but the results should be interpreted with caution
- Ticking many boxes in the study did not necessarily lead to better results in the FAIR evaluations

Conclusions 2/2

- Machine-actionable metadata or globally unique identifiers should be used more widely
  - Persistent identifiers, rich generic and discipline-specific metadata, machine-actionable licenses, and controlled vocabularies expressed in some form of linked open data will quickly increase the FAIR score
- Level of certification relatively low…
  - But how high it should be? Depends on the expectations.
- Funding and coordination of national, regional and international initiatives is essential
  - Repository persistence fluctuates and broader shoulders or sharing the burden might help

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

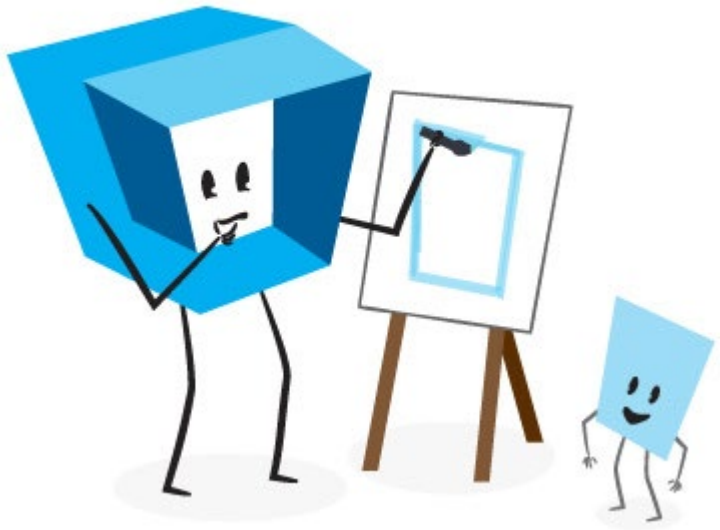
Read more

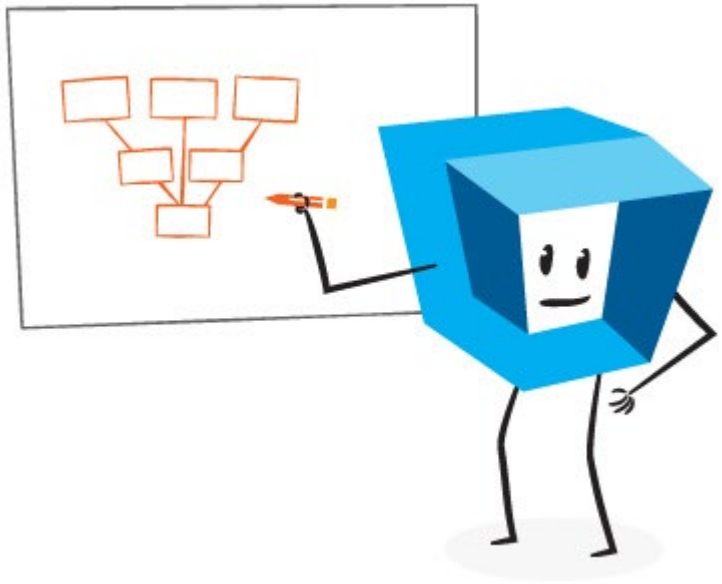
- Alaterä, Tuomas J., Kleemola, Mari, Ala-Lahti, Henri, & Jerlehag, Birger. (2022). D4.5 Report on completed FAIR data standard adoption and certifications of data repositories in the region. https://doi.org/10.5281/zenodo.7303538
- Ala-Lahti, Henri, Mathers, Benjamin Jacob, L'Hours, Hervé, Kleemola, Mari, & Alaterä, Tuomas J. (2022). Data Repositories and Certification in a Diverse Trust Landscape: Results of SSHOC T8.2 Desk Research (v1.0). https://doi.org/10.5281/zenodo.6334025
- Ala-Lahti, Henri, Mathers, Benjamin Jacob, L'Hours, Hervé, Kleemola, Mari, & Alaterä, Tuomas J. (2022). Repositories and Beyond: Analysis of Survey for SSHOC Organisations (v1.0). https://doi.org/10.5281/zenodo.6325149
- Nordling, Josefine, Mihai, Hannah, Meerman, Bert, Alaterä, Tuomas J., Kleemola, Mari, & Livenson, Ilja. (2022). D4.3 Report on Nordic and Baltic repositories and their uptake of FAIR. https://doi.org/10.5281/zenodo.6880904

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

And with that I'm finished

https://www.fsd.tuni.fi/en/





Tuomas J. Alaterä
tuomas.alatera@tuni.fi
ORCID: 0000-0002-3448-3448