Towards an interoperability framework for observable property terminologies

I-ADOPT RDA WG

04 May 2020, 08:30-10:15 CEST

Metadata, Data Models, Semantics, and Collaboration

EGU General Assembly 2020

RESEARCH DATA ALLIANCE
I-ADOPT WG

Mar-2018

Task group formed under Vocabulary Semantic Services Interest Group (VSSIG)

Conceptualisation of measurement parameters - Michael Diepenbroek & Barbara Magagna

Apr-2019

RDA
13th Plenary Meeting
Philadelphia, Pennsylvania USA
2-4 April 2019
Organized by RDA – United States and Drexel University
Drexel

BoF - Harmonizing FAIR descriptions of observational data

New Title of the planned WG:
Interoperability of Observable Property Descriptions WG

Oct-2019

25/10/2019

WG Kick-off meeting
InteroperAble Descriptions of Observable Property Terminology

RDA 14th Plenary Helsinki, I-ADOPT WG introduction by Gwen Moncoiffe

Mar-2020

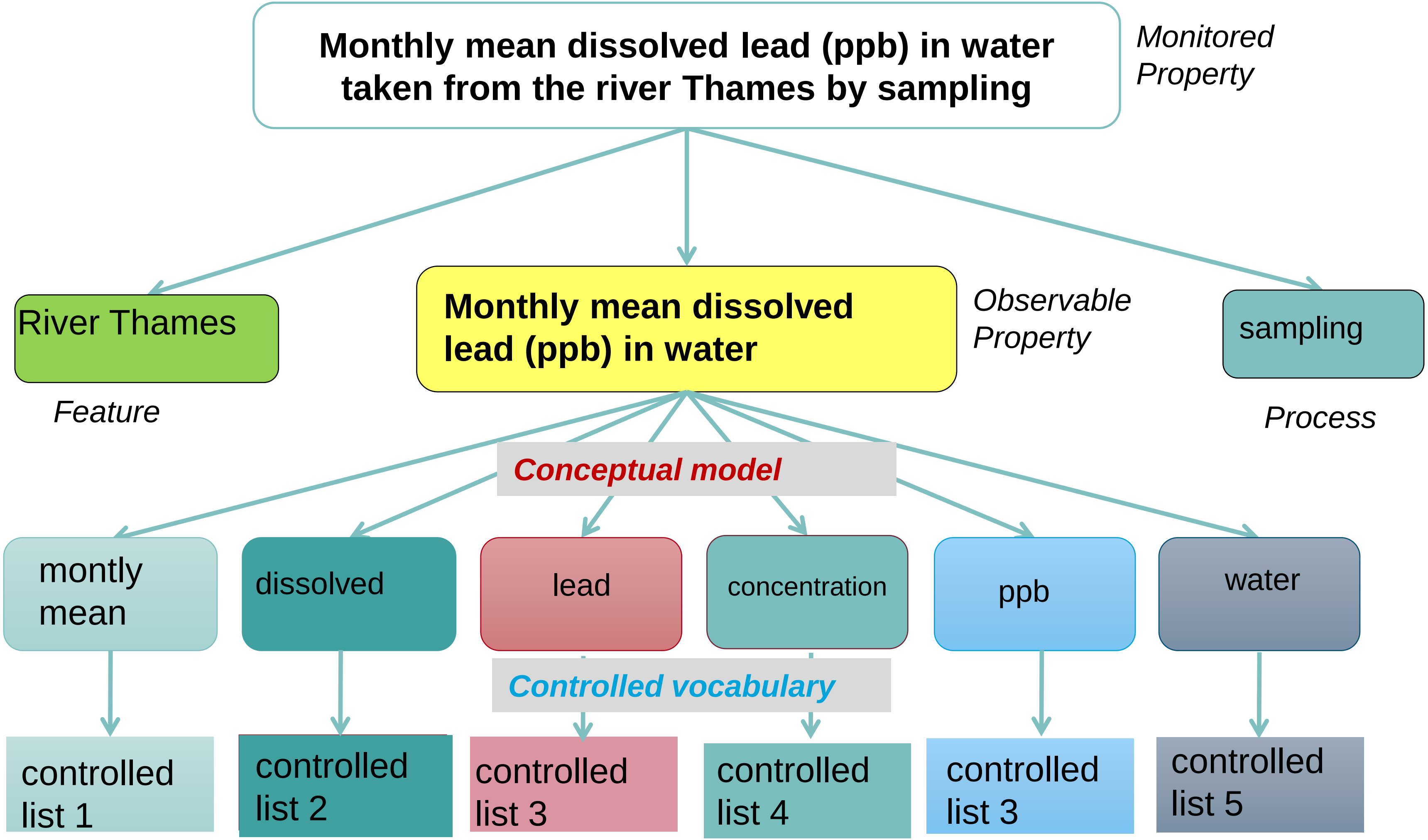
15th Research Data Alliance
Plenary Meeting
Data for Real-World Impact
18-20 MARCH 2020
P15 Virtual Plenary Meeting
from 18 March to 10 April

First I-ADOPT Working Session
RDA 15 Virtual Plenary

Chaired by Barbara Magagna, Gwenaelle Moncoiffe, Michael Diepenbroek, Maria Stoica

What do we mean by observable property?

- Property of the observed object, a natural phenomenon
- The description of what it is and what it represents
- Quantifiable or qualifiable
- Often derived from a representative subset of a feature of interest
  e.g. a physical or digital sample, an individual specimen, a population
- Observed directly or by proxy (modelling/calibration)
- In situ observations, laboratory experiments, remote sensing, modelling
- Also known as “observation type”, “trait”, “variable name”, “parameter”, “measurement“



Monthly mean dissolved lead (ppb) in water taken from the river Thames by sampling

Monitored Property

River Thames

Feature

Monthly mean dissolved lead (ppb) in water

Observable Property

sampling

Process

Conceptual model

montly mean

dissolved

lead

concentration

ppb

water

Controlled vocabulary

controlled list 1

controlled list 2

controlled list 3

controlled list 4

controlled list 3

controlled list 5

Conceptual Models

- Observation and measurement
- OBOE
- Scientific Variable Ontology
- SERONTO
- Complex Property Model
- BODC PUV Semantic model
- Design Patterns for specific parts of the representation of Observable Properties
- Local strategies
- ..

Controlled vocabularies

- SDN vocabularies
- ENVO
- EnvThes
- CHEBI
- OM
- WORMS
- ITIS
- WIGOS
- ..

Challenge: lack of interoperability

Diverse approaches in capturing data semantics:

- At the conceptual level, which model is used to describe the setting of the observation
- At the granularity level, how complex properties are represented
- At the term level, which controlled lists are used to describe what is observed

Motivation

Addressing the “I” of FAIR Data Management

By building a conceptual framework to support interoperability between existing terminologies and address a broad range of known user requirements

By promoting the use of FAIR terminologies to annotate research data with well identified, unambiguous and machine readable vocabularies

I-ADOPT in a nutshell

I-ADOPT will produce an Interoperability Framework for representing observable properties in environmental research (but transferable to other domains)

Task 1: Collect user stories and formalise into use cases	Nov 19 - February 20
Task 2: Survey observation-centric terminologies	Jan 20 - February 20
Task 3: Derive use case requirements	March 20 - May 20
Task 4: Analyse semantic representation of OP against requirements	May 20 - October 20
Task 5: Develop Interoperability Framework	Nov 20 - Feb 21
Task 6: Test local mapping design patterns	March 21 - June 21

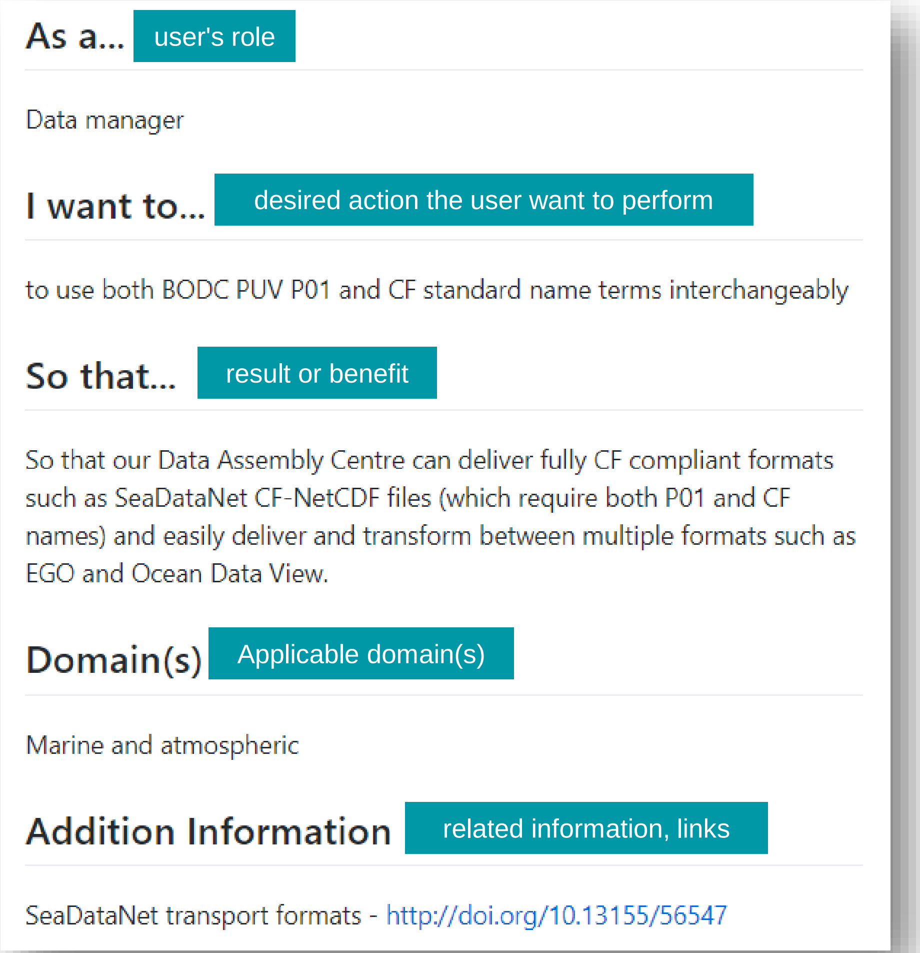
More details to be found in the case statement

Task 1 - user stories and derived use cases

Anusuriya Devaraju
PANGAEA, MARUM - University Bremen, Germany

User Stories

- WG members contributed 19 user stories through Github @ https://github.com/i-adopt/users_stories
- Initial collection period: Nov 2019 – 7th April 2020
- The user stories are not final (will be iteratively improved)
- We welcome new stories!


user's role
desired action the user want to perform
result or benefit
Applicable domain(s)
related information, links

As a...

Data manager

I want to...

to use both BODC PUV P01 and CF standard name terms interchangeably

So that...

So that our Data Assembly Centre can deliver fully CF compliant formats such as SeaDataNet CF-NetCDF files (which require both P01 and CF names) and easily deliver and transform between multiple formats such as EGO and Ocean Data View.

Domain(s)

Marine and atmospheric

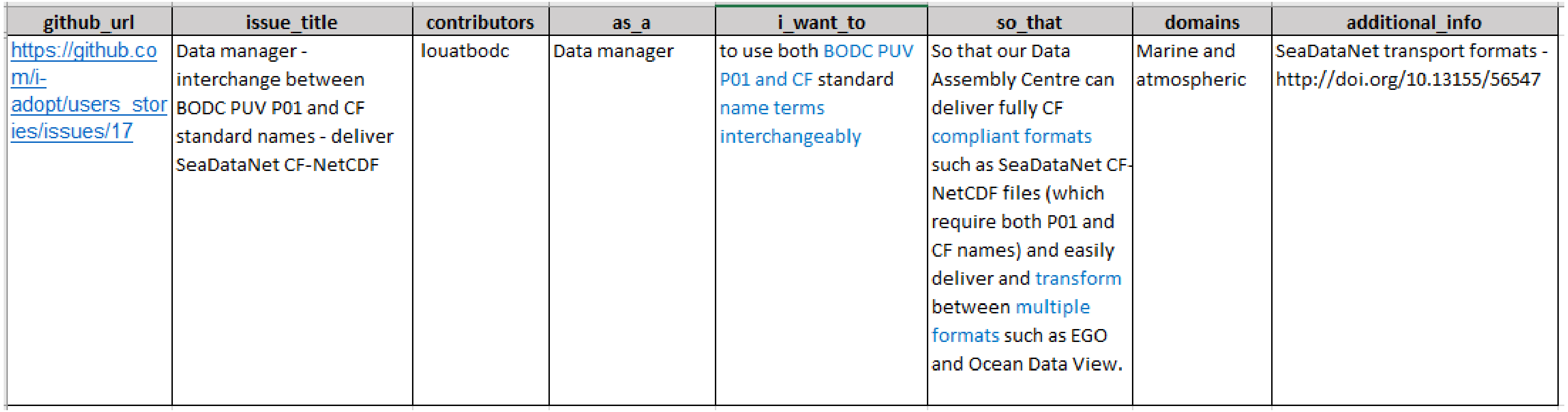
Addition Information

SeaDataNet transport formats - http://doi.org/10.13155/56547

An example of user story, https://github.com/i-adopt/users_stories/issues/17

User Story Analysis (Approach)

Step 1. Label important keywords



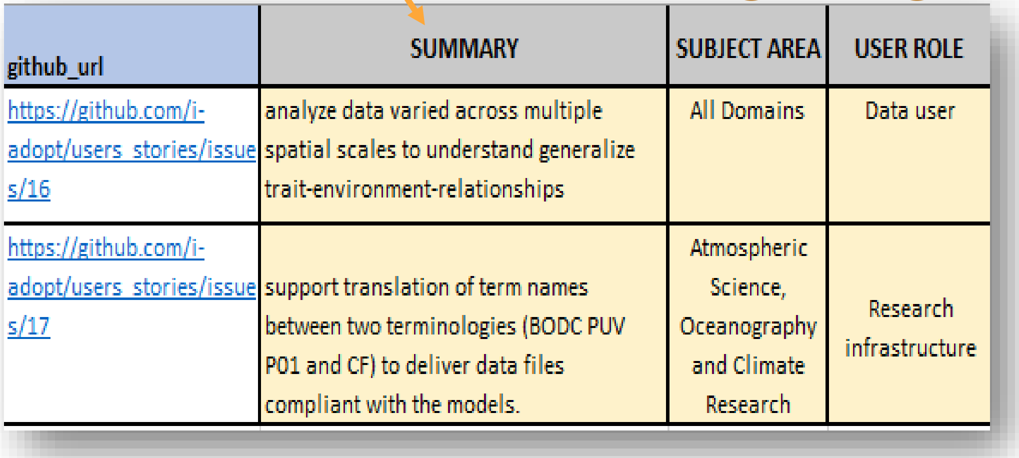
github_url	issue_title	contributors	as_a	i_want_to	so_that	domains	additional_info
https://github.com/i-adopt/users_stories/issues/17	Data manager - interchange between BODC PUV P01 and CF standard names - deliver SeaDataNet CF-NetCDF	louatbodc	Data manager	to use both BODC PUV P01 and CF standard name terms interchangeably	So that our Data Assembly Centre can deliver fully CF compliant formats such as SeaDataNet CF-NetCDF files (which require both P01 and CF names) and easily deliver and transform between multiple formats such as EGO and Ocean Data View.	Marine and atmospheric	SeaDataNet transport formats - http://doi.org/10.13155/56547

User Story Analysis (Approach)

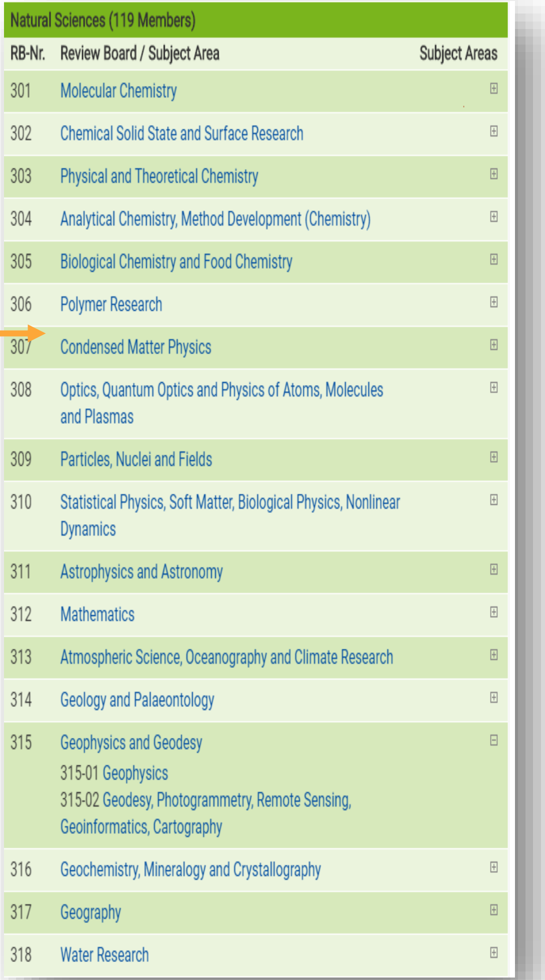
Step 2. Summarize user stories and standardize :

1 Subject Area(s) ([DFG Classification of Subject Area](#))

2 User Role



github_url	SUMMARY	SUBJECT AREA	USER ROLE
https://github.com/i-adopt/users_stories/issues/16	analyze data varied across multiple spatial scales to understand generalize trait-environment-relationships	All Domains	Data user
https://github.com/i-adopt/users_stories/issues/17	support translation of term names between two terminologies (BODC PUV P01 and CF) to deliver data files compliant with the models.	Atmospheric Science, Oceanography and Climate Research	Research infrastructure

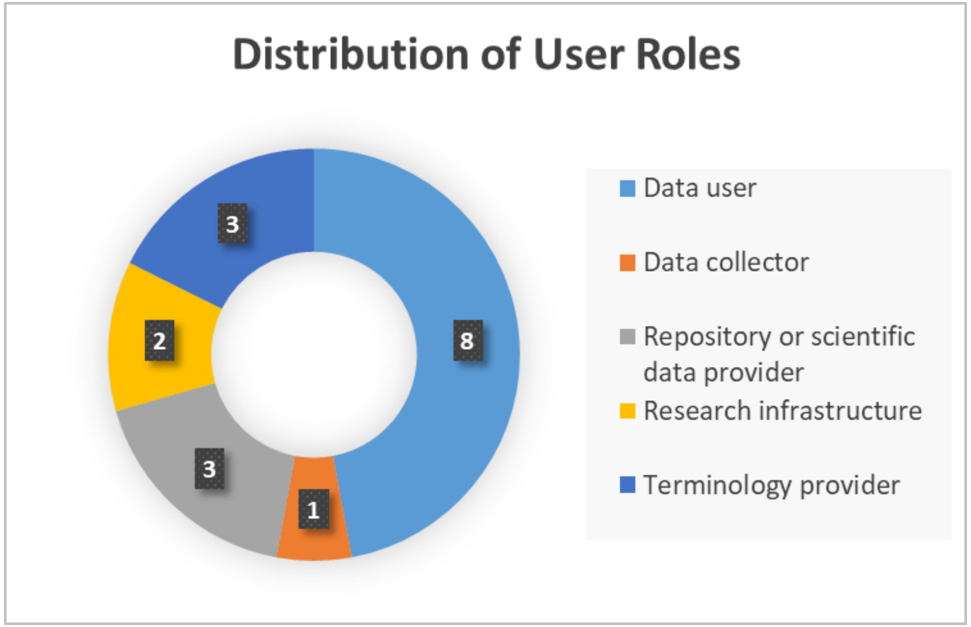


Natural Sciences (119 Members)		
RB-Nr.	Review Board / Subject Area	Subject Areas
301	Molecular Chemistry	⊞
302	Chemical Solid State and Surface Research	⊞
303	Physical and Theoretical Chemistry	⊞
304	Analytical Chemistry, Method Development (Chemistry)	⊞
305	Biological Chemistry and Food Chemistry	⊞
306	Polymer Research	⊞
307	Condensed Matter Physics	⊞
308	Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas	⊞
309	Particles, Nuclei and Fields	⊞
310	Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics	⊞
311	Astrophysics and Astronomy	⊞
312	Mathematics	⊞
313	Atmospheric Science, Oceanography and Climate Research	⊞
314	Geology and Palaeontology	⊞
315	Geophysics and Geodesy 315-01 Geophysics 315-02 Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartography	⊟
316	Geochemistry, Mineralogy and Crystallography	⊞
317	Geography	⊞
318	Water Research	⊞

A partial view of DFG Subject Areas

User Story Analysis (Approach)

User Roles
Data user
Data collector
Repository or scientific data provider
Research infrastructure
Terminology provider



User Story Analysis (Approach)

- Step 3. Derive **use cases** from user stories* and **group** the use cases

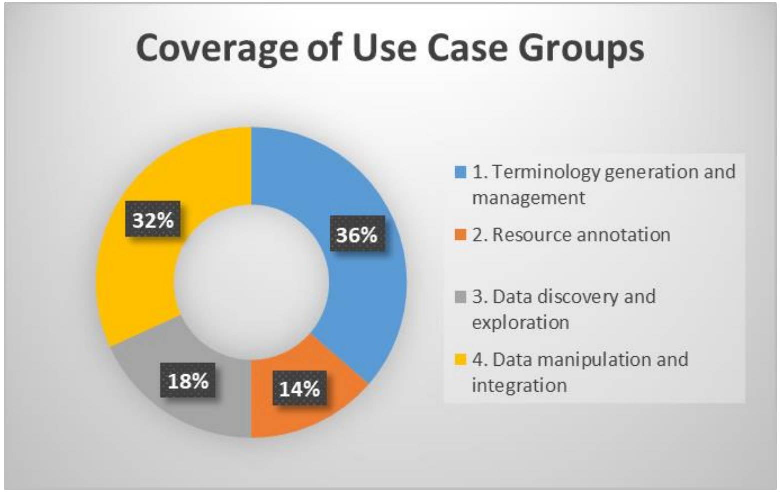
Use Case Groups	Group description	Use Case	Use case descriptions	User Stories
1. Terminology generation and management	This group contains use cases in which the requirements are to generate, curate, align and maintain observable property terminologies.	1.1 Semantic modelling	Develop formal terminologies to represent the concepts being described and the relationships between those concepts.	US3, US13, US14, US18
		1.2 Terminology management	Gather, curate and maintain the individual terms within a terminology.	US1, US15
		1.3 Semantic alignment	Create mappings between terminologies using established relationships; record and preserve the mappings.	US3, US6, US15, US17
		1.4 Terminology search	Search for relevant terminologies and/or terms within terminologies; retrieve the search results.	US5, US6
		1.5 Multilingual concepts	Provide multilingual representations of the concepts within a terminology.	

*Other relevant use cases are also included!

From User Story to Use Case

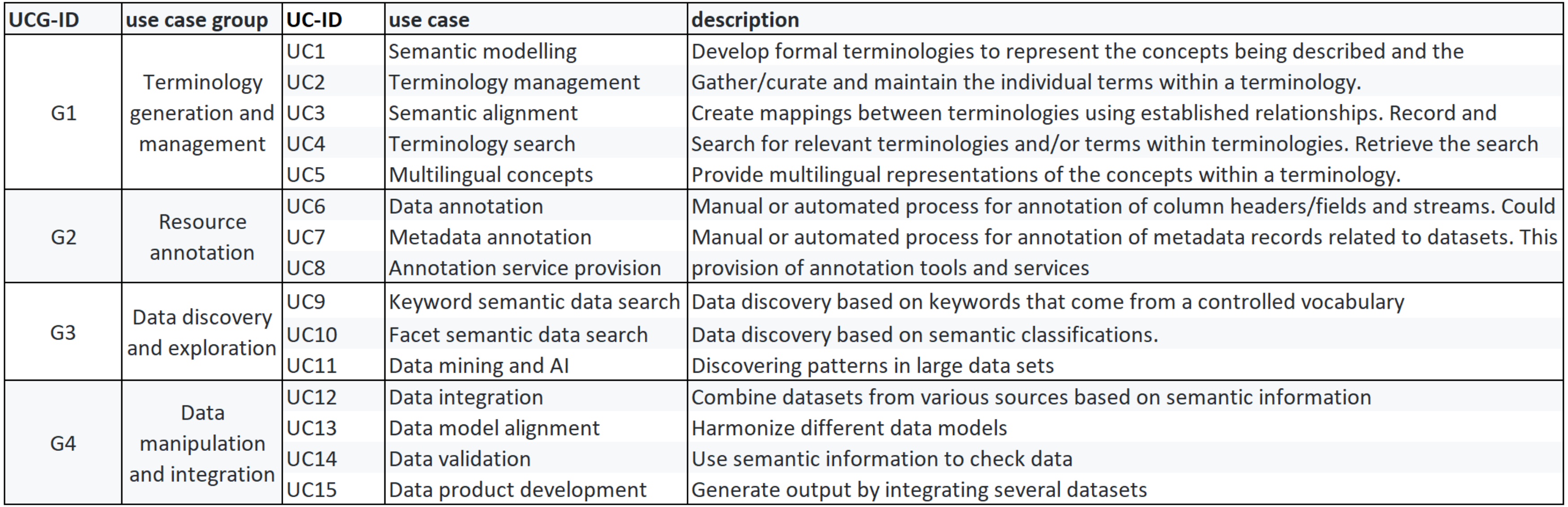
- There are 4 use case groups:

Use Case Groups	Group description	# of use cases
#1 - Terminology generation and management	This group contains use cases in which the requirements are to generate, curate, align and maintain observable property terminologies.	5
#2 - Resource annotation	This group contains use cases that require human and machine-readable identification of observed properties in datasets or parts thereof	3
#3 - Data discovery and exploration	This group contains use cases that require the user to search across multiple sources	2
#4 - Data manipulation and integration	This group contains use cases that require the combination of multiple datasets from various sources	5



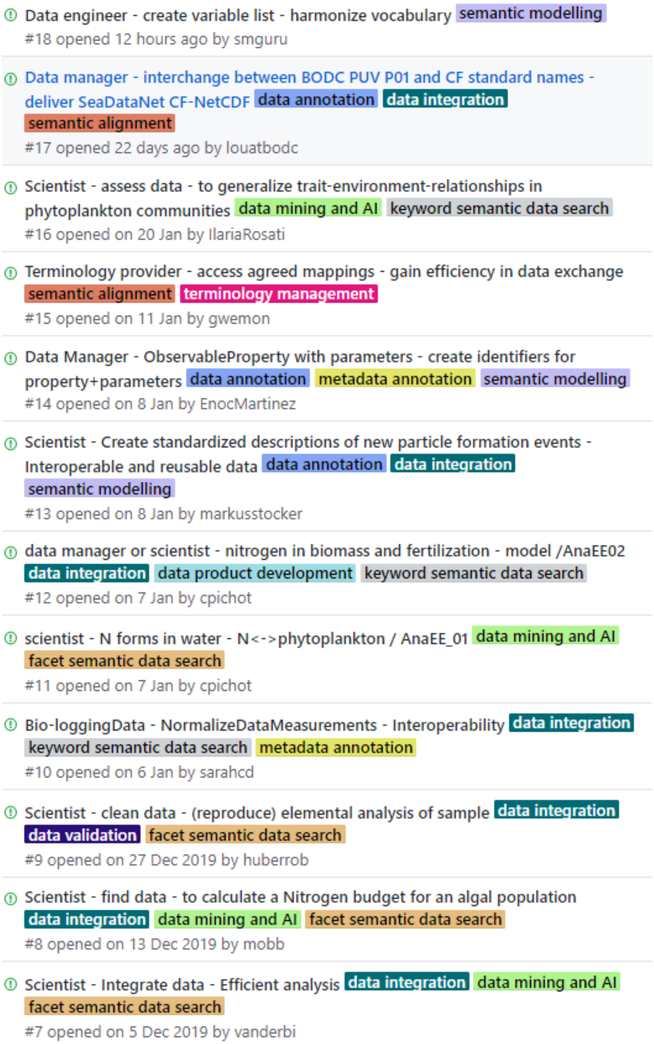
(Note: One user story may belong to one or more use cases)

Overview use cases



UCG-ID	use case group	UC-ID	use case	description
G1	Terminology generation and management	UC1	Semantic modelling	Develop formal terminologies to represent the concepts being described and the
		UC2	Terminology management	Gather/curate and maintain the individual terms within a terminology.
		UC3	Semantic alignment	Create mappings between terminologies using established relationships. Record and
		UC4	Terminology search	Search for relevant terminologies and/or terms within terminologies. Retrieve the search
		UC5	Multilingual concepts	Provide multilingual representations of the concepts within a terminology.
G2	Resource annotation	UC6	Data annotation	Manual or automated process for annotation of column headers/fields and streams. Could
		UC7	Metadata annotation	Manual or automated process for annotation of metadata records related to datasets. This
		UC8	Annotation service provision	provision of annotation tools and services
G3	Data discovery and exploration	UC9	Keyword semantic data search	Data discovery based on keywords that come from a controlled vocabulary
		UC10	Facet semantic data search	Data discovery based on semantic classifications.
		UC11	Data mining and AI	Discovering patterns in large data sets
G4	Data manipulation and integration	UC12	Data integration	Combine datasets from various sources based on semantic information
		UC13	Data model alignment	Harmonize different data models
		UC14	Data validation	Use semantic information to check data
		UC15	Data product development	Generate output by integrating several datasets

User stories in github



Data engineer - create variable list - harmonize vocabulary semantic modelling
#18 opened 12 hours ago by smguru

Data manager - interchange between BODC PUV P01 and CF standard names - deliver SeaDataNet CF-NetCDF data annotation data integration semantic alignment
#17 opened 22 days ago by louatbodc

Scientist - assess data - to generalize trait-environment-relationships in phytoplankton communities data mining and AI keyword semantic data search
#16 opened on 20 Jan by IlariaRosati

Terminology provider - access agreed mappings - gain efficiency in data exchange semantic alignment terminology management
#15 opened on 11 Jan by gwemon

Data Manager - ObservableProperty with parameters - create identifiers for property+parameters data annotation metadata annotation semantic modelling
#14 opened on 8 Jan by EnocMartinez

Scientist - Create standardized descriptions of new particle formation events - Interoperable and reusable data data annotation data integration semantic modelling
#13 opened on 8 Jan by markusstocker

data manager or scientist - nitrogen in biomass and fertilization - model /AnaEE02 data integration data product development keyword semantic data search
#12 opened on 7 Jan by cpichot

scientist - N forms in water - N<->phytoplankton / AnaEE_01 data mining and AI facet semantic data search
#11 opened on 7 Jan by cpichot

Bio-loggingData - NormalizeDataMeasurements - Interoperability data integration keyword semantic data search metadata annotation
#10 opened on 6 Jan by sarahcd

Scientist - clean data - (reproduce) elemental analysis of sample data integration data validation facet semantic data search
#9 opened on 27 Dec 2019 by huberrob

Scientist - find data - to calculate a Nitrogen budget for an algal population data integration data mining and AI facet semantic data search
#8 opened on 13 Dec 2019 by mobb

Scientist - Integrate data - Efficient analysis data integration data mining and AI facet semantic data search
#7 opened on 5 Dec 2019 by vanderbi

Task 2: Annotation practices - observable property models and terminologies in use

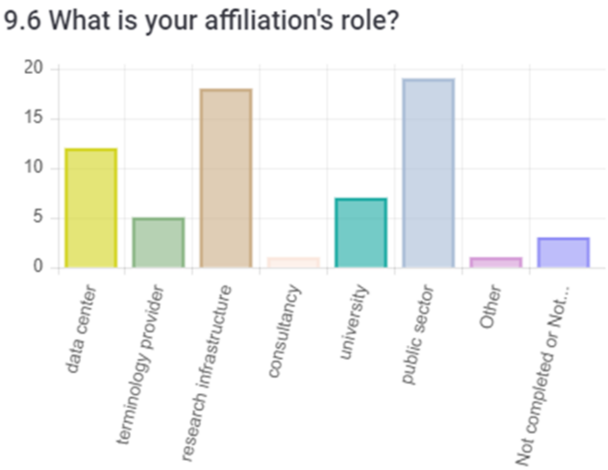
Gwen Moncoiffé
British Oceanographic Data Centre
National Oceanography Centre
United Kingdom

Results of survey *(ongoing!)*

- 33 valid responses received between 23 January - 01 March 2020
  - 21 were both consumers and providers of terminologies
  - 6 were consumers only
  - 6 were providers only
- 25 terminologies >> catalogue
- Mainly english language
  - some bi-lingual French/English terminologies
  - some supporting multilingual translations

Work in progress!
→ Preliminary results

Responders' affiliations



9.6 What is your affiliation's role?

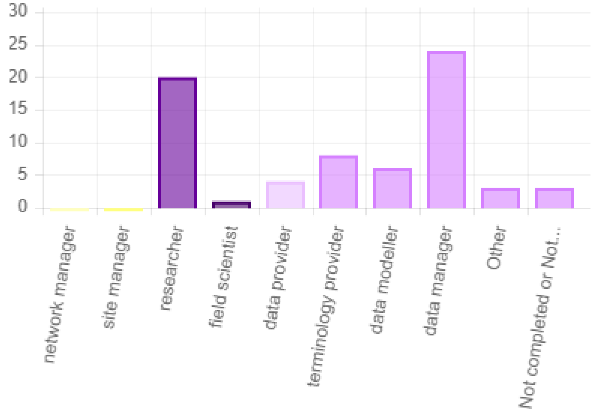
→ RIs and data centres

→ Geographic coverage mainly English speaking and European countries

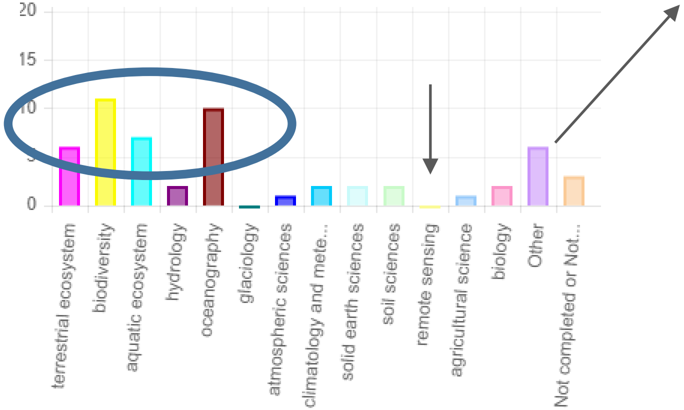
Countries	# submissions All
Australia	2
Austria	1
Canada	1
France	8
French polynesia	1
Germany	2
International	1
Ireland	1
Italy	6
Norway	1
South Africa	1
UK	3
USA	5

Profile of responders

9.2 Which of the following describes your job best?



9.3 If you are a researcher, in which research domain(s) do you operate?



computer science and informatics (x4)
ecology
forest sciences and genetics

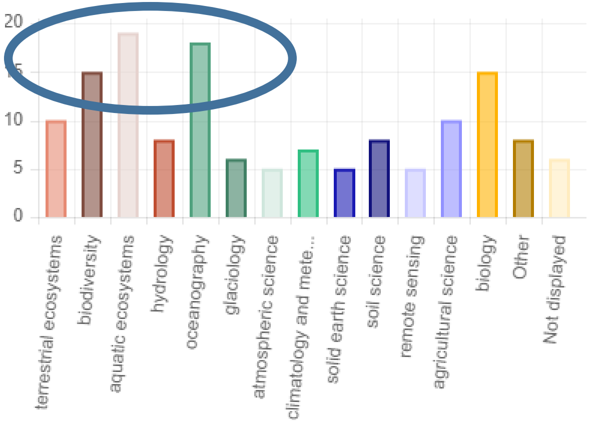
- researchers and data managers in about equal representation
- All except 1 agreed to being contacted/mentioned

- terrestrial and aquatic ecosystems, oceanography and biodiversity

Domain coverage

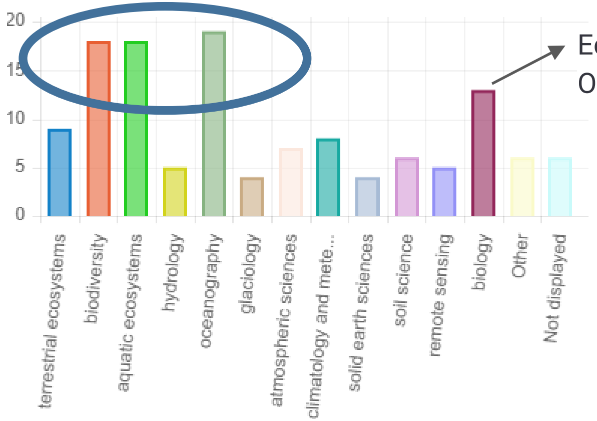
Terminologies

2.5 Which domain(s) is/are the terminology representing?



Consumers

4.3 Which domain is the data you are working with representing?



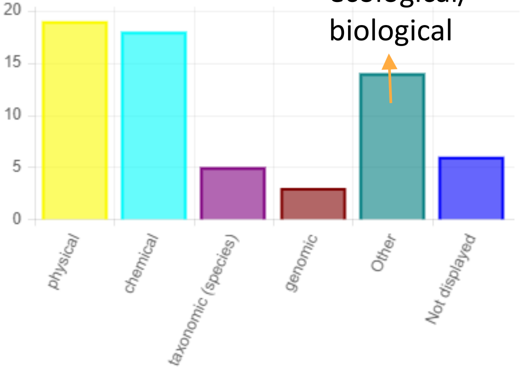
Economic
Ocean contaminants

- → very similar distributions for providers and consumer
- → strong representation from biology / biodiversity / aquatic ecosystems / oceanography
- → many terminologies are described as being multidisciplinary (or “generic”)

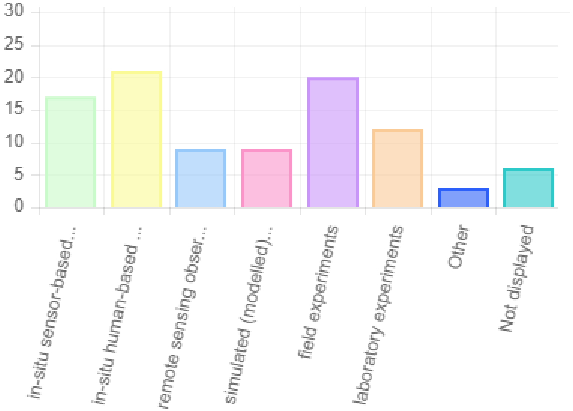
Observations types supported by existing terminologies

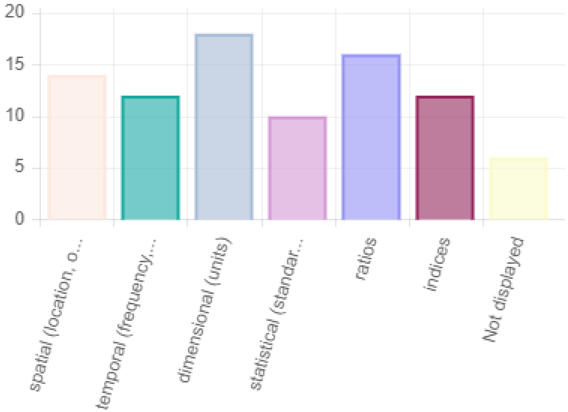
Results for terminologies

Very similar distributions for repositories


3.1 Which of these cross-domain concepts does your terminology support?

ecological/ biological

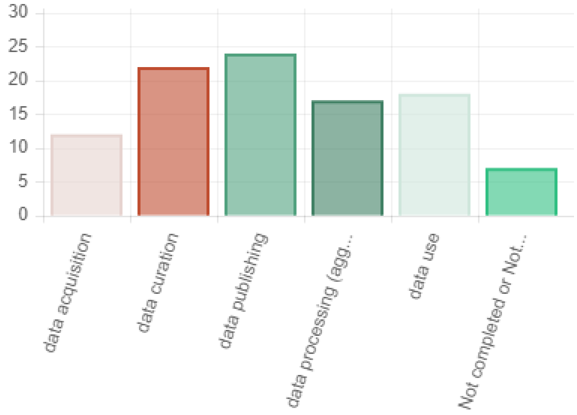

3.2 Which of these types of observations does your terminology support?


3.3 Does your terminology contain concepts related to the f type of properties or quantities?

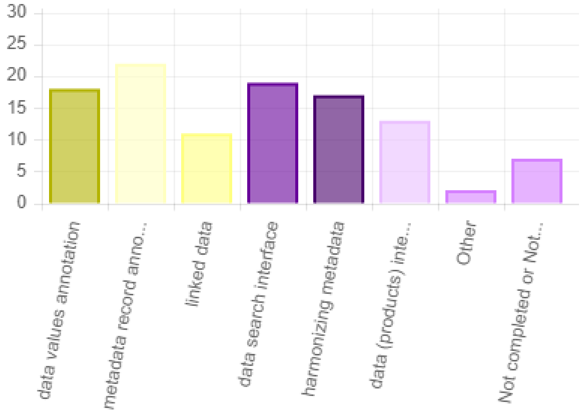
3.4 All but 1 supported both quantitative and qualitative observations

Relation to data life cycle and main purpose of the terminologies

Data Life Cycle phases (ENVRI) 5.1 At which phase of data life cycle do you use terminologies for observable properties?

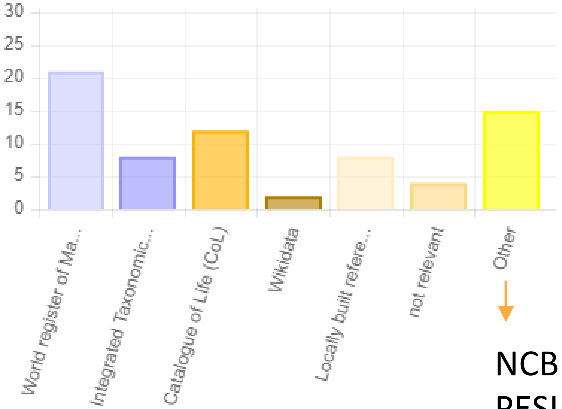


5.2 For which purpose do you use observable property terminologies?



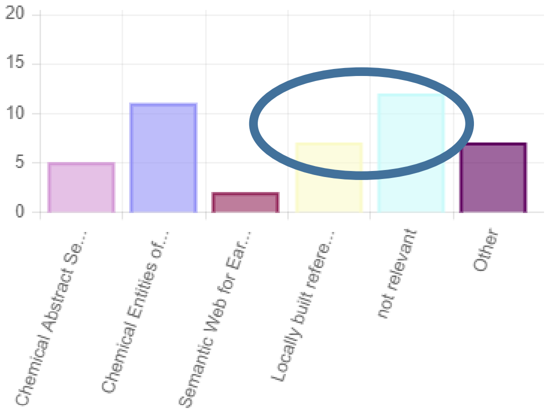
External reference for biological and chemical names

6.1 Which registry of biological taxonomy do you use, if any?



NCBI taxonomy x6
PESI x3

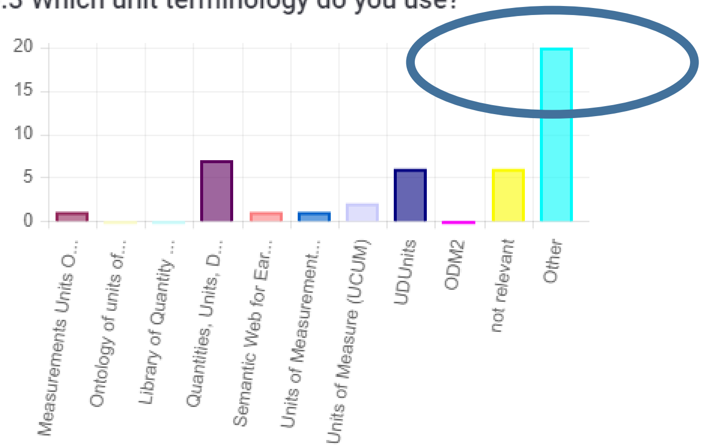
6.2 To what chemical database(s) do you refer for the chemical substance name?



- → WoRMS and ChEBI well used when applicable
- → Locally built reference list counts is high for chemical substances and moderate for biological names
- → Opportunity to look at use of common reference lists

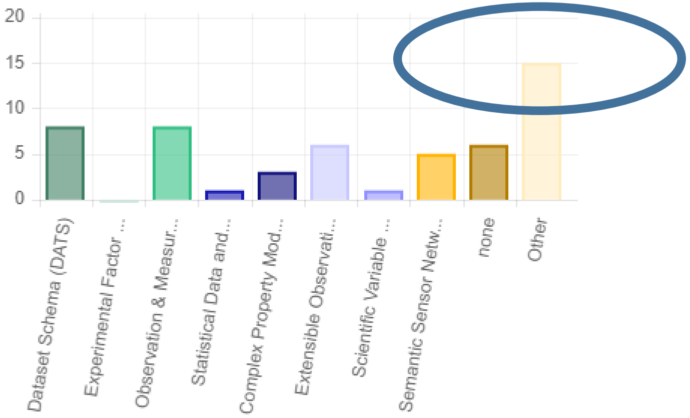
Units and conceptual models

6.3 Which unit terminology do you use?



- → QUDT and UDUnits
- → Other: NVS P06 (x3), OBOE units (x7), partial QUDT (x2)

7.1 What semantic or conceptual model(s) if any do you use to describe your data?



- → DATS, O&M, OBOE, and none
- → Other: at least 10 mentioned

Work in progress!

Task 3: Requirements

first ideas

Barbara Magagna
Environment Agency Austria

Requirements

What does a **terminology** need to **provide**
to **support** a given **use case**?
(“We can do this with that model.”)

Goal:

- to test the suitability of existing models/ontologies and
- a set of requirements = requirement specification for the interoperability framework

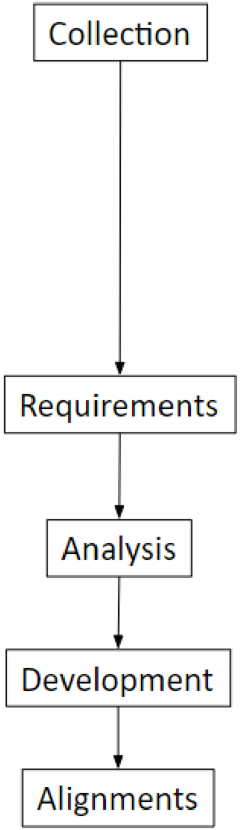
Requirements

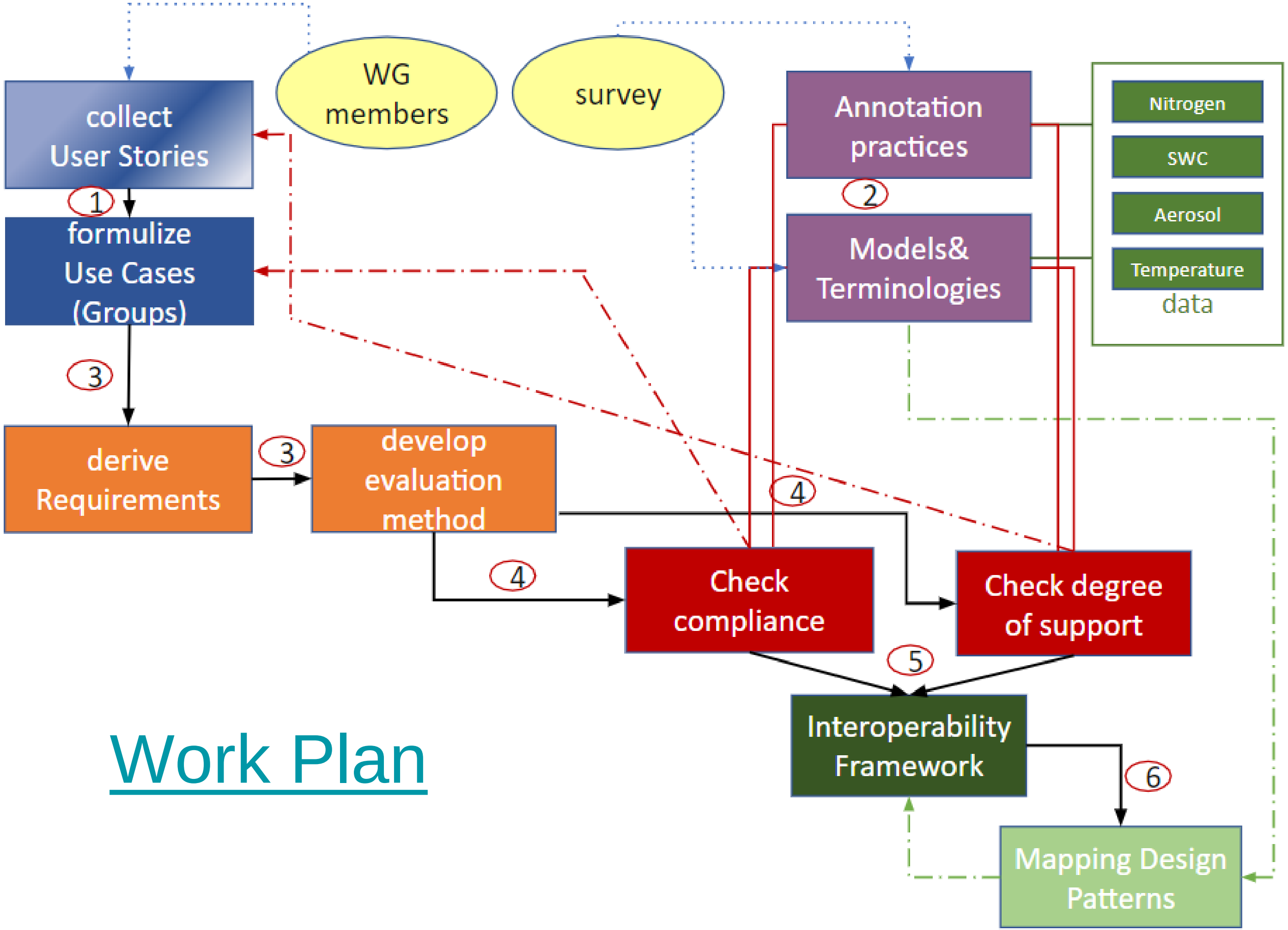
For each use case we aim to collect necessary and optional requirements

Necessary requirement: required features, if missing, the model fails to support the use case

Optional requirement: not necessary features that simplify the implementation of a use case or increase its usefulness

Work Plan


Collection
Requirements
Analysis
Development
Alignments


WG members
survey
collect User Stories
1
formulize Use Cases (Groups)
3
derive Requirements
3
develop evaluation method
4
4
Check compliance
Check degree of support
5
Interoperability Framework
6
Mapping Design Patterns
Annotation practices
2
Models& Terminologies
Nitrogen
SWC
Aerosol
Temperature
data

Road Map

Requirements analysis:

1. agree on use case definitions and involved actors
   - ask user story contributors to check the allocation to use cases
2. define requirements for each use case (necessary/optional)
   - ask user story contributors to check requirements
3. develop evaluation method for the requirements analysis
4. analyse suitability for each pair of OP model and use case
5. analyse degree of support for each pair of OP model and user story

Validation:

1. ask user story contributors for data sets to support the requirement analysis
2. select first N-based user stories with datasets for the analysis
3. validate analysis results with actual datasets and use cases

Evaluation method for requirement analysis

- check methodologies in ontology engineering
  - competency questions
  - metrics
  - … etc.
- decide which methodology to apply
- develop/adopt methodology for I-ADOPT

Working Group modalities

- 2 telcos per month:
  - first Thursday at 18:00 CEST (US-friendly)
  - third Tuesday at 10:00 CEST (Australia-friendly)
- Material to be found in Google Drive
- Ongoing work to be followed in GitHub

Want to participate and contribute?

- subscribe to I-ADOPT
- check the I-ADOPT WIKI
- visit us in Twitter
- contribute a user story in GitHub
- participate at the I-ADOPT survey about terminologies and annotation practices of observable properties

Thank you very much!