Enabling FAIR Data Principles in the Era of Big Data

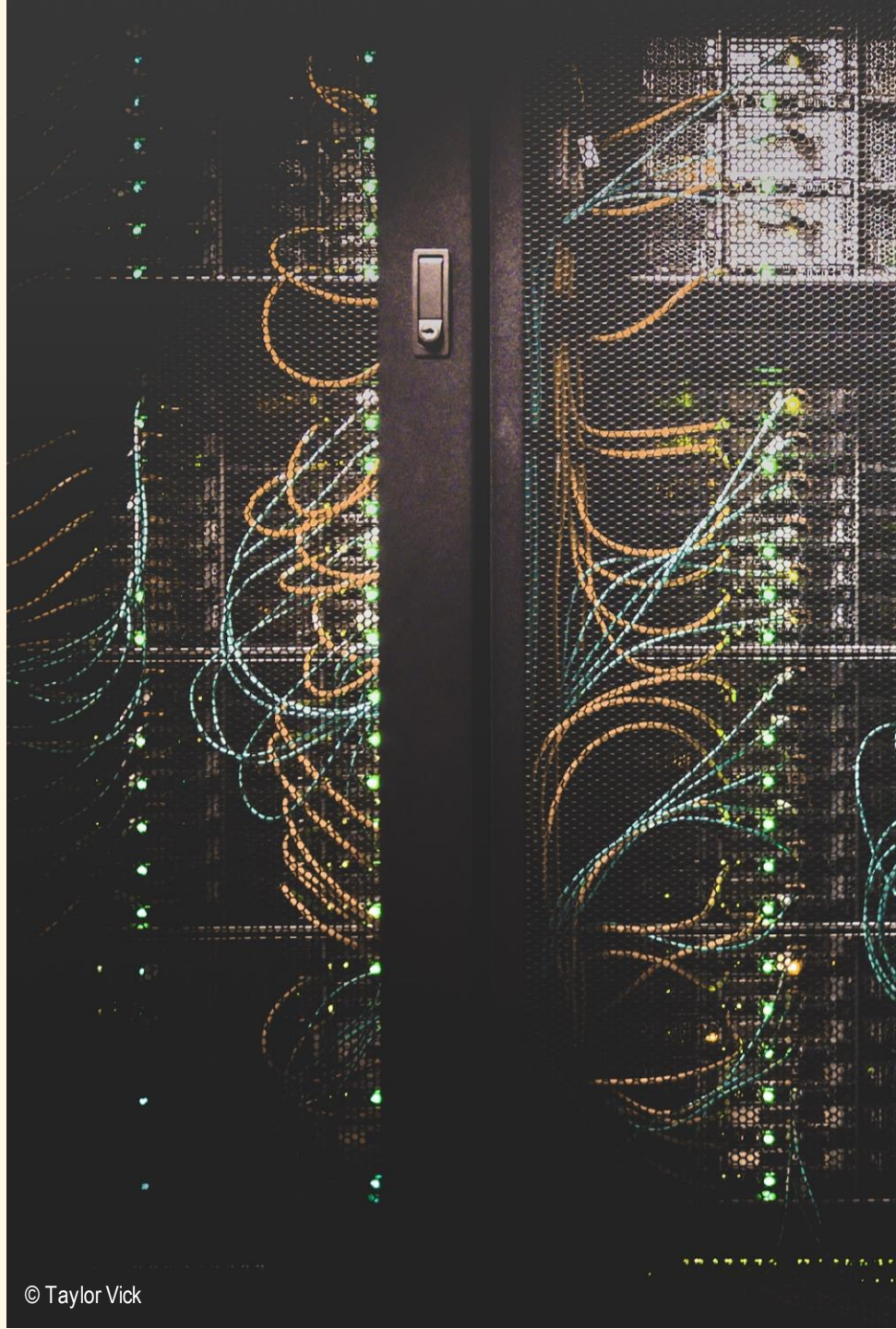
Strategies, challenges and implications

Bernat Montaña

Juan-José Boté-Vericad

Universitat de Barcelona
Facultat d'Informació i Mitjans Audiovisuals

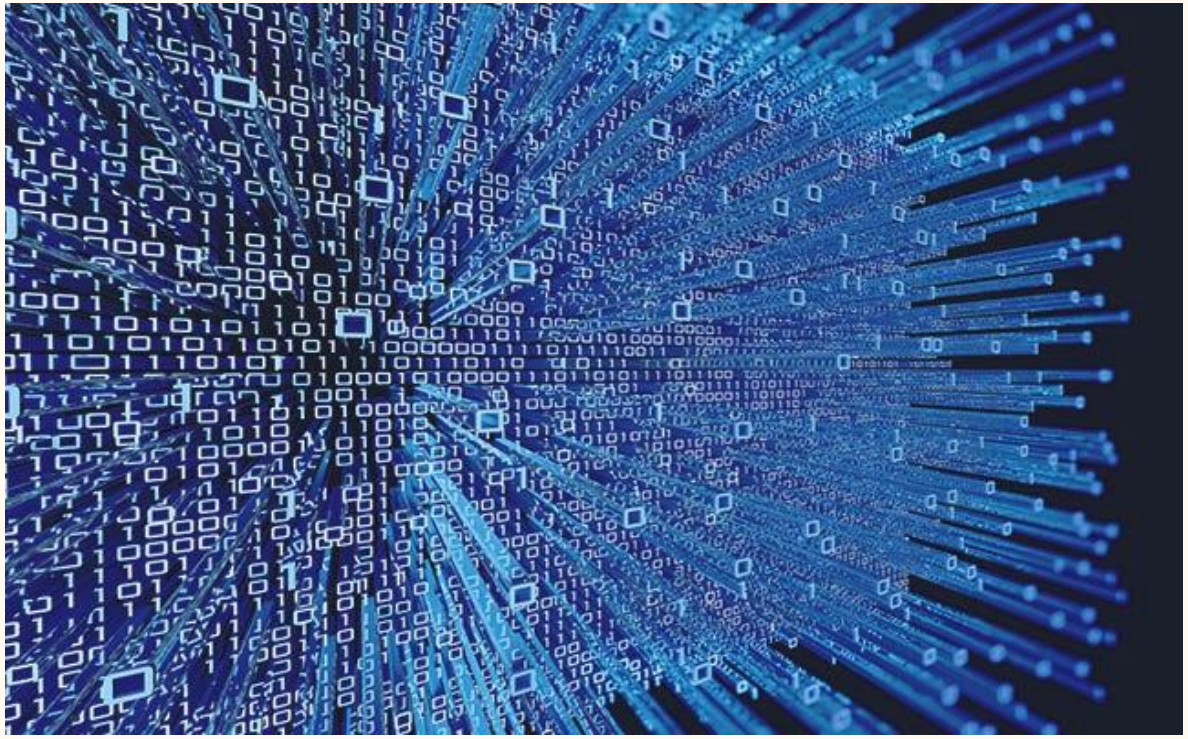
BOBCATSSS 2024
Universidade de Coimbra
Portugal

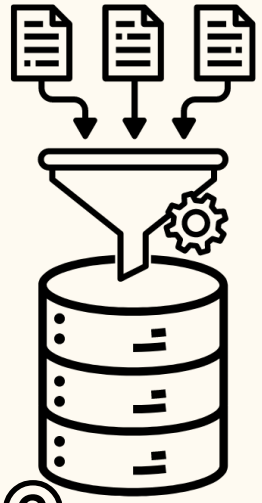


© Taylor Vick

The importance of data

- 328.77 million terabytes of data are created each day or around 3.28771341e+20 Bytes
- "Data is the new oil" [...valuable but if unrefined it cannot really be used]
- 90% of the world's data was created in the last 2 years





Objectives and research problem

Research objectives:

- Understand the FAIR Principles
- Explore effective research data management strategies
- Analyze the FAIR Principles implementation on existing datasets

What's the current state of the FAIR Principles implementation?

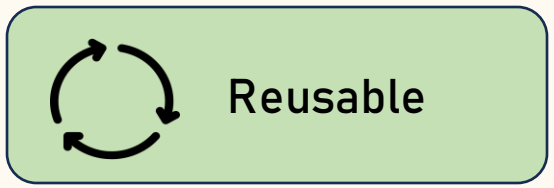
Is there a positive trend in the implementation of said principles over the years?

How can datasets be effectively optimized to align with and practically apply the FAIR Principles?

FAIR Data Principles

Introduced in 2016, the FAIR Data Principles are a transformative framework for scientific data management

This globally embraced principles, advocating for its 4 key points, enhance transparency and collaboration in research



DUTCH TECHCENTRE FOR LIFE SCIENCES

RESEARCH DATA ALLIANCE

Australian Research Data Commons

National Institutes of Health

EUROPEAN OPEN SCIENCE CLOUD

Methodology I. Rubrica

Findable		Accessible		Interoperable		Reusable	
Metadata uses standardized terms a nd formats	Grade 0 - 4	Data is stored in a secure and accessible location	Grade 0 - 4	Open and widely accepted formats and standards used	Grade 0 - 4	Methodology and code scripts are included	Grade 0 - 4
Persistent identifier is included in metadata	Grade 0 - 4	Data is available on a recognized data repository	Grade 0 - 4	Data is compatible with a variety of platforms and tools	Grade 0 - 4	Liscencing information is clearly specified	Grade 0 - 4
Keywords and tags are used effectively in metadata	Grade 0 - 4					Usage restrictions are clearly communicated	Grade 0 - 4
Total findable grade 0 - 12		Total accessible grade 0 - 8		Total interoperable grade 0 - 8		Total reusable grade 0 - 12	

Total grade
0 - 40

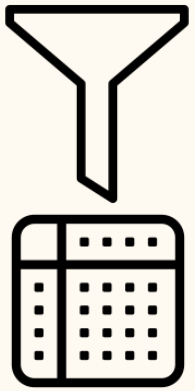
Methodology II. Sample collection

A dataset is a structured and organized collection of data, often presented in tabular form, that represents information about a particular domain or subject.

kaggle

Google Dataset Search Beta

DATA HUB



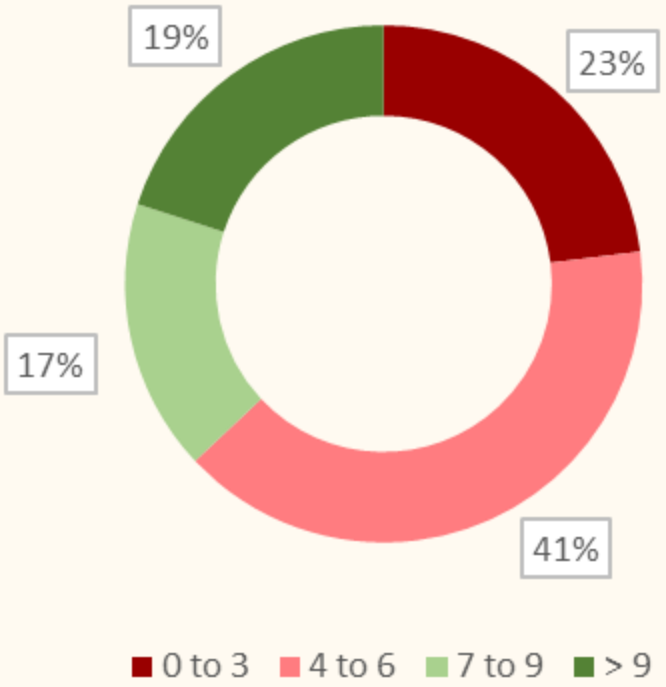
X 120 datasets

Methodology III. Sample analysis

Dataset	Findable			Accessible		Interoperable		Reusable			Grade				
Name/Title	Metadata uses standardized terms and formats	Persistent identifier is included in metadata	Keywords and tags are used effectively in metadata	Data is stored in a secure yet accessible location	Data is available online, preferably in a recognized data repository	Open and widely accepted data standards and formats are used	Data is compatible with a variety of platforms and tools	Data dictionaries, code scripts, and relevant methodologies are included	Licensing information is clearly specified	Any restrictions on data usage are clearly communicated.	Findable grade (0 - 12)	Accessible grade (0 - 8)	Interoperable grade (0 - 8)	Reusable grade (0 - 12)	**Total grade (0 - 40)**
Holstein Cattle Recognition	4	4	4	4	3	4	4	3	4	4	12	7	8	11	38
Flota de taxis segons combustible	1	0	1	2	2	3	4	0	0	0	2	4	7	0	13
Price of Used Toyota Corolla Cars	3	0	3	3	3	4	4	3	4	3	6	6	8	10	30
Transport Networks of Spain	1	0	1	3	2	2	3	0	0	3	2	5	5	3	15
Library loans (books only)	3	0	2	2	2	3	4	2	3	3	5	4	7	8	24
FC Barcelona total spend (2010- 2023	1	0	2	0	1	3	3	0	0	0	3	1	6	0	10
Cloud types	2	0	1	2	1	3	2	1	0	0	3	3	5	1	12
Library Usage	3	0	1	2	4	3	3	2	3	3	4	6	6	8	24
Universitats subvencionades amb est	3	0	4	3	2	4	4	1	3	3	7	5	8	7	27
Ontario public library statistics	4	0	4	3	2	3	3	2	3	4	8	5	6	9	28
Library Collection Inventory	4	0	4	4	3	4	3	3	0	1	8	7	7	4	26
Green pea production in Romania(20	1	0	2	2	1	3	3	0	0	2	3	3	6	2	14
Collection of cow's milk	3	0	3	3	3	4	4	3	3	3	6	6	8	9	29
Transport Accessibility Data	3	0	2	3	2	3	4	3	3	2	5	5	7	8	25
Road freight transport	3	0	3	3	2	4	4	2	1	2	6	5	8	5	24
NOAA - Severe weather warnings tor	4	0	3	4	3	4	4	1	0	0	7	7	8	1	23
Historical Tornado Tracks	4	0	3	4	3	4	4	2	2	3	7	7	8	7	29

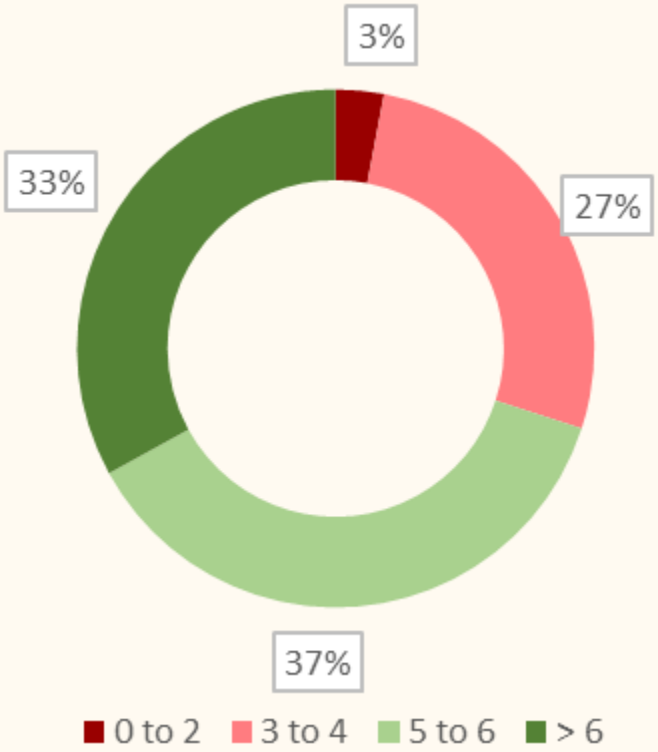
Results

Findability rubrica points analysis
0 - 12



*General lack of identifiers

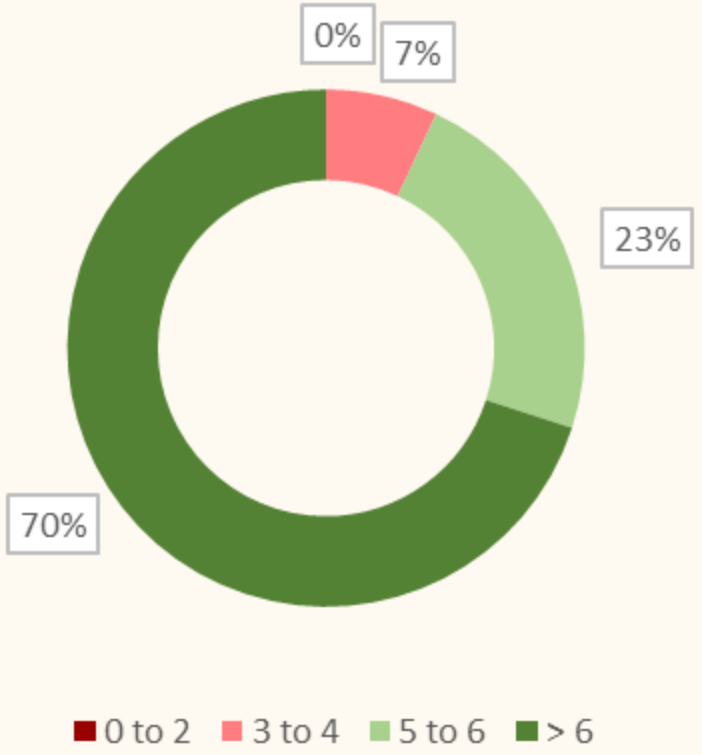
Accessibility rubrica points analysis
0 - 8



*Slight trend to not use recognized repositories

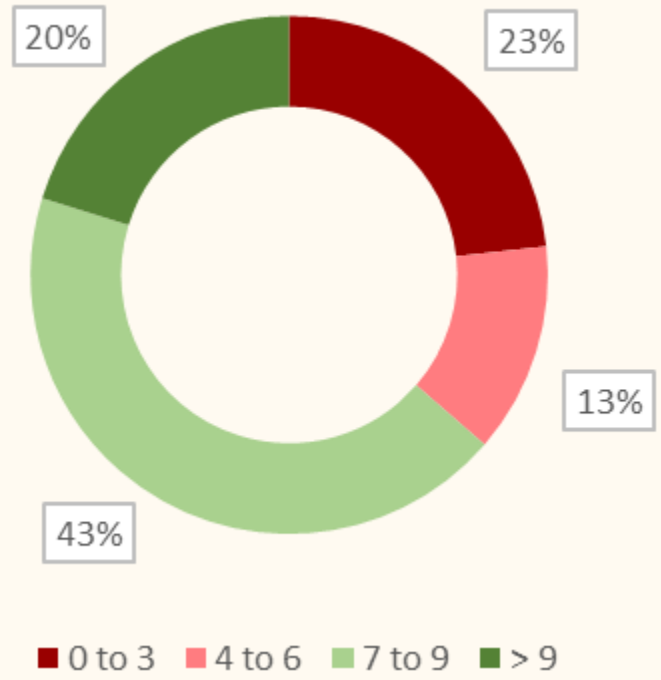
Results II

Interoperability rubrica points analysis
0 - 12



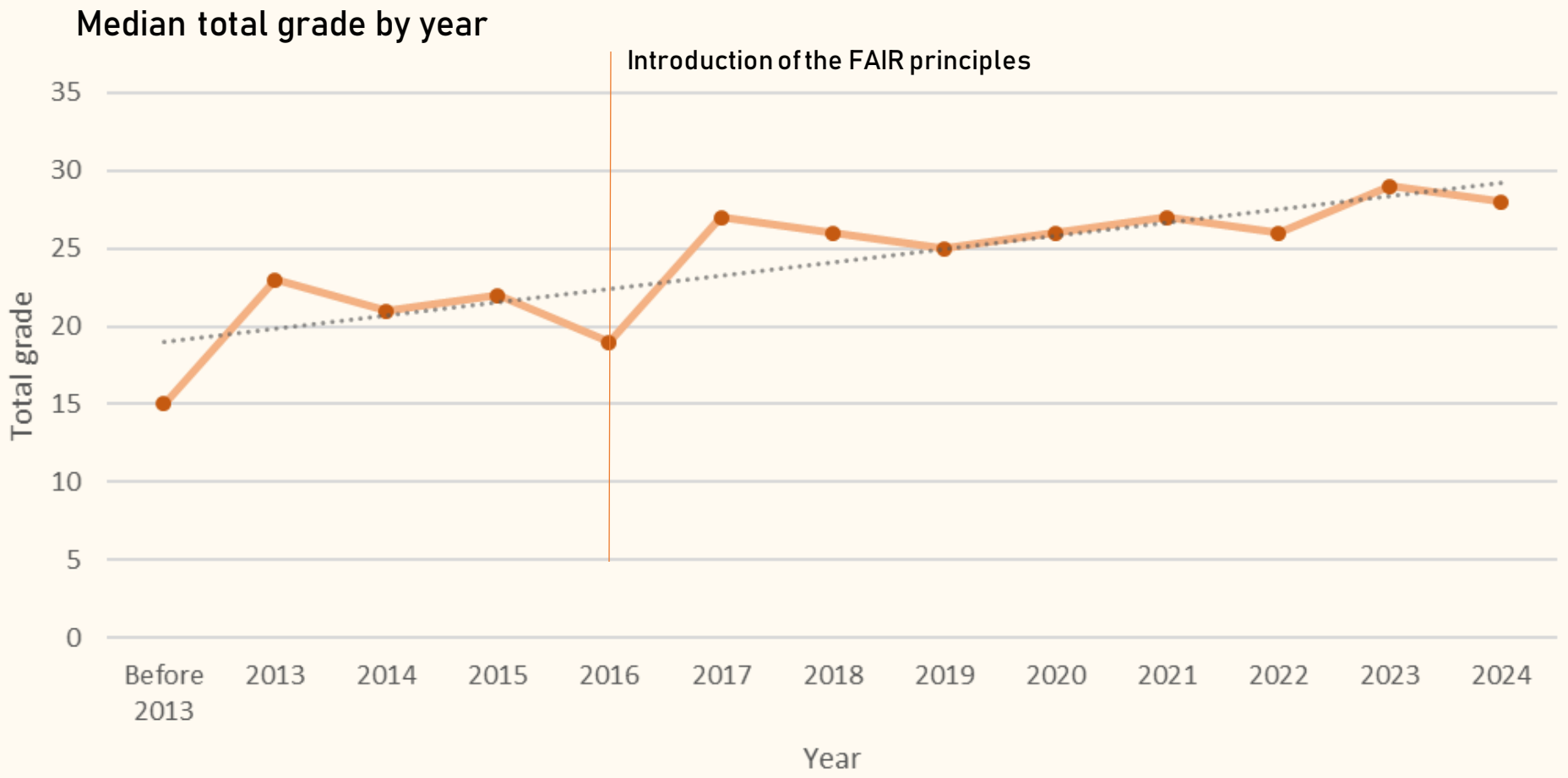
*Most of them comply

Reusability rubrica points analysis
0 - 12



*General lack of methodologies

Results III



Conclusion

Great progress is being achived, especially by on the accessibility and interoperability sections

Achived goals	Work in progress
Follow standard data types and formats	Include methodologies about data
Data stored in a secure and accessible locations	Add a unique identifier to the dataset

Future research could focus on more specific samples to get an in depth view of that field. The rubrica could also be modified for future studies

Thank you!

linkedin.com/in/bernat-montana/

linkedin.com/in/juan-jos%C3%A9-bot%C3%A9-vericad/

UNIVERSITAT DE BARCELONA

Facultat d'Informació i Mitjans Audiovisuals

BOBCATSSS 2024 UNIVERSIDADE D COIMBRA Portugal