Metadata Best Practices Illustrated with the CEDAR Workbench

November 2018

John Graybeal
jgraybeal@stanford.edu

BMIR
Stanford Center for
Biomedical Informatics Research

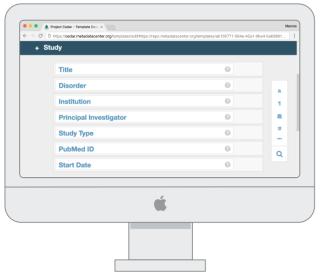
What is the CEDAR Workbench?

CEDAR

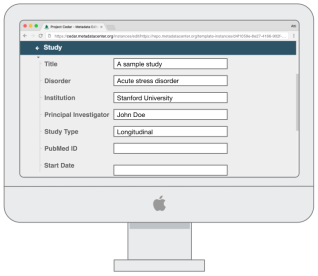
metadatacenter.org

CEDAR Workbench

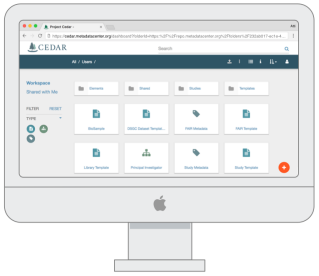
CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL



Template Designer



Metadata Editor

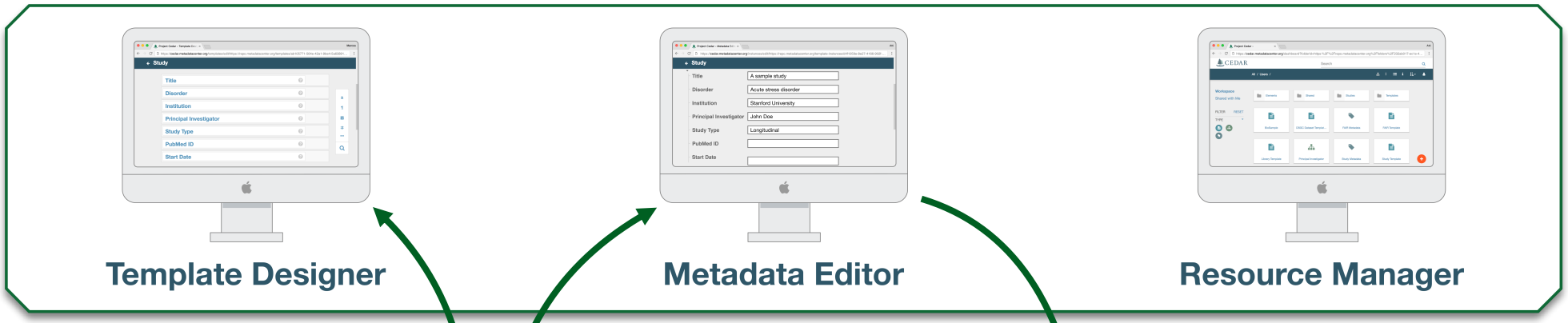


Resource Manager

1. **A Template Designer to create forms.**
2. **A Metadata Editor to fill out those forms.**
3. **A Resource Manager to manage the forms and metadata.**

CEDAR Workbench

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL



Terms

Metadata

BioPortal

NCBO BioPortal
Open Services

Public Repositories

db1 db2 . . . dbn

With semantic services (vocabularies) from BioPortal

And APIs to access metadata remotely or submit them to external repositories

Our Best Practices

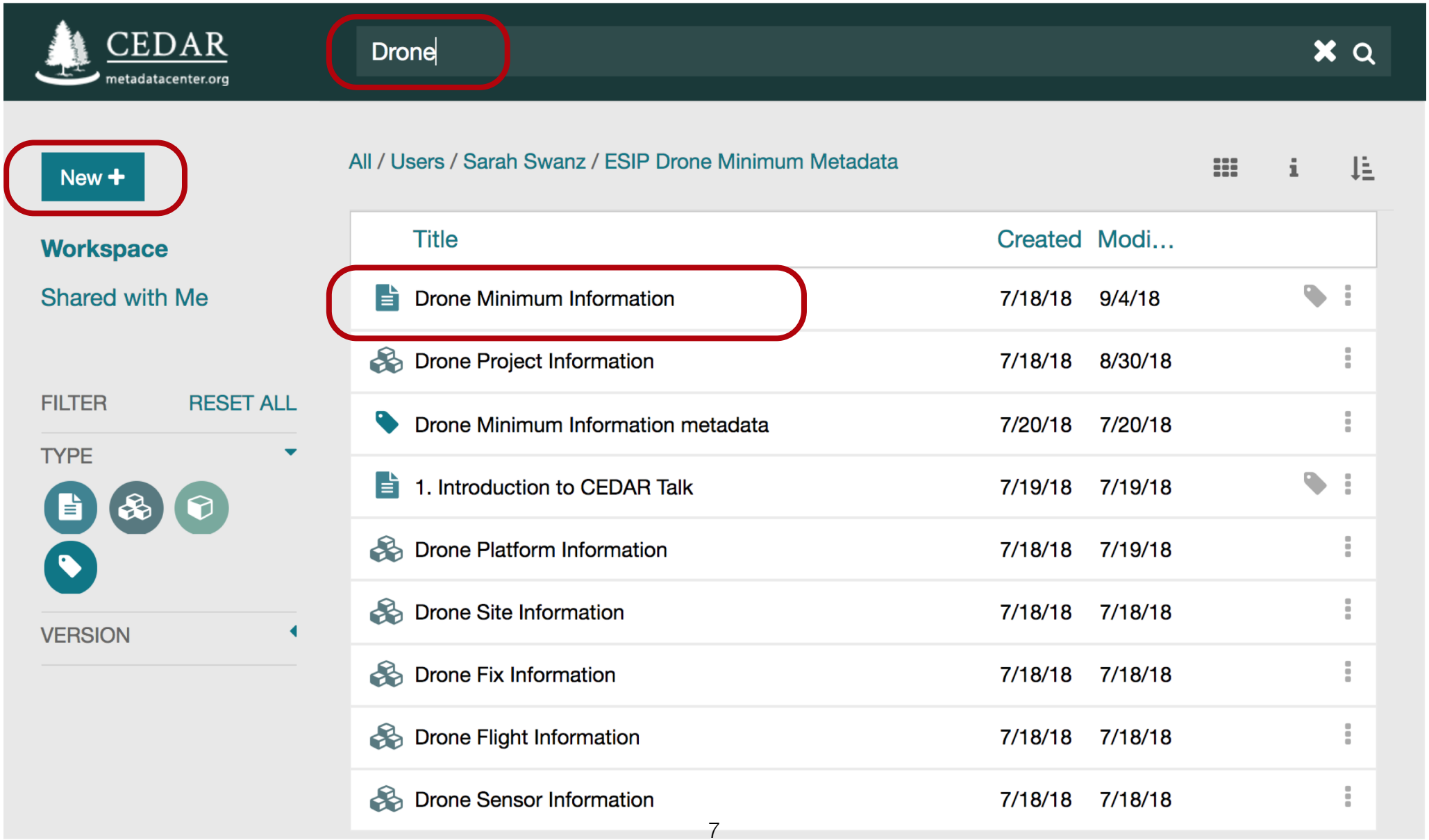
1. Quickly target your team's metadata standards
2. Make metadata entries consistent and accurate
3. Enter and verify metadata as quickly and easily as possible
4. Drive search with well-defined vocabularies and mappings

BP1: Quickly target your team's metadata standards

- GIVEN: A defined standard, community practice, or external requirement for metadata content
- GOAL: Quickly set up a web service that lets teams enter and verify metadata that meets the specification
- EXAMPLE: Minimal metadata that must meet project requirements and be submitted to an external repository
- APPROACH: Define a metadata form satisfying your metadata content using CEDAR's Template Builder.
- EXTRA BENEFITS:
  - A sharable computable specification in JSON Schema
  - Support for manual and automated metadata entry.

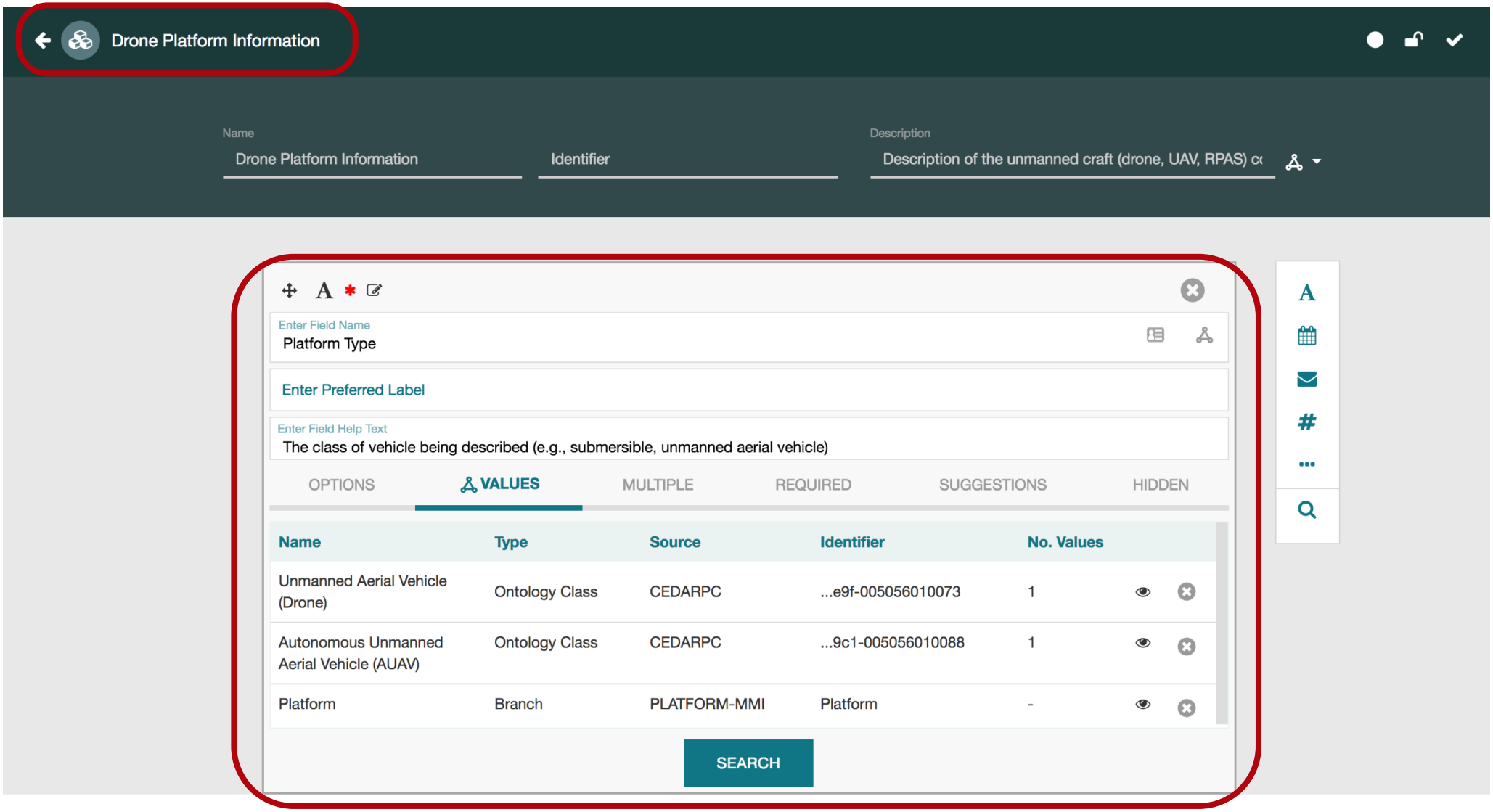
BP1: Target your metadata standards

Find a similar template, or create your own

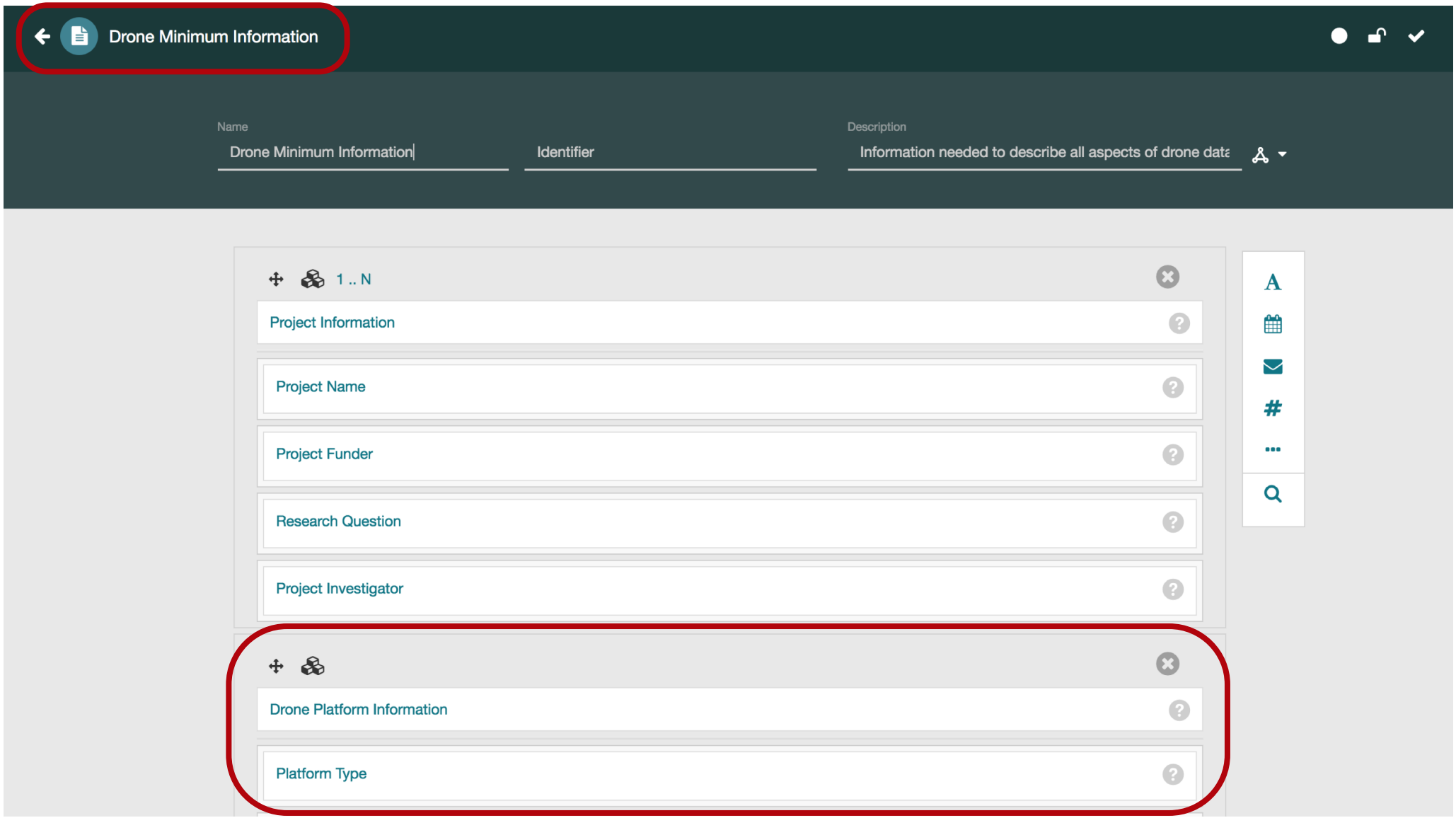


BP1: Target your metadata standards

Create fields and elements that match your needs

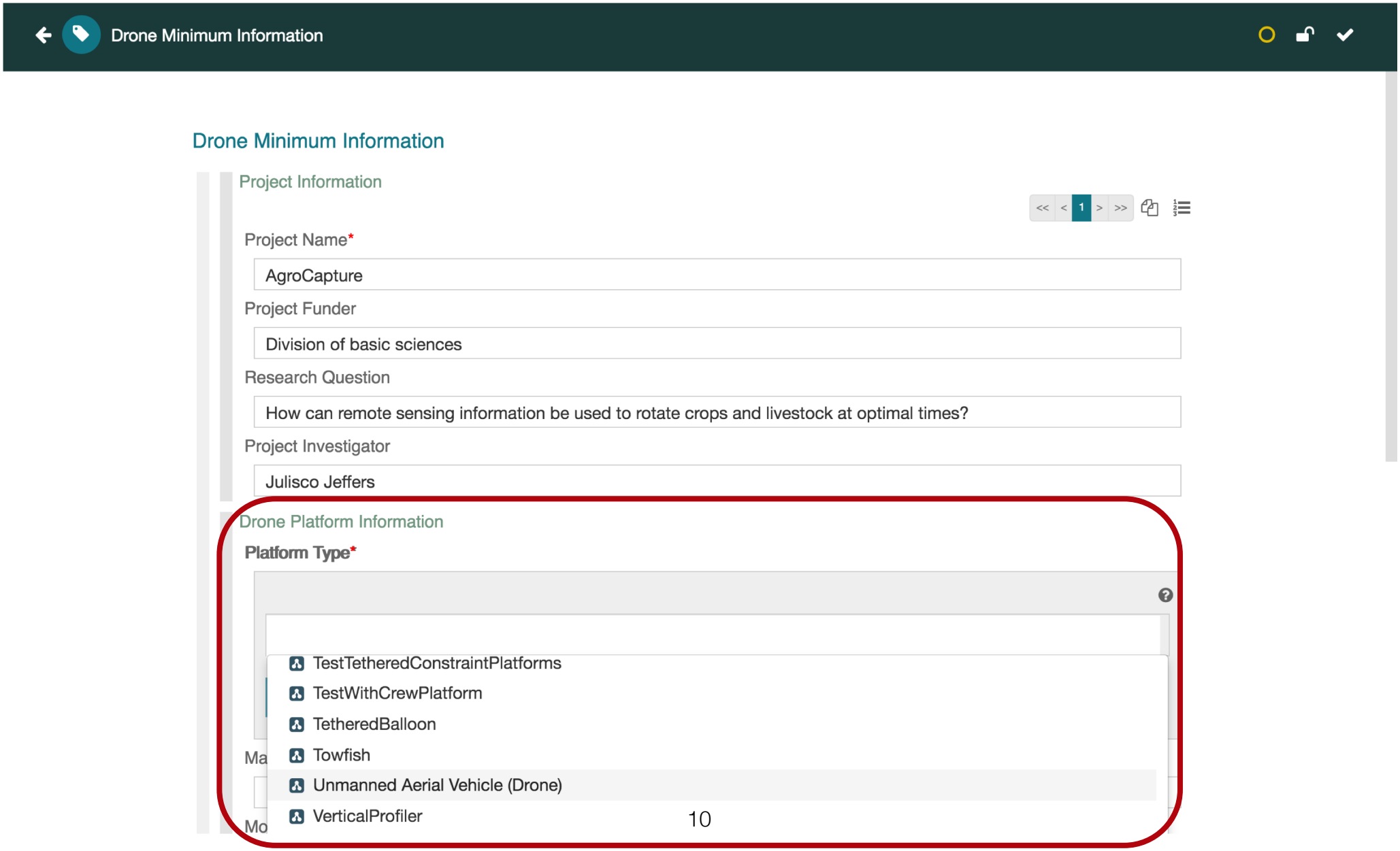


Include fields and elements in your template



BP1: Target your metadata standards

Start collecting your metadata



BP2: Make metadata entries consistent and accurate

- GIVEN: Complicated field values that must be exactly right
- GOAL: Get the metadata entered correctly
- EXAMPLE: Data product descriptions using complex terms
- APPROACH: Semantic Terms from Controlled Vocabularies, Auto-completion, Field Validation, Field Tips
- EXTRA BENEFITS:
  - Interoperability with semantic web (JSON-LD or RDF)
  - Early confirmation of many typographic errors
  - Less experienced users more confident in their ability to enter good metadata, and more motivated to do so.

BP3: Enter & verify metadata as quickly and easily as possible

- GIVEN: Many assets requiring a lot of metadata entries
- GOAL: Enter metadata quickly with minimal pain
- EXAMPLE: Describe 40 similar files using complex values
- APPROACH: Ordered Controlled Terms, Suggestions, and an 'Instance example' (with Hidden fields)
- EXTRA BENEFITS:
  - Obtain benefit from earlier work by other contributors
  - Ability to blend automated, manual metadata entry
  - Can include 'provenance fields' in each filled-out form

BP4: Drive search with well-defined vocabularies and mappings

- GIVEN: Metadata from varied sources using terms relatable to other terms (e.g., synonym or parent/child relations)
- GOAL: To find all applicable matches across term sets
- EXAMPLE: Data in Google's Data Search is indexed with terms from GCMD, CF, and SWEET. Find all data that includes air temperature.
- APPROACH: Make sure that keyword and parameter description fields in CEDAR templates require selection from well-known controlled vocabularies (or vocabularies mapped to them).
- EXTRA BENEFITS:
  - Meaning of concepts in metadata sources is defined.
  - Value of vocabularies enhanced by usage and mappings.

CEDAR References

- Sign up and use CEDAR: **https://cedar.metadatacenter.org**
- Learn about CEDAR: **https://metadatacenter.org**
- More CEDAR references: **https://metadatacenter.org/refs**
- On GitHub (and social media) at **metadatacenter**