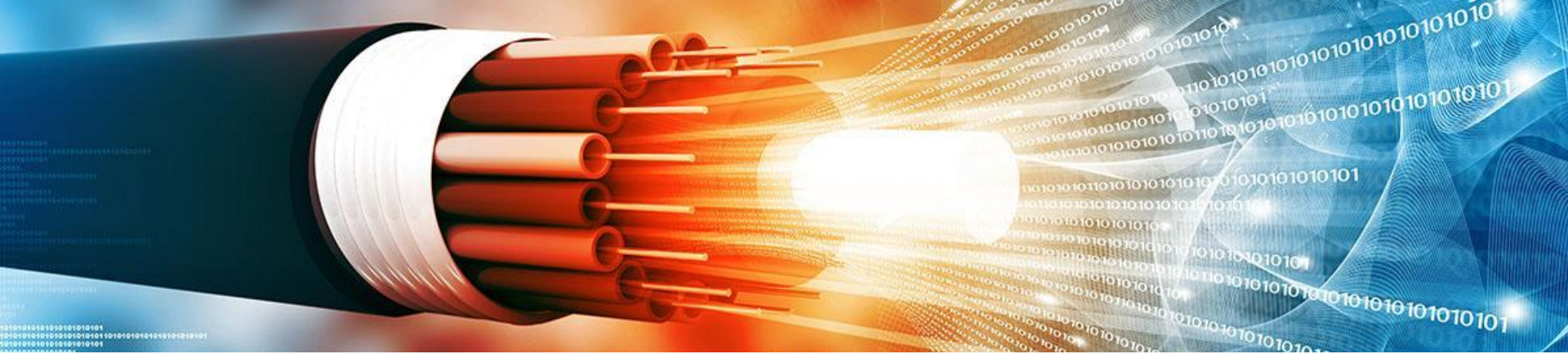


EOSC NORDIC

Real world experience with evaluating repositories

February 8th, 2022

Hannah Mihai, DeiC
Tuomas J. Alaterä, FSD

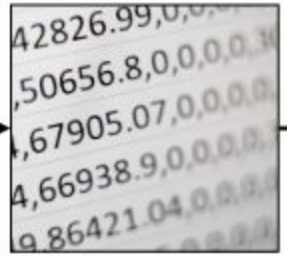
EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

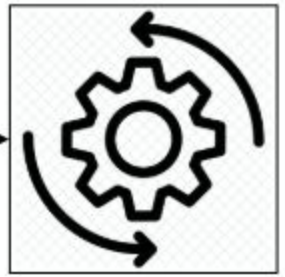
neic
NORDIC E-INFRASTRUCTURE COLLABORATION

WP4: FAIR maturity of Nordic and Baltic Repositories

WP4 activities

Excluded repositories that don't assign a GUID to each individual dataset



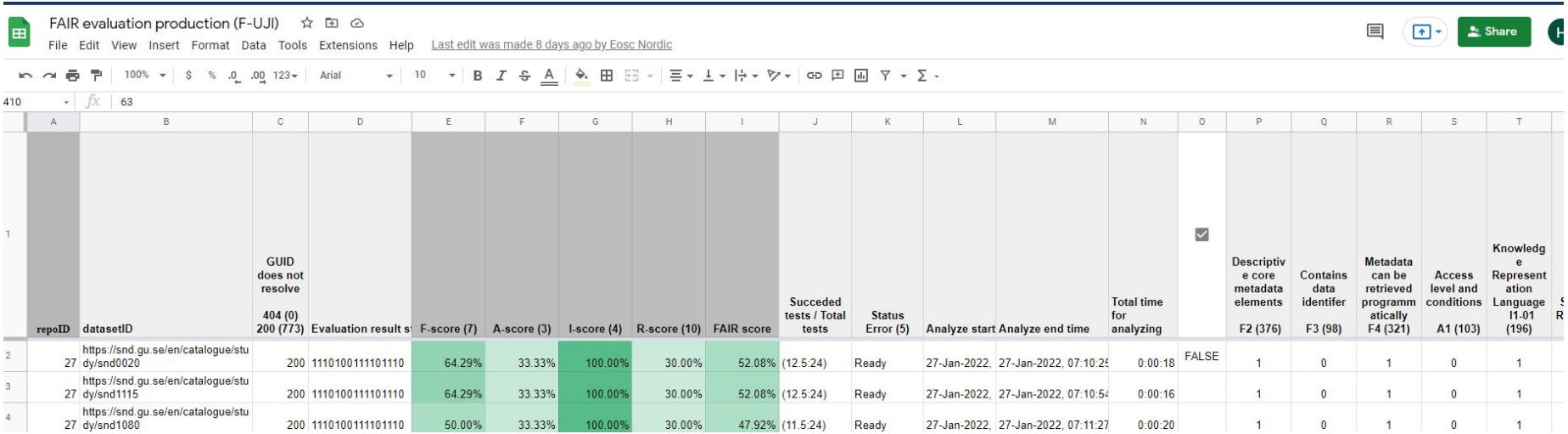


Surveyed the Nordics and Baltics for 100(+) research repositories

Evaluation of 10 datasets per repository

Monthly automated evaluations of these datasets, in the beginning with the Wilkinson tool, now with the F-UJI tool (FAIRsFAIR)

The workflow



FAIR evaluation production (F-UJI)

File Edit View Insert Format Data Tools Extensions Help Last edit was made 8 days ago by Eosc Nordic

Share

410 | fx | 63

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	repoID	datasetID	GUID does not resolve 404 (0) 200 (773)	Evaluation result s	F-score (7)	A-score (3)	I-score (4)	R-score (10)	FAIR score	Succeded tests / Total tests	Status Error (5)	Analyze start	Analyze end time	Total time for analyzing	☑	Descriptive core metadata elements F2 (376)	Contains data identifer F3 (98)	Metadata can be retrieved programmatically F4 (321)	Access level and conditions A1 (103)	Knowledge Representation Language I1-01 (196)	
2	27	https://snd.gu.se/en/catalogue/study/snd0020	200	1110100111101110	64.29%	33.33%	100.00%	30.00%	52.08%	(12.5:24)	Ready	27-Jan-2022,	27-Jan-2022, 07:10:25	0:00:18	FALSE	1	0	1	0	1	
3	27	https://snd.gu.se/en/catalogue/study/snd1115	200	1110100111101110	64.29%	33.33%	100.00%	30.00%	52.08%	(12.5:24)	Ready	27-Jan-2022,	27-Jan-2022, 07:10:54	0:00:16		1	0	1	0	1	
4	27	https://snd.gu.se/en/catalogue/study/snd1080	200	1110100111101110	50.00%	33.33%	100.00%	30.00%	47.92%	(11.5:24)	Ready	27-Jan-2022,	27-Jan-2022, 07:11:27	0:00:20		1	0	1	0	1	

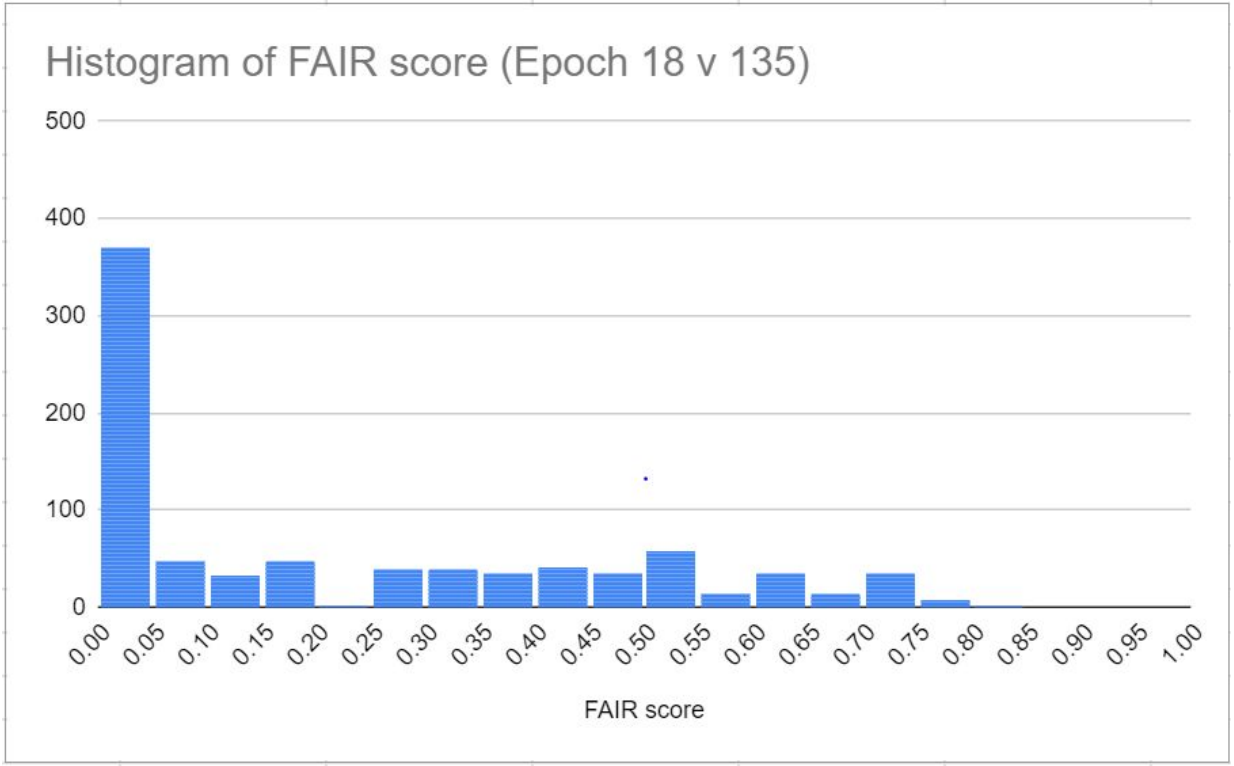
- Analysis is getting started in GoogleSheets (2 modes, with and without DataCite metadata)
- GoogleScripts run in the background
- One analysis takes around 20 seconds, for ca. 800 datasets it takes 4-5 hours

The workflow

	38%	41%	37%	23%						
	(73)	(45)	(37)	(47)		72	72	72	72	72
	0.380	0.251	0.185	0.151	0.236	0.007	0.018	0.013	0.014	0.010
850	(73)	(73)	(73)	(73)						
								74%	<33%	
MAX	100.0%	100.0%	75.0%	0.5(3)	70.8%			15%	33%<X<50%	
MIN	14.3%	0.0%	0.0%	0.0%	4.2%			13%	>50%	

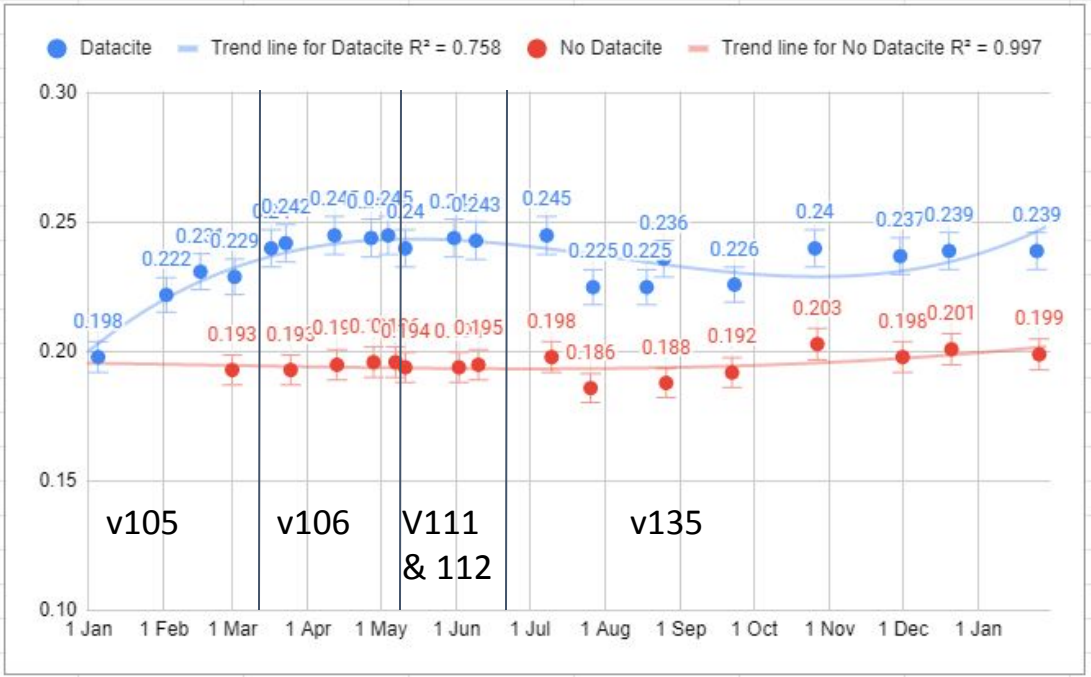
- Summary for the entire sample is generated automatically
- More data-analysis needs manual work

Histogram of FAIR scores of all evaluated repositories*



*(incl. DataCite metadata)

Preliminary results



- DataCite metadata gives added FAIR-value
- Especially I and R scores are affected
- general (slight) increase over time
- Affected by change of version in F-UJI

Experiences from FSD

Finnish Social Science Data Archive

- 20+ years of experience in data archiving and opening research data
  - Strong in metadata (production, development and reuse)

- Project partner in EOSC-Nordic
  - Support provider for selected repositories

- At the same time one of the evaluated repositories

   Internal goal: be FAIR and perform well in the evaluation

Evaluation stages

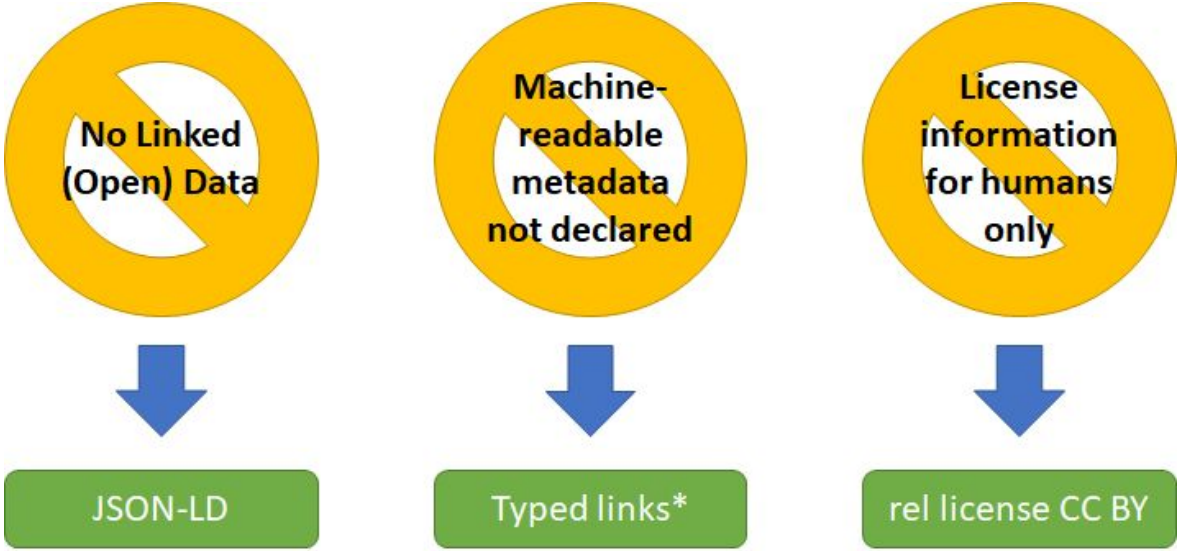
Already familiar with the 16 requirements. Preliminary (human) assessment done. Expectation of being FAIR, interoperability perhaps an issue.

Initial automated evaluation: poor results almost in all tests. Evident that actions are needed.

Closer examination of the results a relief: evaluation failures not fundamental but require rethinking. Focus on the benefits.

Weaknesses identified. Evaluation result greatly improved with justified changes to metadata. Work in progress…

Key findings



DOI vs URN? At first seemed like a battle lost

Lessons learnt

DO

- focus on metadata
  - Metadata available, but only some datasets can be downloaded without registration → FAIRness of metadata is crucial
- take basic steps:
  - Embedded JSON → multilingual and vocabulary based
  - Enriched DublinCore
  - Typed links / signposting
  - Vocabularies, ontologies, keywords, mappings…
- produce valid metadata

DON'T

- do it for the evaluator
- worry if not reaching 10/10 : understand the results and limitations
- think FAIR only now. Keeping data FAIR needs to be addressed