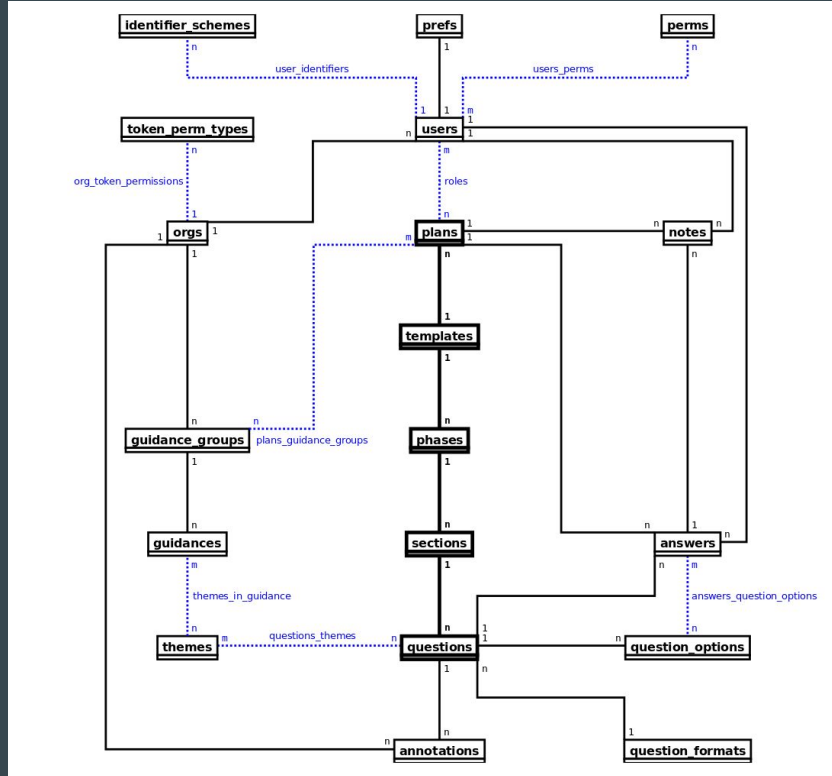


Implementing maDMPS in DMPonline



Sam Rust - Systems Developer, DCC
31 July 2019

DMPRoadmap Data Model



See: <https://github.com/DMPRoadmap/roadmap/wiki/Data-model> For a full-sized image

Targeting a minimal DMP

Required objects are a DMP, with at least one Contact and Dataset

```
1  {
2  "dmp": {
3    "title": "Minimal DMP",
4    "language": "en",
5    "created": "2018-07-23T10:10:23.6",
6    "modified": "2019-02-06T10:10:23.6",
7    "ethical_issues_exist": "unknown",
8    "contact": {
9      "mail": "cc@example.com",
10     "name": "Charlie Chaplin",
11     "contact_id": {
12       "contact_id": "http://orcid.org/0000-0000-0000-0000",
13       "contact_id_type": "HTTP-ORCID"
14     }
15   },
16   "dataset": [
17     {
18       "title": "Placeholder dataset",
19       "type": "(value from dictionary)",
20       "personal_data": "unknown",
21       "sensitive_data": "unknown"
22     }
23   ]
24 }
25 }
```

RDA-DMP-Common-Standard - Minimal DMP:

<https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/examples/JSON/ex8-dmp-minimal-content.json>

What We Already Have From DMPRoadmap

- DMP
 - ◆ Title from Plan's title
 - ◆ Description from Plan's description
 - ◆ Creation/modification dates from Plan
 - ◆ Language from plan owner's language
- Contact
 - ◆ Plan's owner's name, email, and ORCID identifier

```
{  
  "dmp": {  
    "title": "Example DMP",  
    "description": "An in-depth description",  
    "language": "en",  
    "created": "2018-07-23T10:10:23.6",  
    "modified": "2019-02-06T10:10:23.6",  
    "contact": {  
      "mail": "cc@example.com",  
      "name": "Charlie Chaplin",  
      "contact_id": {  
        "contact_id": "http://orcid.org/0000-0000-0000-0000",  
        "contact_id_type": "HTTP-ORCID"  
      }  
    }  
  }  
}
```

Making a Few Assumptions to get a Minimal DMP

→ Datasets

- ◆ DMPRoadmap currently does not have the concept of a Dataset
- ◆ Let's assume that all answers correspond to a single dataset
 - Title from Plan's title
 - Type from the most generic at http://vocabularies.coar-repositories.org/puby/resource_type.html
 - Personal and sensitive data collection can safely be set as 'unknown'

→ DMP

- ◆ Ethical issues can safely be set to 'unknown'

```
{
  "dmp": {
    "title": "Example DMP",
    "description": "An in-depth description",
    "ethical_issues_exist": "unknown",
    "language": "en",
    "created": "2018-07-23T10:10:23.6",
    "modified": "2019-02-06T10:10:23.6",
    "contact": {
      "mail": "cc@example.com",
      "name": "Charlie Chaplin",
      "contact_id": {
        "contact_id": "http://orcid.org/0000-0000-0000-0000",
        "contact_id_type": "HTTP-ORCID"
      }
    }
  }
  "dataset": {
    "title": "Example DMP Dataset"
    "type": "dataset"
    "personal_data": "unknown",
    "sensitive_data": "unknown"
  }
}
```

Comparing to a Minimal maDMP

```
1 {
2   "dmp": {
3     "title": "Minimal DMP",
4     "language": "en",
5     "created": "2018-07-23T10:10:23.6",
6     "modified": "2019-02-06T10:10:23.6",
7     "ethical_issues_exist": "unknown",
8     "contact": {
9       "mail": "cc@example.com",
10      "name": "Charlie Chaplin",
11      "contact_id": {
12        "contact_id": "http://orcid.org/0000-0000-0000-0000",
13        "contact_id_type": "HTTP-ORCID"
14      }
15    },
16    "dataset": [
17      {
18        "title": "Placeholder dataset",
19        "type": "(value from dictionary)",
20        "personal_data": "unknown",
21        "sensitive_data": "unknown"
22      }
23    ]
24  }
25 }
```

```
{
  "dmp": {
    "title": "Example DMP",
    "description": "An in-depth description",
    "ethical_issues_exist": "unknown",
    "language": "en",
    "created": "2018-07-23T10:10:23.6",
    "modified": "2019-02-06T10:10:23.6",
    "contact": {
      "mail": "cc@example.com",
      "name": "Charlie Chaplin",
      "contact_id": {
        "contact_id": "http://orcid.org/0000-0000-0000-0000",
        "contact_id_type": "HTTP-ORCID"
      }
    }
  },
  "dataset": {
    "title": "Example DMP Dataset",
    "type": "dataset",
    "personal_data": "unknown",
    "sensitive_data": "unknown"
  }
}
```

RDA-DMP-Common-Standard - Minimal DMP:

<https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/examples/JSON/ex8-dmp-minimal-content.json>

Can We Do Better With The Existing Data? -- Themes

→ Themes

- ◆ <https://github.com/DMPRoadmap/roadmap/wiki/Themes>
- ◆ Currently => A Tagging system between questions and guidance
- ◆ But, some map quite well to “string” fields on the maDMPS
 - Quickly a grey-area
- ◆ Data Description => Dataset description
- ◆ Ethics & Privacy => DMP ethicalIssuesDescription
 - ? Ethics & Privacy => Security and Privacy object ?
- ◆ Metadata & Documentation => Metadata object
- ◆ Budget => Cost object
- ◆ Roles & responsibilities => DMStaff object
- ◆ Preservation => Dataset preservationStatement
- ◆ ? Storage & Security => Security and Privacy object

```
"Data description",  
"Data collection",  
"Metadata & documentation",  
"Storage & security",  
"Preservation",  
"Data sharing",  
"Related policies",  
"Data format",  
"Data volume",  
"Ethics & privacy",  
"Intellectual Property Rights",  
"Data repository",  
"Roles & responsibilities",  
"Budget"
```

Can We Do Better With The Existing Data? -- Themes

→ Approach

- ◆ Collect answers to questions with specific themes
- ◆ Concatenate (with separators) if cardinality [0..1]
- ◆ Add separately if cardinality [0..*]

→ Remaining Questions

- ◆ Just add the answer, or the question text as well ...?
 - Answers lose context without the questions.
- ◆ Ambiguous Themes

```
"Data description",  
"Data collection",  
"Metadata & documentation",  
"Storage & security",  
"Preservation",  
"Data sharing",  
"Related policies",  
"Data format",  
"Data volume",  
"Ethics & privacy",  
"Intellectual Property Rights",  
"Data repository",  
"Roles & responsibilities",  
"Budget"
```


Can We Improve The Data? -- Themes

→ Approach

- ◆ Adjust the DMPRoadmap themes to more-closely align with maDMP fields and objects
- ◆ Ex: Ethics & Privacy split into two themes
- ◆ Ex: Data Quality theme added => dataset dataQualityAssurance

→ Result

- ◆ Easier to correspond freetext answers in DMPRoadmap to string fields in the maDMP

Can We Improve The Data? -- Adding Fields

Lots of little ones here, with minor backend/UI changes:

Start and end dates on the Plan would enable Project object

Adding identifiers to organisations (ROR, Funder Registry), along with a funding_status on the plan enables the Funding object

A Common Problem: Controlled Vocabularies

The maDMP format enforces the use of many controlled vocabularies, which DMPRoadmap currently has no concept of:

Cost.currency_code	Allowed values defined by ISO 4217.
DMStaff.contributor_type	Contributor Type. Allowed values as defined by DataCite. See: https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf
Dataset.type	Type according to: http://vocabularies.coar-repositories.org/pubby/resource_type.html
Metadata.identifier	Controlled Vocabulary using an Identifier
Funder.funder_id	Funder ID, recommended to use CrossRef Funder Registry. See: https://www.crossref.org/services/funder-registry/

Can We Improve The Data? -- Extending Roadmap Model

→ New, Structured Question Formats

◆ Cost

- currency type, numerical amount field, and comment
- Maps onto cost object when tagged with budget theme

◆ Boolean

- Simple Yes/No
- in-conjunction with ethics, personal, and sensitive themes, and a potential unanswered status maps onto the corresponding data/issuesExist fields

◆ Staff

- Link to user's identifiers (i.e. ORCID) & list allowed contributor_type options
- Maps onto DMStaff


◆ Metadata

- Restrict choice to a controlled vocabulary to allow saving an identifier


Existing Work - RDA Metadata Standards Question Type

What are the metadata standards you will use?


Your Selected Standards:

- EML (Ecological Metadata Language) 


Please select a subject

Science 

Please select a sub-subject


Environmental sciences and € 

Browse Standards

DIF (Directory Interchange Format) 

An early metadata initiative from the Earth sciences community, intended for the description of scientific data sets. It includes elements focusing on instruments that capture data, temporal and spatial characteristics of the data, and projects with which the dataset is associated. It is defined as a W3C XML Schema.






Sponsored by the Global Change Master Directory, the DIF Writer's Guide Version 6 is from November 2010.

EML (Ecological Metadata Language) 


Ecological Metadata Language (EML) is a metadata specification particularly developed for the ecology discipline. It is based on prior work done by the Ecological Society of America and associated efforts (Michener et al., 1997, Ecological Applications).

Sponsored by ecoinformatics.org, EML Version 2.1.1 was released in 2011.

Comment

B I     

EML is one of the accepted formats used in ecology, and works well for the types of data we will be producing. We will create these metadata using Morpho software, available through KNB.



Provides an extendable way of storing the data for new question-formats.

Example of pulling in controlled vocabulary from an external API.

```
def answer_hash
  default = { "standards" => {}, "text" => "" }
  begin
    h = text.nil? ? default : JSON.parse(text)
    rescue JSON::ParserError => e
      h = default
    end
    h
  end
```

```
def update_answer_hash(standards = {}, text = "")
  h = {}
  h["standards"] = standards
  h["text"] = text
  self.text = h.to_json
end
```

A Big Remaining Question -- Datasets

A large assumption we made at the beginning was that all questions were answered about one dataset. While this is needed with our current data model, to properly support the maDMP standard, we would want to be able to express multiple datasets for a DMP.

DMPOPIDoR team has piloted functionality to support answers for multiple, defined, datasets in their fork of the DMPRoadmap project:

https://github.com/OPIDoR/DMPOPIDoR/tree/dmpopidor_branding

An upshot to the way this is implemented is all of the theme/question mapping logic would remain the same, and just map onto multiple dataset objects.

A Minimal DMP in DMPonline

```
json.prettify!~
~
json.dmp do~
  json.title @plan.title~
  json.description @plan.description if @plan.description.present?~
  json.language @plan_info[:language]~
  json.created @plan.created_at~
  json.modified @plan.updated_at~
  json.ethicalIssuesExist @plan_info[:ethical_issues]~
  ## TODO: add in other ethical issues fields~
  ## TODO: Map answers to question with theme "Ethics & privacy"~
  ## json.ethicalIssuesDescription~
~
  json.contact do~
    json.name @plan_info[:contact][:name]~
    json.mail @plan_info[:contact][:mail]~
    json.contact_id do~
      json.contact_id @plan_info[:contact][:id]~
      json.contact_id_type @plan_info[:contact][:id_type]~
    end~
  end~
~
  ## TODO: make this an array with a single item~
  json.dataset do~
    json.personal_data "unknown"~
    json.sensitive_date "unknown"~
    json.title @plan.title ## NOTE: this is a bad mapping~
    json.type "dataset" ## NOTE: this is a bad mapping~
    json.description @plan_info[:dataset][:desc]~
  end~
end~
```

```
# frozen_string_literal: true~
~
class Api::V0::RdaController < Api::V0::BaseController~
  ~
  before_action :authenticate~
  ~
  def export_dmp~
    @plan = Plan.find(params[:id])~
    @plan_info = {}~
    @plan_info[:ethical_issues] = "unknown"~
    @plan_info[:language] = @plan.owner.language || Language.default~
    @plan_info[:language] = @plan_info[:language].abbreviation~
    @plan_info[:contact] = {}~
    @plan_info[:contact][:name] = @plan.owner.name~
    @plan_info[:contact][:mail] = @plan.owner.email~
    @plan_info[:contact][:id_type] = "HTTP-ORCID"~
    @plan_info[:contact][:id] = "http://orcid.org/0000-0000-0000-0000"~
    @plan_info[:staff] = []~
    @plan_info[:staff] = @plan.roles.editor.not_creator.map(&:user)~
    @plan_info[:dataset] = {}~
    data_desc = Theme.find_by(title: "Data description")~
    @plan_info[:dataset][:desc] = @plan.answers.select { |a| a.question.themes.pluck
  end~
end~
```