ENVRI-FAIR Procedure to extract FAIR Answers from FAIR Questionnaires

Barbara Magagna barbara.magagna@umweltbundesamt.at

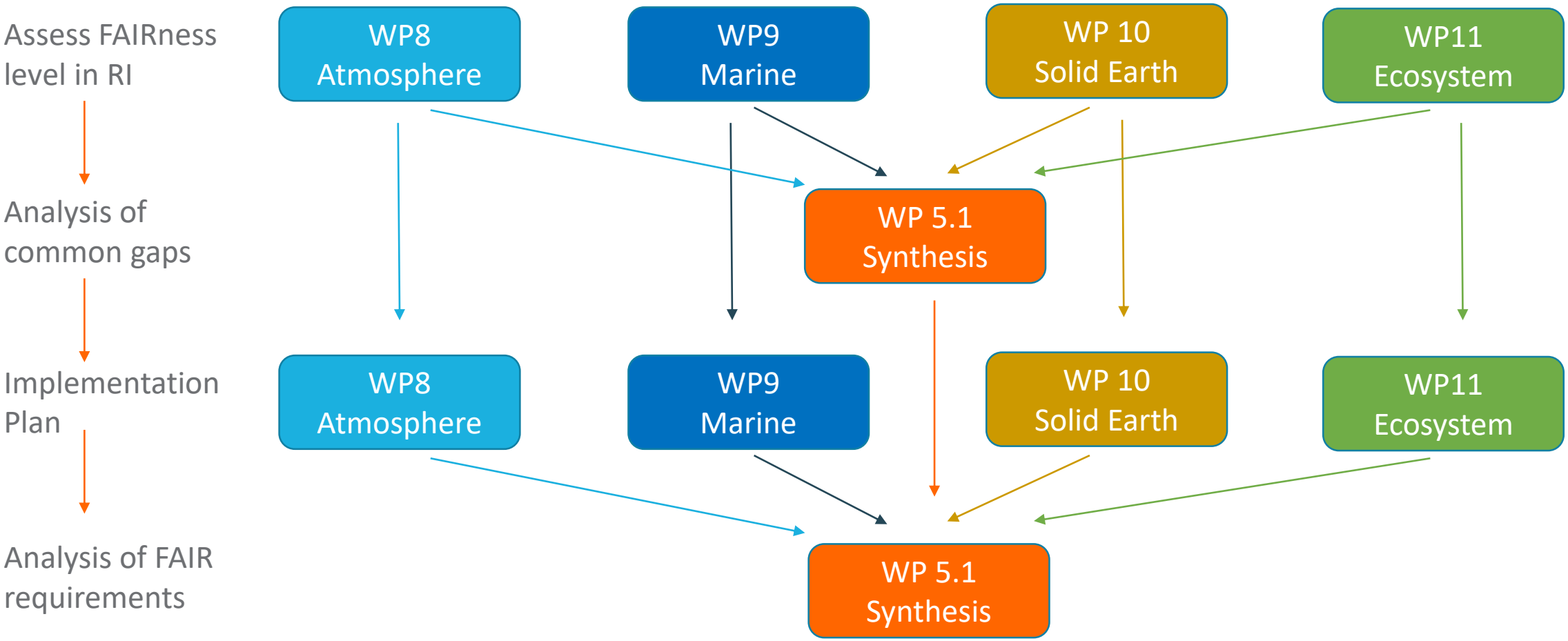
Ecosystem Research and Environmental Information Management

umweltbundesamt
PERSPEKTIVEN FÜR UMWELT & GESELLSCHAFT

Lecce, July 2 2019

ENVRI-FAIR receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824068

Assess FAIRness in ENVRI-FAIR



FAIR Metrics (Gen1)

(14) FAIR Metrics (Gen1, 2018): FAIRness is measurable -> FAIR Metric Authoring Template

Example: F1B, Identifier Persistence
https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_F1B.pdf

Question: Provide URL to a document describing your persistence policy

Test: Does the URL resolve?

-> Semiautomated Test via Questionnaire (1) plus minimal validation

13	5	F1A: Please provide the IRI for a registered identifier schema for your resource's IRI (e.g. DOI, HTTP): *	*What must be provided? URL to a registered identifier scheme.*	
14	6	F1B: Please provide the IRI to the document describing the persistence policy for the identifier of this(meta)data (this may be a document from your service provider, e.g. Zenodo, UniProt, etc.):	*In simple words: What should be provided? A URL that resolves to a document containing the relevant policy.*	

From Metrics to Maturity Indicators

BUT: Questionnaire-based approaches are subjective and do not evaluate the entire point of FAIRness

AND: The questionnaire is not well accepted by non technical experts, as many of the questions are considered to be too specific and difficult to understand

THUS: an objective evaluation system is needed, based on the key question if a machine can find and (re)use data consuming only the GUID of the metadata

FAIR Maturity Indicators (Gen2)

- Maturity Indicator Tests (Gen2, 2019)
  - are standalone Web interfaces that can be created by anyone with a clear workflow
  - return binary pass/fail
  - they log everything they do, reason of failure/success is transparent and helpful for improvement
  - Communities can decide which Maturity Indicators are relevant to them

https://fairsharing.github.io/FAIR-Maturity-FrontEnd/#!/#%2F!

FAIR Evaluation Service

FAIR Evaluation Services | HOME | EVALUATIONS | MI TESTS | COLLECTIONS | ABOUT US | Search

FAIR METRICS GEN2- UNIQUE IDENTIFIER

FAIR METRICS GEN2 - IDENTIFIER PERSISTENCE

FAIR METRICS GEN2 - DATA IDENTIFIER PERSISTENCE

Status: Failure
Principle tested: F1
Description: Metric to test if the unique identifier of the data resource is likely to be persistent. Known schema are registered in FAIRSharing (https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema). For URLs that don't follow a schema in FAIRSharing we test known URL persistence schemas (purl, oclc, fdlp, purlz, w3id, ark).
Metric test created on: May 8, 2019 by Mark D Wilkinson (updated on May 8, 2019).
Test executed on: May 13, 2019

Test results

```
INFO: Found a URI.
INFO:  Attempting to resolve http://data.emso.eu/files/ using HTTP Headers {"Accept"=>"text/turtle, application/ld+json, application/rdf+xml, text/xhtml+xml, application/n3, application/rdf+n3, application/turtle, application/x-turtle, text/n3, text/turtle, text/rdf+n3, text/rdf+turtle, application/n-triples"}.
INFO: Found html text/html type of content when resolving http://data.emso.eu/files/ using HTTP Accept header {"Accept"=>"text/turtle, application/ld+json, application/rdf+xml, text/xhtml+xml, application/n3, application/rdf+n3, application/turtle, application/x-turtle, text/n3, text/turtle, text/rdf+n3, text/rdf+turtle, application/n-triples"}.
INFO: parsing as HTML.
INFO:  Using 'extruct' to try to extract metadata from return value (message body) of http://data.emso.eu/files/.
WARN: the extruct tool failed to find parseable data at http://data.emso.eu/files/
```

ENVRI FAIR

Why FAIR Questionnaires in the ENVRI-FAIR context?

- Understand FAIR principles and their advantages for the Research Infrastructures (RIs)
- Assess the status quo of the RI‘s data and services in FAIR terms
- Detect information and implementation gaps
- Discover strenghts
- Compare the different implementations
- Evaluate possible technology takeups for improvements
- Prioritisize FAIR improvements
- Include chosen FAIR improvements in RI plans

-> FAIR Questionnaire 2 (by Kristina Hettne and Peter Wittenburg)

FAIR MATRIX

Services	Component	Most used	C2CAMP	OPEDAS	PHT	Rare-Diseases	
central to all	DOIP	DOIP	DOIP	DOIP	DOIP	DOIP	
central to all	Metadata format			RDF	RDF	RDF	
central to all	Metadata access protocol			LDP/FDP	LDP/FDP	LDP/FDP	
central to all	Metadata core elements	TBD on M4M		TBD on M4M	TBD on M4M	TBD on M4M	
Technology	Data Format			RDF for interop.	RDF for interop.	RDF for interop.	
Technology	Data Access Protocols (MR/A)			LDP/FDP	PHT-standard	PHT-standard	
Technology	Computer-actionable license description language			RDF	RDF	RDF	
Tooling	Repository (Data/Metadata)		DONA	IFDS Data Station	IFDS Data Station	ERN?	
Tooling(Repository)	https://www.dataone.org						
Tooling	Registry Service		DONA	IFDS Station Registry	IFDS Station Registry	ERN?	
tooling	Metadata forms/creators			CEDAR/CASTOR			
Tooling	Search capability		DOIP	IFDS Station Registry	IFDS Station Registry	IFDS Station Registry	
Policy	Persistence Policy			TBD	TBD	TBD	
Technology	Computer-actionable policy description language			RDF	RDF	RDF	
Tooling	License protocols			TBD	TBD	TBD	

From text to comparable structured information

Questionnaires sent to RI experts

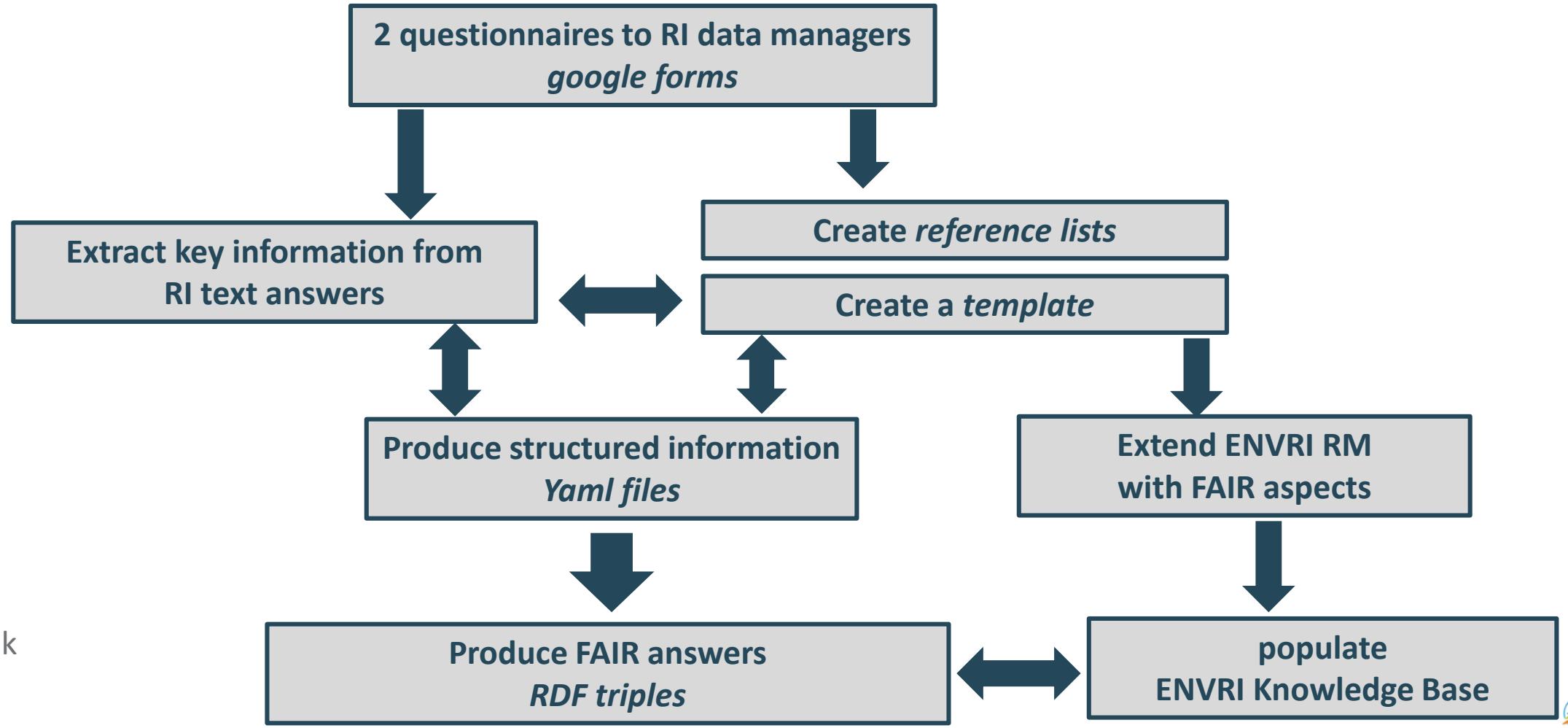
- Questionnaire 1 (25 Q)
- Questionnaire 2 (53 Q)

-> Questionnaire 3 (78 Q) plus online FAIR maturity evaluation

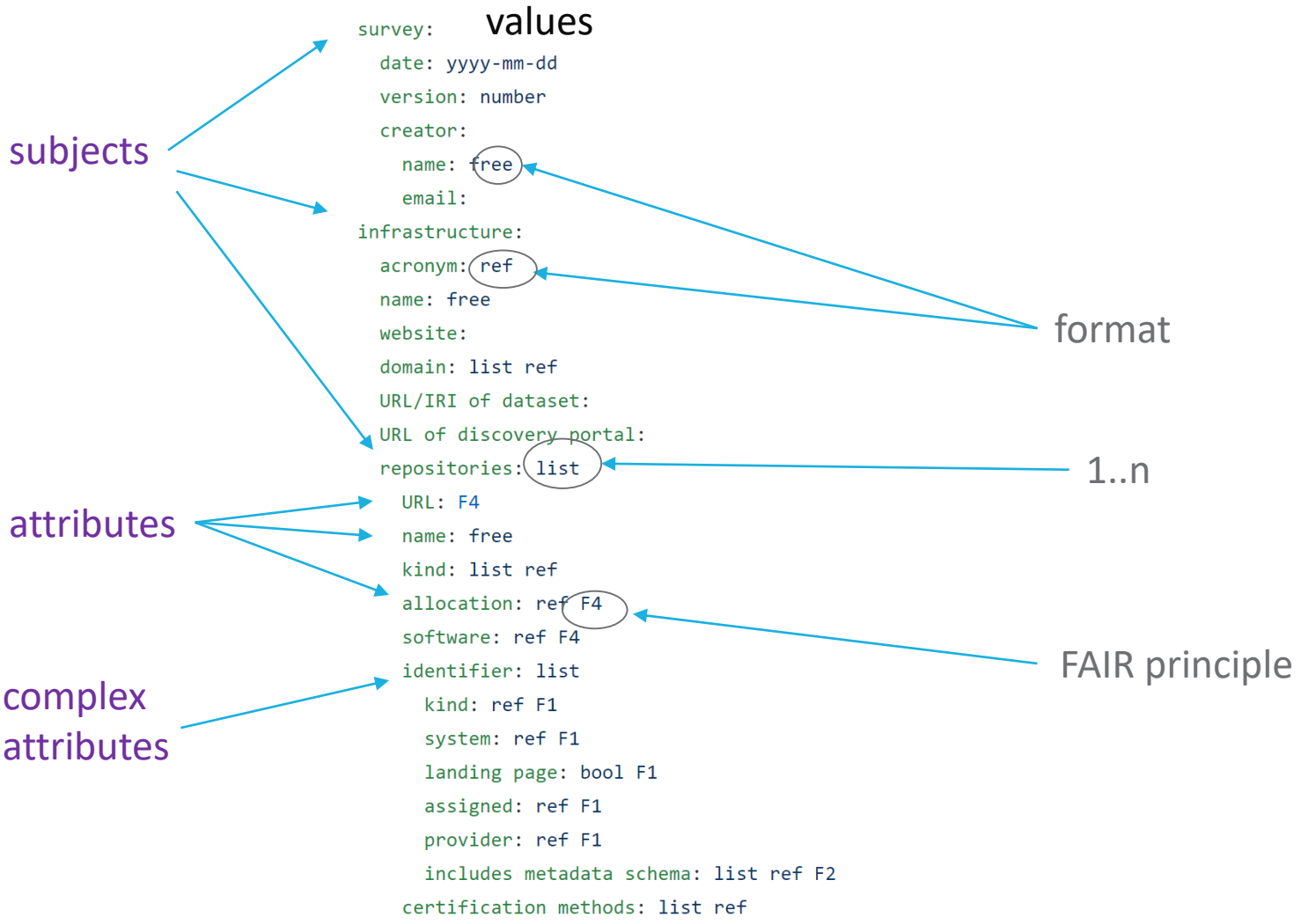
- **-> Spreadsheets with text information**
- Extract key information
- Transform into structured data
- Produce Yaml files
- Produce reference lists
- Produce RDF triples
- SPARQL triples

-> Spreadsheets with FAIR, comparable answers

ENVRI-FAIR – Work Flow FAIR Questionnaires



ENVRI-FAIR – Yaml Template



Answers: values

```
access mechanisms:
  authentication method: VOID
  access protocol URL: https://doi.org/10.17882/42182
  protocol open: yes
  protocol royalty free: yes
  own user database maintained: no
  ORCID used in AAI: no
  major access technology supported: VOID
  authorisation technique: VOID
  authorisation needed for: none
  authorization for accessing content needed: no
  access content process description IRI: VOID
  data licenses in use:
  - CC-BY4.0
  data license IRI:  https://creativecommons.org/licenses/by/4.0/
  metadata openly available: yes
data:
- type name: binary
  preferred formats:
  - format name: NetCDF
    metadata types in data headers:
    - NetCDF ACDD metadata
  registered data schema: no
  search on data: yes
  search engine URL: http://www.argodatamgt.org/Access-to-data/Argo-data-selection
metadata:
  schema:
  - URL: http://dx.doi.org/10.13155/29825
    name: Argo user manual
    provenance fields included:
    - text only
  categories defined in registries: planned
  PIDs included: yes
  primary storage format: NetCDF CF Argo
```

Possible Answers:

- free: free text (literal)
- URL/IRI: website address
- list: use "-" for each entry of a list
- bool: yes/no
- date: yyyy-mm-dd
- ref: reference list
- if no answer is given: NULL
- if no answer can be given: VOID
- if the answer is that it is planned to provide a solution: planned
- if you should use a reference list but your answer is negative: none
- if answer is defined bool, it is also possible to use: partially

Reference lists

repository software:

- 52NORTH SOS
- Sextant
- MS SQL
- D-SPace
- Fedora
- CDI system

identifier kind:

- PID
- local ID

identifier system:

- Handle
- LSR URN
- DOI

ENVRI-FAIR – Extract key information

1.10 Do you assign PIDs manually or automatically?	The Argo DOI fragments are assigned automatically	automatically	
1.11 Which PID registration provider do you use?	DataCite	SEANOE	
1.12 Do you use the PID Record to store attributes about the data?	"Yes for the monthly snapshot (the DOI+ fragment) No otherwise. "	yes	
1.13 Are these repositories certified? If so, which methods are used?	"Yes, Ifremer is DSA and IODE certified. Ifremer-Sismer is in certification process as “RDA-Trustworthy repository” "	- Data Seal of Approval	
1.14 Are repository policies mentioned at the website? If so, indicate the major ones.	"Yes https://creativecommons.org/licenses/by/4.0/ "	- data access	
1.15 Are your repositories registered in a registry? If so which registry?	Yes, GEO registry	GEO	
1.16 Which persistency guaranties are typically given?	"The Argo long term archive is managed by US-NCEI. US-NCEI has a Unesco-WMO mandate as world data centre (WDC-A). "	NULL	
1.17 Which are the most popular data types used?	The self-describing NetCDF CF format Argo implementation	binary	

```
repositories:
- URL: http://doi.org/10.17882/42182
  name: Euro-Argo Data
  kind: data repository
  data repository type: central
  metadata repository type: central
  software: NetCDF file
  identifier:
  - IRI: http://doi.org/10.17882/42182
    persistency-policy document IRI: https://doi.org/10.13155/44515
    kind: DOI
    system: SEANOE
    assigned: automatically
    provider: SEANOE
    includes-attributes: yes
  certification methods:
  - Data Seal of Approval
  - IODE certified
  - RDA Trustworthy Data Repository
  policies:
  - data access
  registries:
  - GEO
  persistency-guaranty: NULL
  access mechanisms:
    authentication method: VOID
    access protocol URL: https://doi.org/10.17882/42182
```

Questionnaire evolution

- Questionnaire (Q) 1 and 2 from Go-FAIR
- At WP9 workshop discussed usefulness of Q1
- Q1 partly replaced by the online assessment service:

-> https://fairsharing.github.io/FAIR-Maturity-FrontEnd/#!/#%2F!

- Some of the questions of Q1 integrated in Q3
- Q3 simplified and aligned to logic of the yaml structure: https://docs.google.com/spreadsheets/d/14omLaR1gbmd5v14aVfw79272JerJEI3l/edit#gid=722055771 see new version tab!
  - Unambiguous questions with only one correspondent answer requested
  - More than one answer organised as lists
  - Sequence following the logic of the yaml template

What is needed

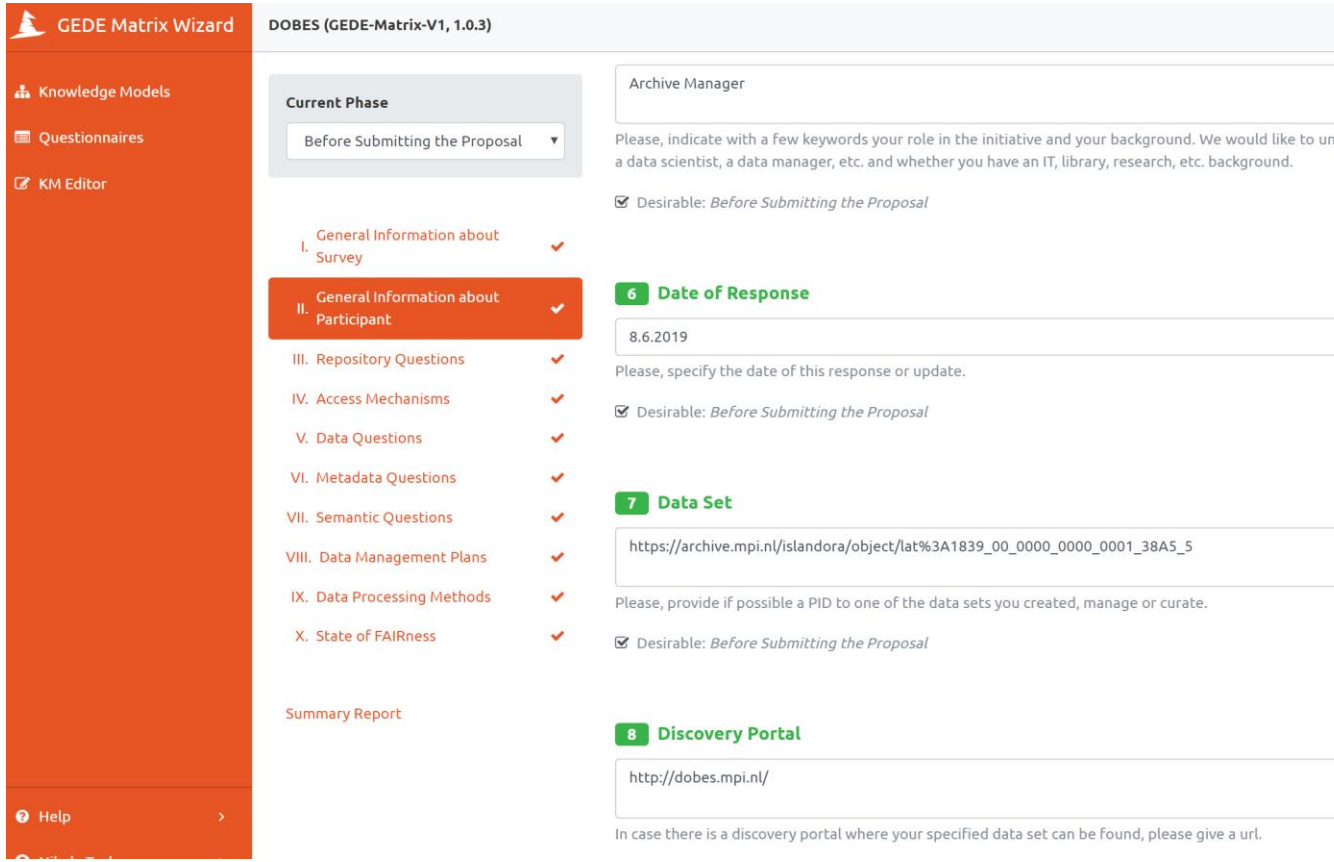
- To understand the questions
- To deliver concrete answers for each repository
- To differentiate between NULLs (no answer), misinterpretations, none(s)
- To identify real gaps
- To understand what the RIs planned
- To produce overviews and synthesis of the questionnaire outputs
- To understand in which direction the RI wants to enhance their FAIRness
- To repeat this FAIRness assessment in two years using an automated process (Wizard!)

Important links and documents

- GitHub: https://github.com/envri-fair/fairness-scorer
- Google drive: https://drive.google.com/drive/folders/1kRJmDr2oMEOJ4_hAxDxvBL-Oe-TyN0Os?usp=sharing
- New questionnaire: https://drive.google.com/file/d/14omLaR1gbmd5v14aVfw79272JerJEI3l/view?usp=sharing
- Yaml Template: https://docs.google.com/document/d/1W6sVFcPpqjPHY9dZkIWypBl3jsXW0xts-to8k7zgueA/edit?usp=sharing
- Reference lists: https://docs.google.com/document/d/1lKadyZII9S8vsiGWRJEQx_YRGjjGhd3IW0hi25qcK4Y/edit?usp=sharing

Collaboration in the FAIR Matrix Team

- Consolidate Questionnaire
- Adopt the Data Stewardship WIZARD:
  - Use drop down lists
  - Allow additional entries
  - Export answers as RDF triples
  - Will be used for ENVRI-FAIR (V2)
- Align possible answers and produce controlled lists
- Build a common knowledge model
- Include in ENVRI RM
- Include ENVRI choices in the FAIR Matrix



ENVRI
FAIR

THANKS

envri.eu/envri-fair

@ENVRIcomm

ENVRI community

facebook.com/ENVRIcomm