Metadata-powered characterization of Digital Twins in DT-GEO

Pablo Orviz <orviz@ifca.unican.es>

IFCA-CSIC

on behalf of DT-GEO WP4 team

DT-GEO

This project has received funding from the European Union's Horizon research and innovation programme under the grant agreement No 101058129

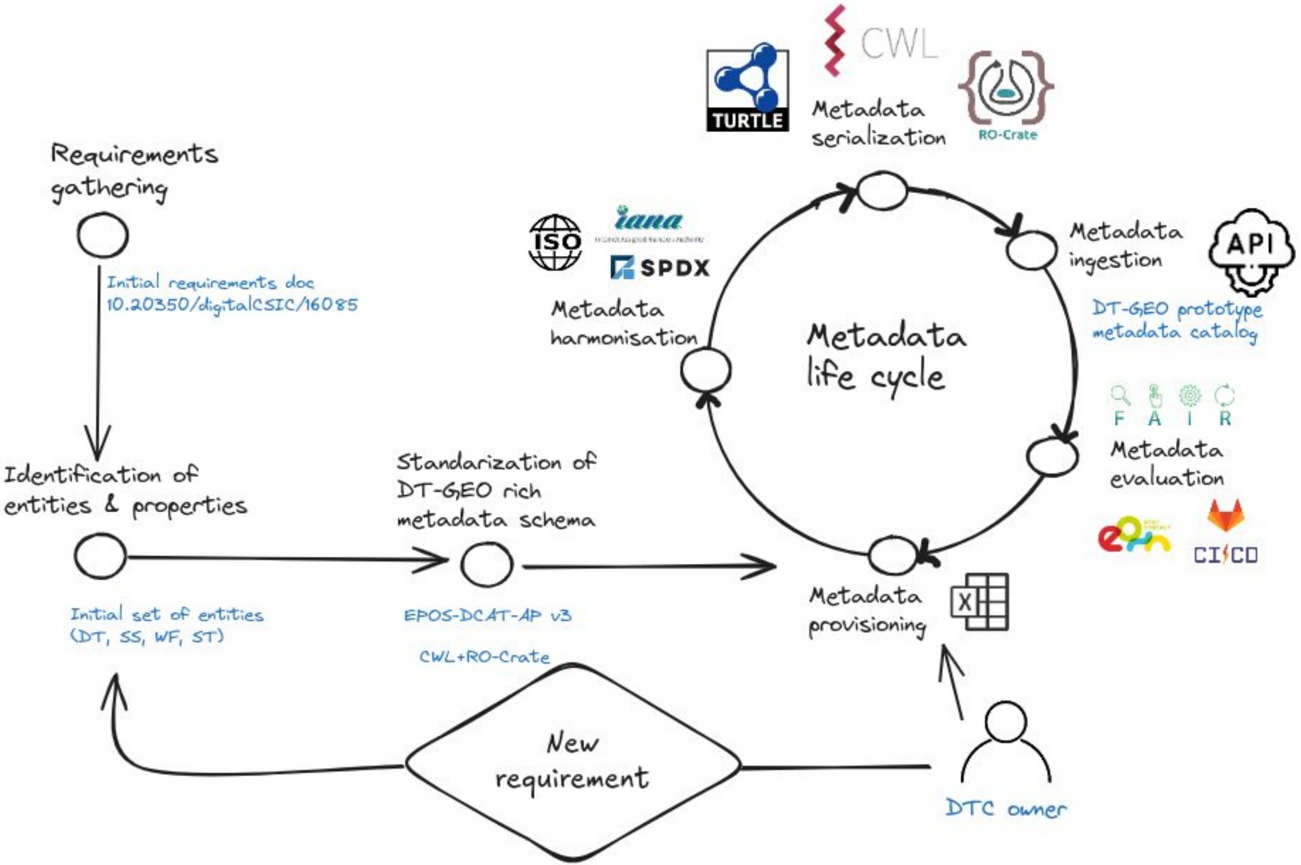
Context

- The goal of the project: develop a **prototype for a digital twin** on geophysical extremes (earthquakes, volcanoes, tsunamis, and anthropogenic-induced extreme events)
  - **12 Digital Twin Components (DTCs)** are being developed embedding flagship simulation codes that address specific scientific questions
  - DTCs will be **verified at 13 Site Demonstrators (SD)**

- The role of metadata:
  - **Characterise (and keep track of) the variety of digital assets** used by the DTCs into efficient workflows
  - Allow sufficient **richness of expression** to allow automated or semi-automated workflow orchestration
  - Promote **adherence with FAIR and quality assurance principles** of the digital assets

Metadata management in DT-GEO

Approach



Requirements gathering

Initial requirements doc
10.20350/digitalCSIC/16085

Identification of entities & properties

Initial set of entities
(DT, SS, WF, ST)

Standarization of DT-GEO rich metadata schema

EPOS-DCAT-AP v3

CWL+RO-Crate

Metadata life cycle

Metadata harmonisation

Metadata serialization

Metadata ingestion

DT-GEO prototype metadata catalog

Metadata evaluation

Metadata provisioning

New requirement

DTC owner

#1 Defining the DT-GEO metadata schema [M1-M6]

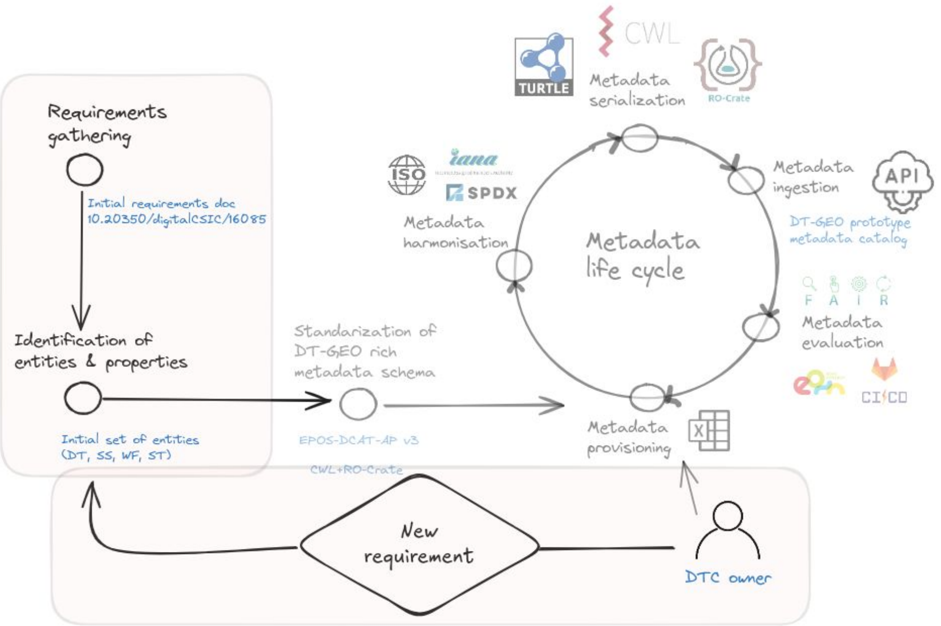
Timeline:

- **[M1 to M3] Requirements gathering**: joint effort among horizontal & vertical WPs (https://doi.org/10.20350/digitalCSIC/16085)
- **[M3 to M6] Initial definition of the DT-GEO metadata schema** (theoretical)
  - Metadata knowledge graph in accordance with CERIF (Common European Research Information Format)
  - Extension of the schema used under the European Plate System (EPOS ERIC)

Structure of the schema:

- **Base entities (aka "digital assets")**
  - **Datasets (DT) and Software-services (SS)**
  - **Workflow** (**WF**) and **Step** (**ST**)
- **Link entities** or Relationships
- Semantic rich identifiers (see table)

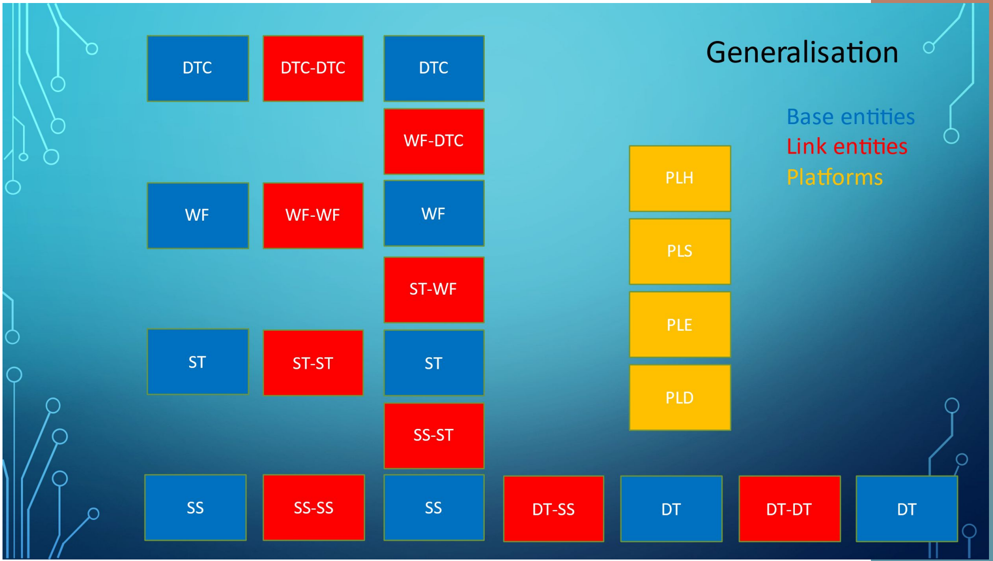
DTWnn	DTW	Digital Twin
DTC<WPn><DTCn>	DTC	Digital Twin Component
WF<WPn><DTCn><WFnn>	WF	Workflow
ST<WPn><DTCn><WFnn><STnn>	ST	Step
SS<WPn><DTCn><SSnn>	SS	Software Service (i.e. executable code)
DT<WPn><DTCn><DTnn>	DT	Dataset
DP<WPn><DTCn><DPnn>	DP	Data Product
SO<WPn><DTCn><SOnn>	SO	Source code of software



#1 Defining the DT-GEO metadata schema [M1-M6]

Timeline:

- **[M1 to M3] Requirements gathering**: joint effort among horizontal & vertical WPs (https://doi.org/10.20350/digitalCSIC/16085)
- **[M3 to M6] Initial definition of the DT-GEO metadata schema** (theoretical)
  - Metadata knowledge graph in accordance with CERIF (Common European Research Information Format)
  - Extension of the schema used under the Europea Plate System (EPOS ERIC)



Structure of the schema:

- **Base entities (aka "digital assets")**
  - **Datasets (DT) and Software-services** (**SS**)
  - **Workflow** (**WF**) and **Step** (**ST**)
- **Link entities** or Relationships
- Semantic rich identifiers (see table)

DTWnn	DTW	Digital Twin
DTC<WPn><DTCn>	DTC	Digital Twin Component
WF<WPn><DTCn><WFnn>	WF	Workflow
ST<WPn><DTCn><WFnn><STnn>	ST	Step
SS<WPn><DTCn><SSnn>	SS	Software Service (i.e. executable code)
DT<WPn><DTCn><DTnn>	DT	Dataset
DP<WPn><DTCn><DPnn>	DP	Data Product
SO<WPn><DTCn><SOnn>	SO	Source code of software

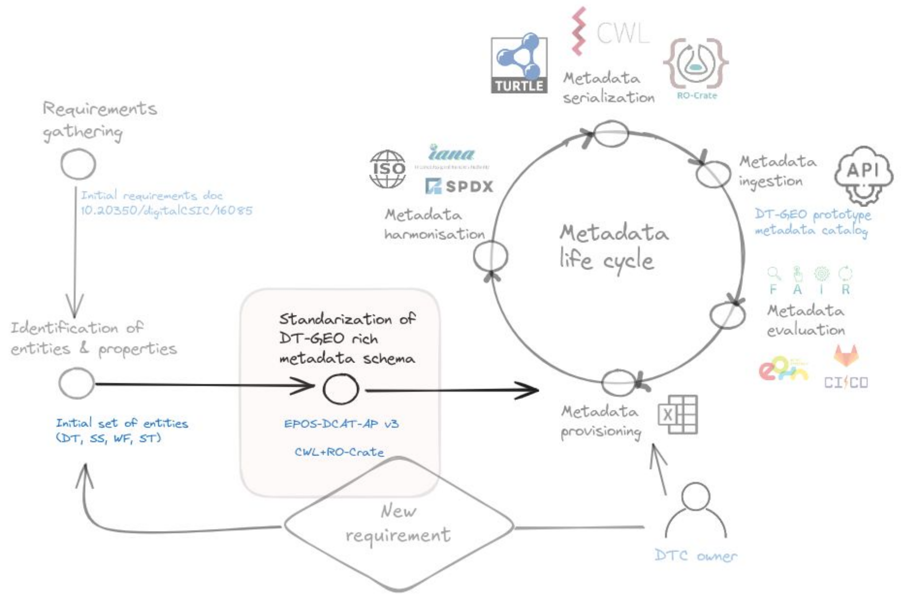
#2 Standarization phase: DTs and SSs [M6-M18]

EPOS-DCAT-AP v3 released:

https://epos-eu.github.io/EPOS-DCAT-AP/v3/

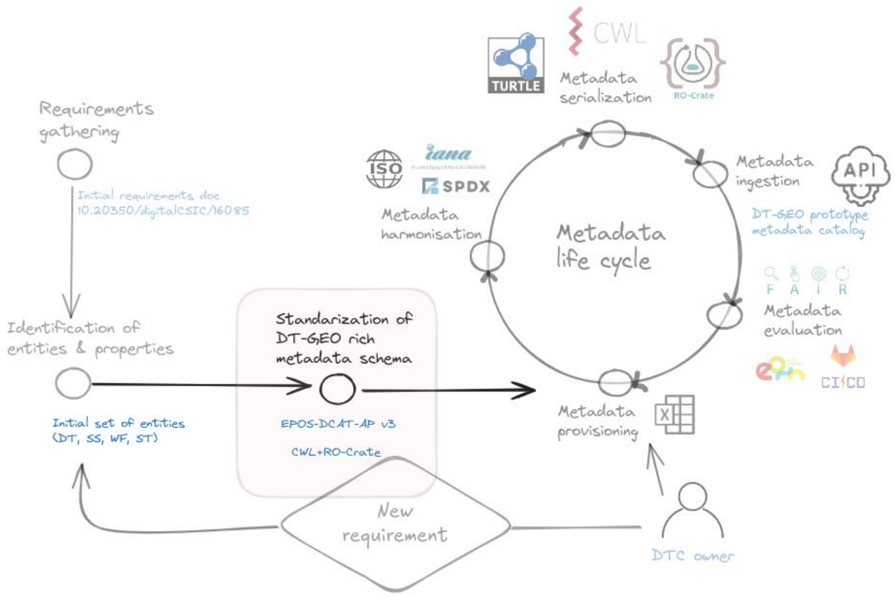
- Only for pre-existing entities in EPOS: **DTs and SSs**
- **Mappings** between DT-GEO schema and current production version of EPOS-DCAT-AP vocabulary (see table below)
  - **Extensions** done to EPOS-DCAT-APv2 (orange)
- **Controlled vocabularies (CVs)** for the main properties (Keywords, IDs, Person and Organisation, File formats, ..) were identified

DT-GEO extended schema	EPOS-DCAT-AP mapping class	EPOS-DCAT-AP mapping property	Controlled vocabularies
Unique ID	Dataset	dct:identifier adms:identifier	UUID HTTP URI URN + OID IRI
Name	Dataset	dct:title	ASCII, unicode, UTF-16
Type	Dataset DataService	dct:type	MX_ScopeCode codelist (ISO 19115, 19115-2)
..	..	..	
Maturity level	Distribution	adms:status	TRL levels



#2 Standarization phase: WFs (and STs) [M18-M24]

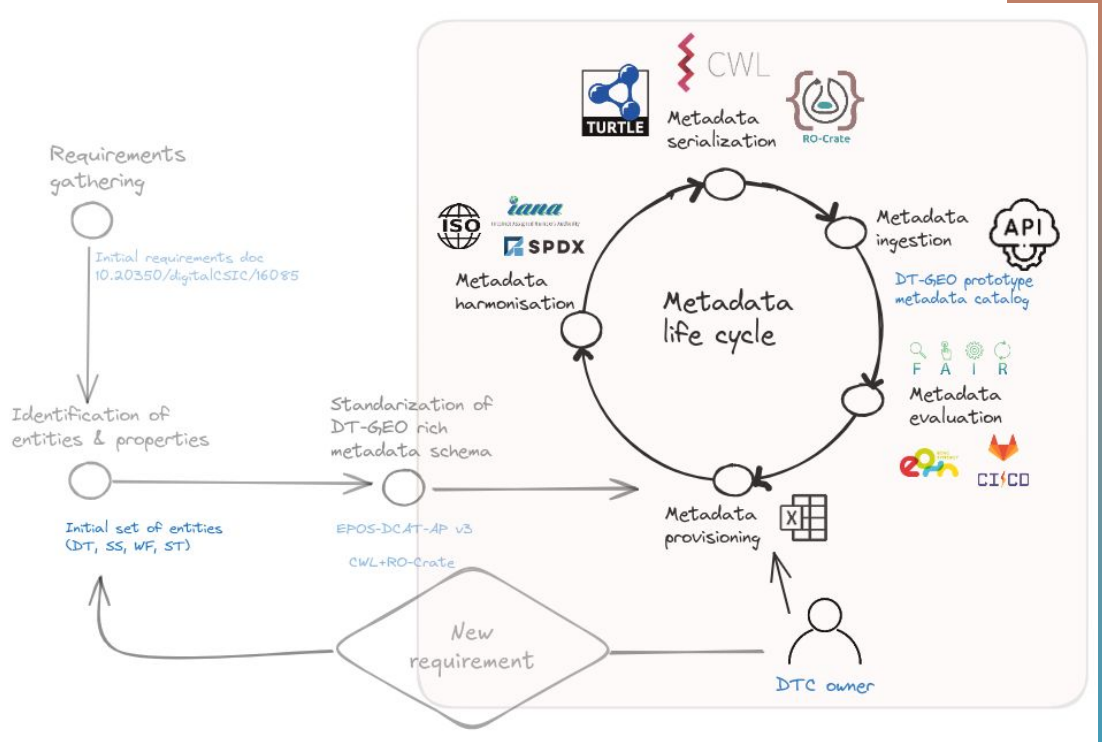
- Metadata description of workflows follows CWL+RO-Crate solution
  - Abstract descriptions of workflows
    - Prospective provenance
  - Uses CWL for defining the graph of relationships (link entities) among objects (base entities)
    - Worfklows, subworkflows and steps
    - Software and input/output data consumed/produced within the workflow steps
  - Uses RO-Crate to package CWL + research (meta)data
    - **RO-Crates references** the base entities managed through **EPOS-DCAT-AP v3**



#3 Metadata life cycle [M12-today]

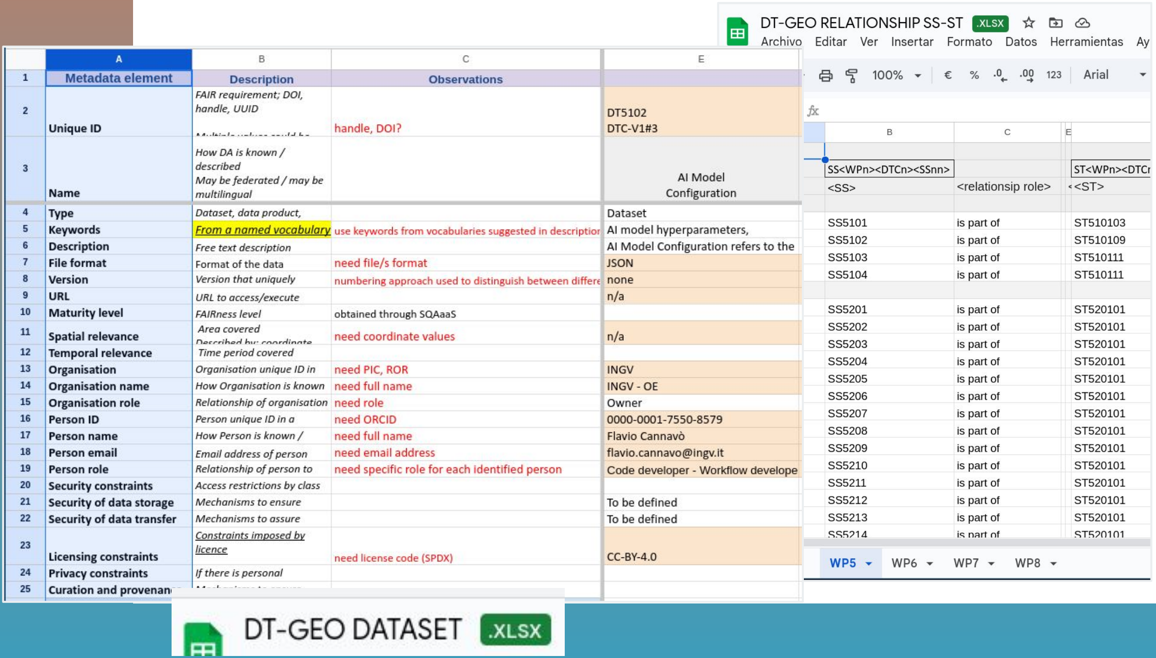
Continuous improvement of metadata; comprises:

1. Metadata provisioning
2. Metadata harmonisation
3. Metadata serialisation
4. Metadata ingestion
5. Metadata evaluation



#3.1 Metadata provisioning (life cycle)

Metadata is **provided by DTC owners through shared spreadsheets**: comprehensive characterisation of base and link entities

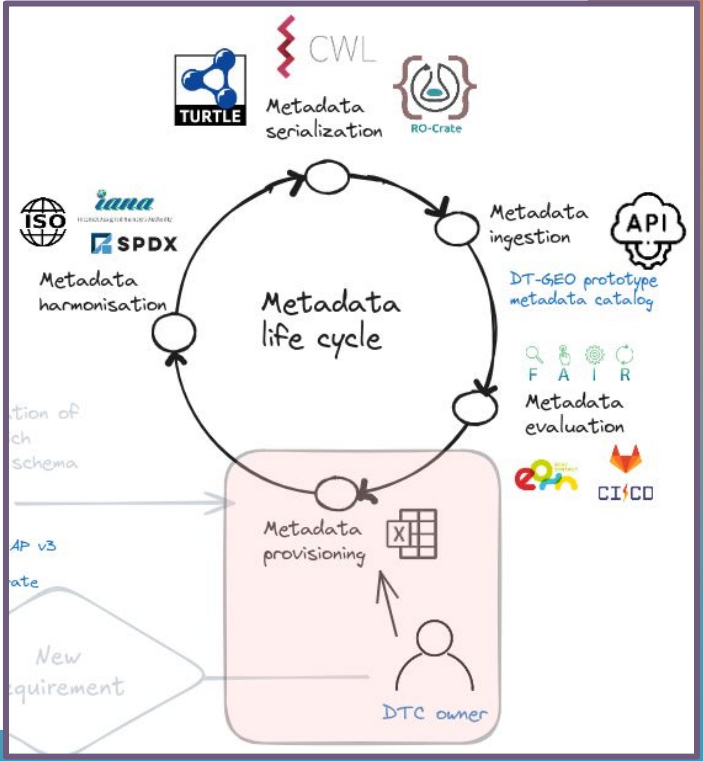


DT-GEO DATASET .XLSX

	A	B	C	E
1	Metadata element	Description	Observations	
2	Unique ID	FAIR requirement; DOI, handle, UUID	handle, DOI?	DT5102 DTC-V1#3
3	Name	How DA is known / described May be federated / may be multilingual		AI Model Configuration
4	Type	Dataset, data product,		Dataset
5	Keywords	From a named vocabulary	use keywords from vocabularies suggested in descriptior	AI model hyperparameters,
6	Description	Free text description		AI Model Configuration refers to the
7	File format	Format of the data	need file/s format	JSON
8	Version	Version that uniquely	numbering approach used to distinguish between differe	none
9	URL	URL to access/execute		n/a
10	Maturity level	FAIRness level	obtained through SQAaaS	
11	Spatial relevance	Area covered Described by: coordinate	need coordinate values	n/a
12	Temporal relevance	Time period covered		
13	Organisation	Organisation unique ID in	need PIC, ROR	INGV
14	Organisation name	How Organisation is known	need full name	INGV - OE
15	Organisation role	Relationship of organisation	need role	Owner
16	Person ID	Person unique ID in a	need ORCID	0000-0001-7550-8579
17	Person name	How Person is known /	need full name	Flavio Cannavò
18	Person email	Email address of person	need email address	flavio.cannavo@ingv.it
19	Person role	Relationship of person to	need specific role for each identified person	Code developer - Workflow develope
20	Security constraints	Access restrictions by class		
21	Security of data storage	Mechanisms to ensure		To be defined
22	Security of data transfer	Mechanisms to assure		To be defined
23	Licensing constraints	Constraints imposed by licence	need license code (SPDX)	CC-BY-4.0
24	Privacy constraints	If there is personal		
25	Curation and provenan	Mechanisms to ensure		

DT-GEO RELATIONSHIP SS-ST .XLSX

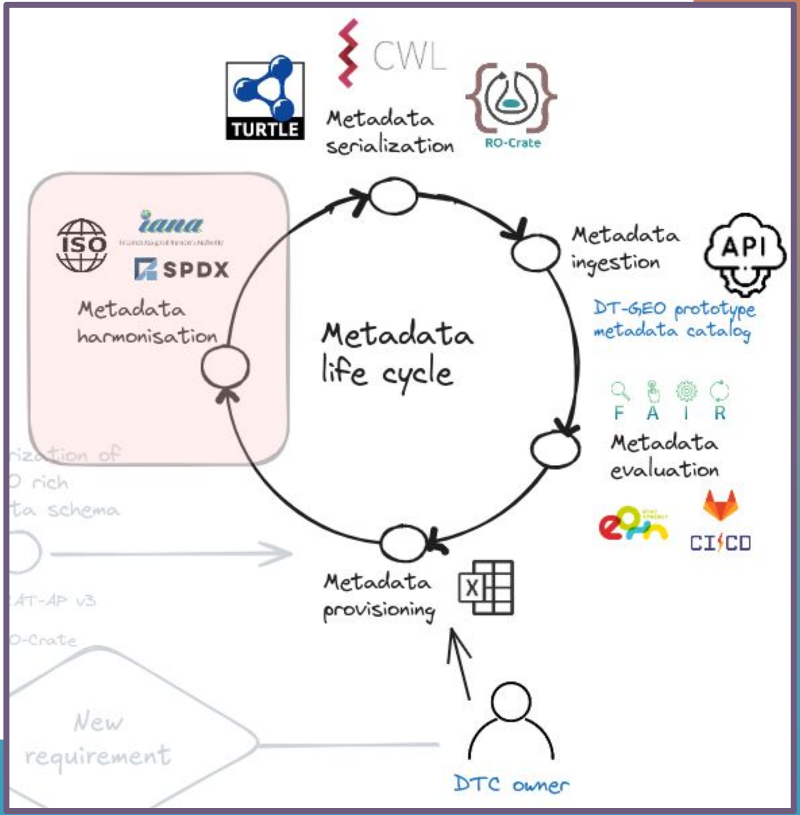
B	C	E
SS<WPn><DTCn><SSnn>		ST<WPn><DTCr
<SS>	<relationsip role>	<ST>
SS5101	is part of	ST510103
SS5102	is part of	ST510109
SS5103	is part of	ST510111
SS5104	is part of	ST510111
SS5201	is part of	ST520101
SS5202	is part of	ST520101
SS5203	is part of	ST520101
SS5204	is part of	ST520101
SS5205	is part of	ST520101
SS5206	is part of	ST520101
SS5207	is part of	ST520101
SS5208	is part of	ST520101
SS5209	is part of	ST520101
SS5210	is part of	ST520101
SS5211	is part of	ST520101
SS5212	is part of	ST520101
SS5213	is part of	ST520101
SS5214	is part of	ST520101



#3.2 Metadata harmonisation (life cycle)

Manual **curation of metadata** by the Data management team (WP4) for **efficient interoperability**

- Syntax (structure)
  - ✓ Avoidance of duplication of the same entity, attribute and/or instance
  - ✓ Ensure referential and functional dependency on the unique (semantic rich) identifiers
- Semantics (meaning)
  - ✓ Values provided are compliant with CVs:
    - ISO19115 Codelist (type, person, organisation)
    - UNDRR/ISC Hazard Information Profiles (keywords)
    - IANA media types (format)
    - SPDX (license)
    - ..



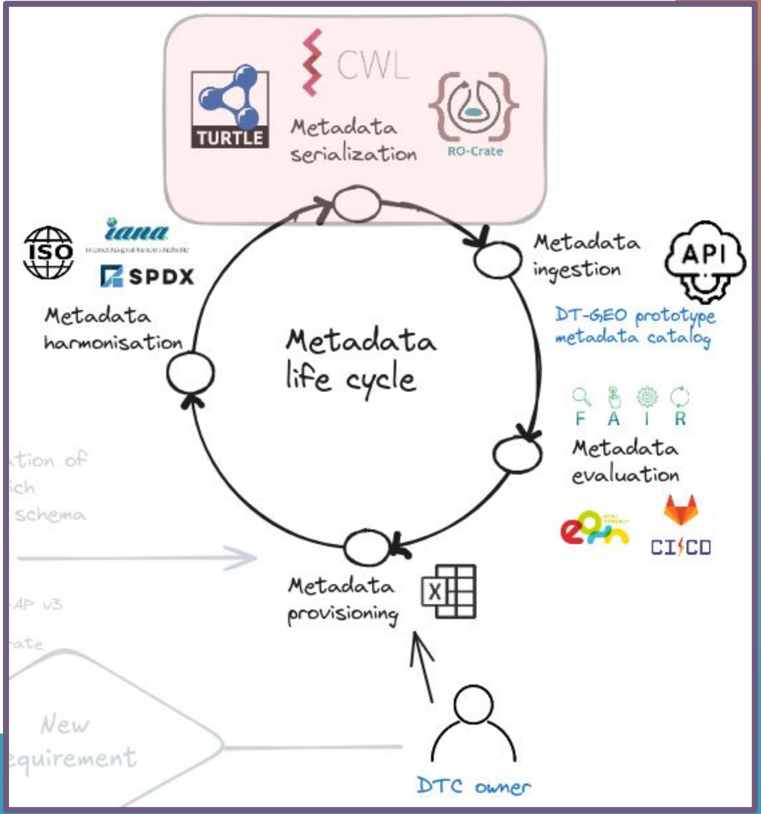
#3.3 Metadata serialisation (life cycle)

Serialisation implies translating the data model into a file format structure, as a previous step before the ingestion

- DTs and SSs: RDF-based **Turtle format (TTL)**
- WFs (and STs): **CWL+RO-Crate**

Files maintained in **Git repositories** (**[M]**anual, **[A]**utomated)**:**

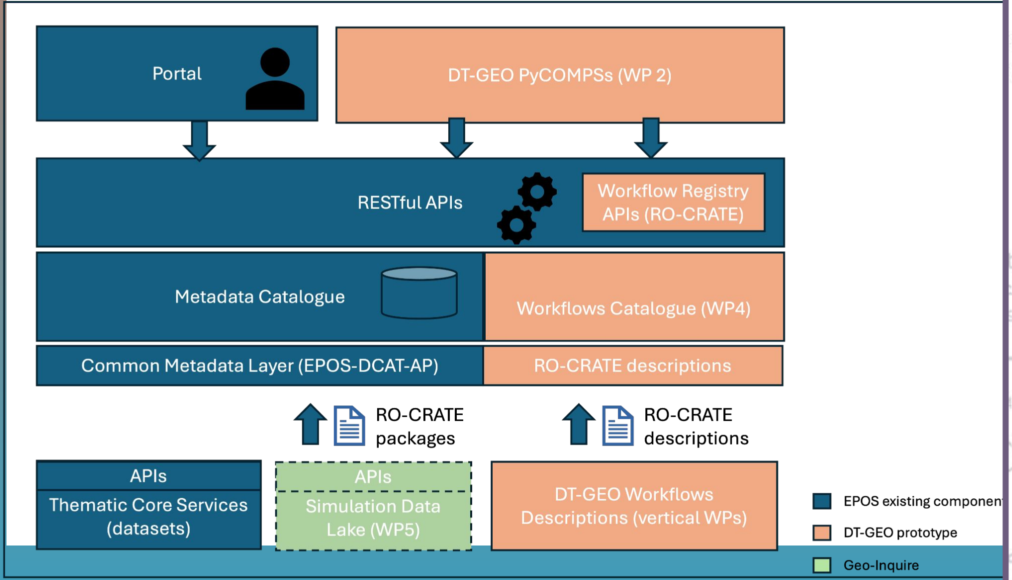
1. **[M]** Peer review of each change
2. **[A]** Validation of syntax
   - SHACL for TTLs
   - JSON-LD for RO-Crates
3. **[A]** Sync with upstream EPOS repositories



#3.4 Metadata ingestion (life cycle)

Metadata ingestion → **DT-GEO prototype metadata catalog**

- **Interfaces: REST API and Web portal**
- **REST APIs are populated by Git repositories**
    - **2 separated APIs: (i) DT+SS and (ii) WF**



Portal

DT-GEO PyCOMPSs (WP 2)

RESTful APIs

Workflow Registry APIs (RO-CRATE)

Metadata Catalogue

Workflows Catalogue (WP4)

Common Metadata Layer (EPOS-DCAT-AP)

RO-CRATE descriptions

RO-CRATE packages

RO-CRATE descriptions

APIs

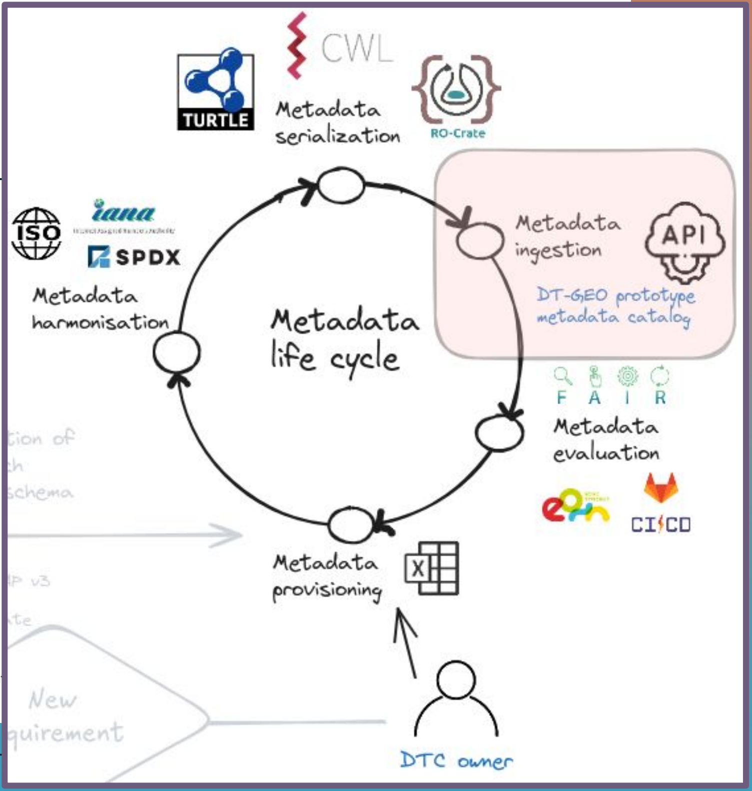
Thematic Core Services (datasets)

APIs

Simulation Data Lake (WP5)

DT-GEO Workflows Descriptions (vertical WPs)

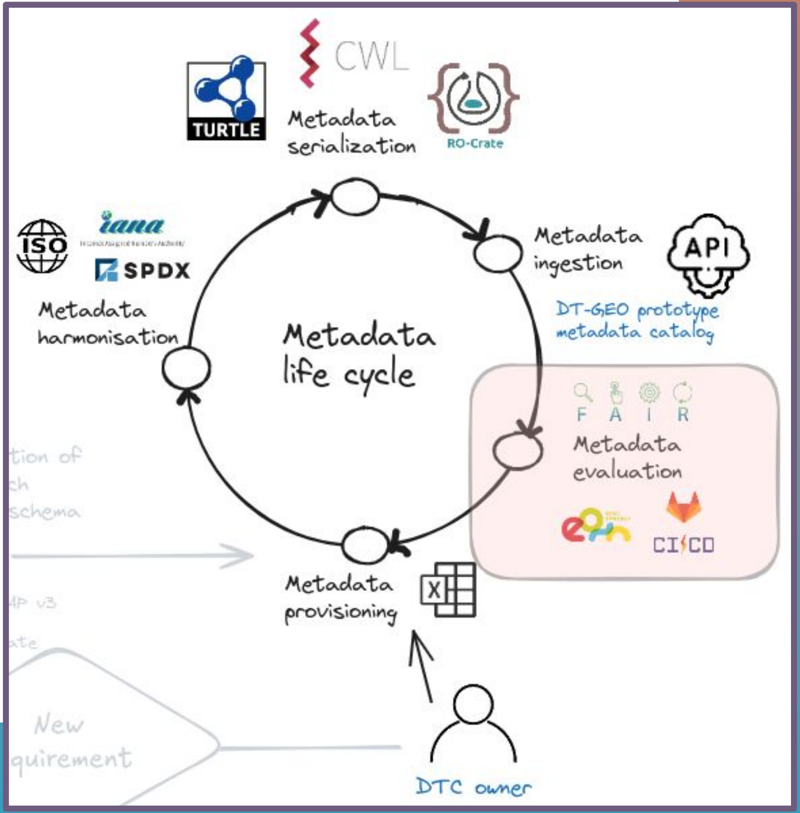
- 🟦 EPOS existing componen
- 🟧 DT-GEO prototype
- 🟩 Geo-Inquire



#3.5 Metadata evaluation (life cycle)

Metadata evaluation is **done in an automated fashion by requesting DT-GEO prototype APIs**

- Data FAIR maturity levels
  - Tool: FAIR-EVA evaluator
  - [Mon 11:30] *“FAIR-EVA : Fair data in the DT_GEO project” (Iván Palomo)*
- Source code QA
  - Tool: SQAaaS
  - [Mon 10:30] *“Mastering the SQAaaS platform: a Software Quality Assurance as a Service tutorial”* (Pablo Orviz, Samuel Bernardo)
- Workflow execution
  - Tools: GitLab CI + Container Image Creation + SQAaaS + PyCOMPSs
  - [Tue 15:00] *“SQAaaS as the quality gate for Digital Twins” (Pablo Orviz)*



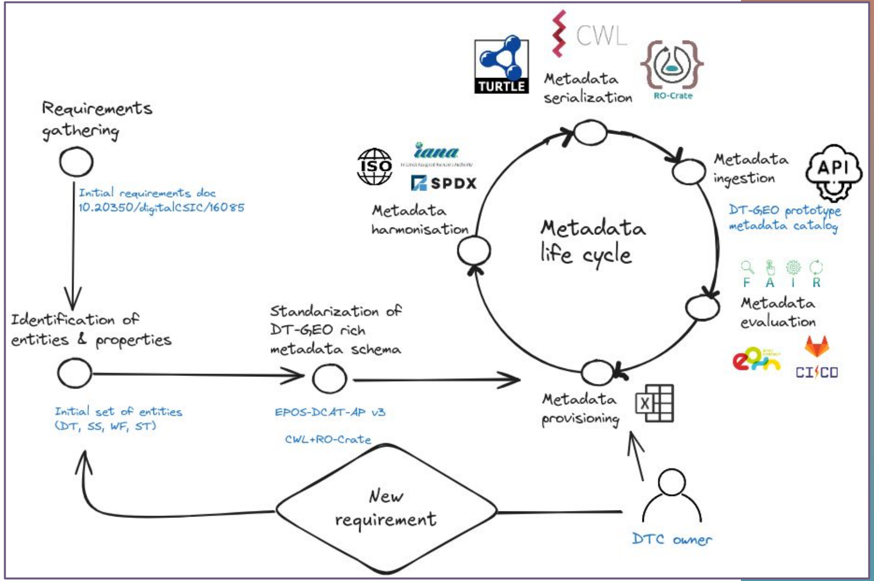
Summary and Highlights

Set up a solution of **continuous improvement of metadata that fully characterises the DTCs**

- Phases: **provision→harmonisation→serialisation→ingestion→evaluation**
- Actors:
  - **DTC owners (coordinators, developers)**
  - **Data Management Team**
  - **Research Infrastructure (EPOS IT)**
- **Extended adaptability:** react to new requirements

DT-GEO prototype metadata catalog

- **Standard-based:** CERIF, EPOS-DCAT-AP v3, CWL+RO-Crate
- **Promoting FAIR & QA**: data (RDA FAIR maturity), code (SQAaaS)
- **Ready for production** ⇒ EPOS ERIC data portal (peer review, automated validation)
- **[in the making] Active population of metadata into WfMS (eFlows4HPC, PyCOMPs) registries and catalogs**



THANK YOU

Q&A

✉ orviz@ifca.unican.es

@dtgeo_eu

linkedin.com/company/dt-geo/

European Commission

This project has received funding from the European Union's Horizon research and innovation programme under the grant agreement No 101058129