RESEARCH DATA ALLIANCE

BoF: Using schema.org and enriched metadata to enable/boost FAIRness on research resources

Leyla Garcia, John Graybeal, Josef Hardi, Nick Juty, Siri Jodha Khalsa,
Fotis Psomopoulos, Adam Shepherd, Mingfang Wu

research data sharing without barriers
rd-alliance.org

3 April, *2019, RDA 13th Plenary Meeting, Philadelphia*

Collaborative Session Notes

Session notes: https://goo.gl/q7d1te

Session page:
https://www.rd-alliance.org/bof-using-schemaorg-and-enriched-metadata-enableboost-fairness-research-resources-rda-13th-plenary

(please sign in)

RESEARCH DATA ALLIANCE

Motivation & Goals

Nick Juty
University of Manchester
ELIXIR-UK

Horizon 2020
654248

CORBEL

research data sharing without barriers
rd-alliance.org

Motivation & goals

- Schema.org well-established
  - Leverages on Schema.org
    - Simple to adopt, moderately expressive
    - Extendable
    - Indexed (discoverability)
- Becoming a popular mechanism
  - DataCite DOIs : schema.org (JSON-LD) through content negotiation
  - DataVerse : citation metadata to dataset landing pages
- Other research domains (Survey)
  - Will have their own ‘best practice’
    - commonalities across domains
    - Key differences expressing domain relevant information
- ResearchSchemas

Motivation & goals

- DDP IG TF: "Using schema.org for research data discovery"
- Propose creation of a focused and formal RDA working group to establish ‘ResearchSchemas’ as a community

- Initial landscaping survey to determine & refine objectives
  - Which metadata schemas are used across research areas
    - Not necessarily schema.org
  - Mapping across equivalent properties (crosswalk)
    - Facilitate interoperability & MI recommendations
  - Identify gaps
    - Proposing new types, properties or relations

- Metadata & standards => improve **FA**IR

Agenda

- Welcome, goal of the session
- Current practices in working on schemas to describe datasets
  - Bioschemas (https://github.com/bioschemas)
  - Science-on-schema (https://github.com/ESIPFed/science-on-schema.org)
- Tools to facilitate schema converting, validation, etc
  - From existing metadata schemas to schema.org
- Report on the survey
- Process for setting up the research schemas WG
- Discussion
  - WG case statement, Schemas and extensions, Tooling, Guidelines, FAIR alignment in research schemas, ..
- Summary & next steps

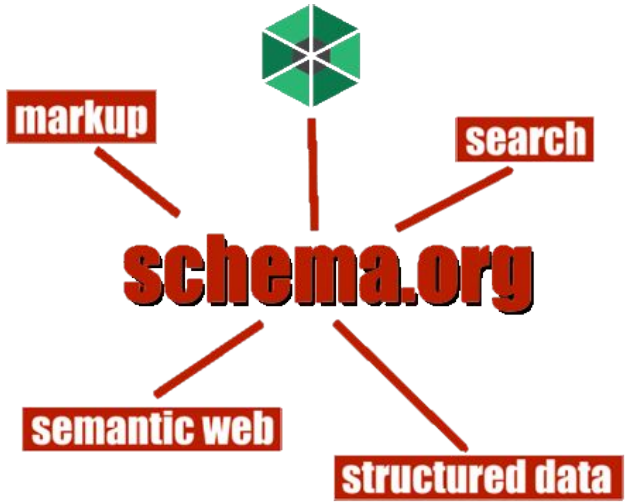
Bioschemas

Leyla Garcia
Elixir Hub

research data sharing without barriers
rd-alliance.org

Background schema.org

- Collaborative, community activity with a mission to create, maintain, and promote **schemas** for **structured data** on the Internet, on **web pages**, in email messages, and beyond



Background

schema.org

Types

What we are talking about

Core vocabulary

- Thing
  - Action
    - AchieveAction
      - LoseAction
      - TieAction
      - WinAction
    - AssessAction
      - ChooseAction
        - VoteAction
      - IgnoreAction
      - ReactAction
        - AgreeAction
        - DisagreeAction
        - DislikeAction
        - EndorseAction
        - LikeAction
        - WantAction
      - ReviewAction
    - ConsumeAction
      - DrinkAction
      - EatAction
      - InstallAction
      - ListenAction
      - ReadAction
      - UseAction
        - WearAction
      - ViewAction
      - WatchAction
    - ControlAction
      - ActivateAction
      - DeactivateAction
      - ResumeAction

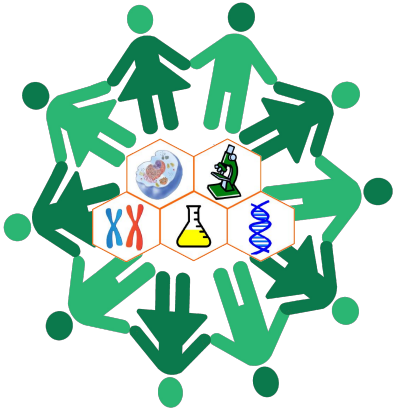
Properties

What we can say about it

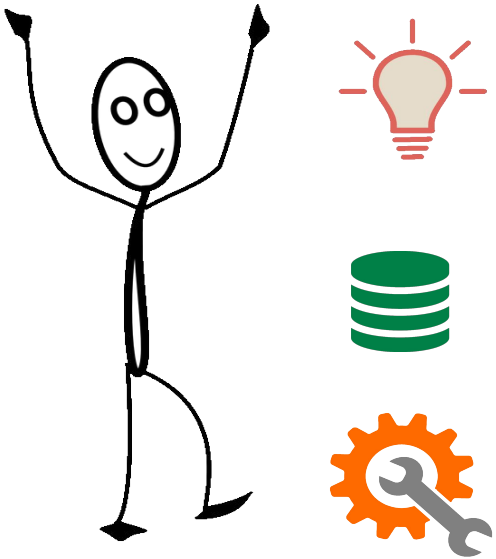
Property	Expected Type	Description
Properties from Person		
additionalName	Text	An additional name for a Person, can be used for a middle name.
address	PostalAddress or Text	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
alumniOf	EducationalOrganization or Organization	An organization that the person is an alumni of. Inverse property: alumni.
award	Text	An award won by or for this item. Supersedes awards.
birthDate	Date	Date of birth.
birthPlace	Place	The place where the person was born.
brand	Brand or Organization	The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person.
children	Person	A child of the person.
colleague	Person or URL	A colleague of the person. Supersedes colleagues.
contactPoint	ContactPoint	A contact point for a person or organization. Supersedes contactPoints.
deathDate	Date	Date of death.
deathPlace	Place	The place where the person died.
duns	Text	The Dun & Bradstreet DUNS number for identifying an organization or business person.
email	Text	Email address.
familyName	Text	Family name. In the U.S., the last name of an Person. This can be used along with givenName instead of the name property.
faxNumber	Text	The fax number.
follows	Person	The most generic uni-directional social relation.
	Organization or	A person or organization that supports (sponsors) something

Bioschemas

- Community initiative built on top of **schema.org**
- Aim
  - Improve data discoverability and interoperability in Life Sciences
- How
  - Adding Life Science types to schema.org
  - Providing usage guidelines, examples and tools



Motivation



- Researcher looking for help to
  - sequence a cactus → people, expertise
  - find similar sequences → datasets
  - compare and visualize → tools

Use cases

- Generic
  - Search engines
  - Google dataset specialized search
- Life sciences
  - Data depositions can harvest markup from small providers → small providers can benefit back from data depositions
  - Tess (training and events registry) → Automatic population → trainers add markup
  - Common/rare terms → Resource index

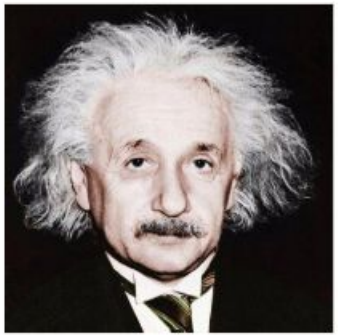
It is all about context

Albert

Einstein

Birth date: 14 March 1879

Birthplace: Ulm, Kingdom of Württemberg, German Empire



Height: 1.72 cm

Alumni of: University of Zurich

Weight: unknown

Death date: 18 April 1955

Work location: 1 Einstein Drive Princeton, New Jersey 08540 USA

Know about: Physics, relativity, mass energy equivalence

email

CONTEXT

Honorific prefix: Dr. Prof.

Knows language: German, English

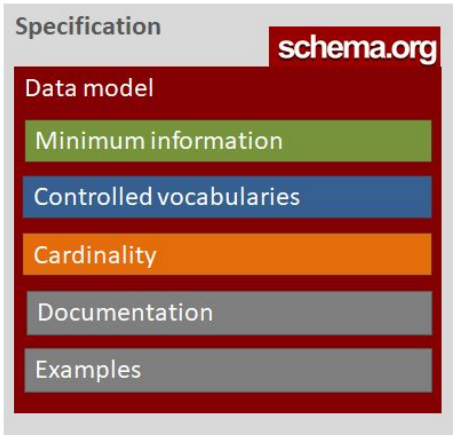
Award: Nobel prize in physics

orcid

Siblings: Maria

isicv4: 7210

Profiles: types in context



Dataset Profile

Version: 0.2 (25 February 2018)

Bioschemas specification for describing a dataset in the life-science.

Description | Contributors | Links

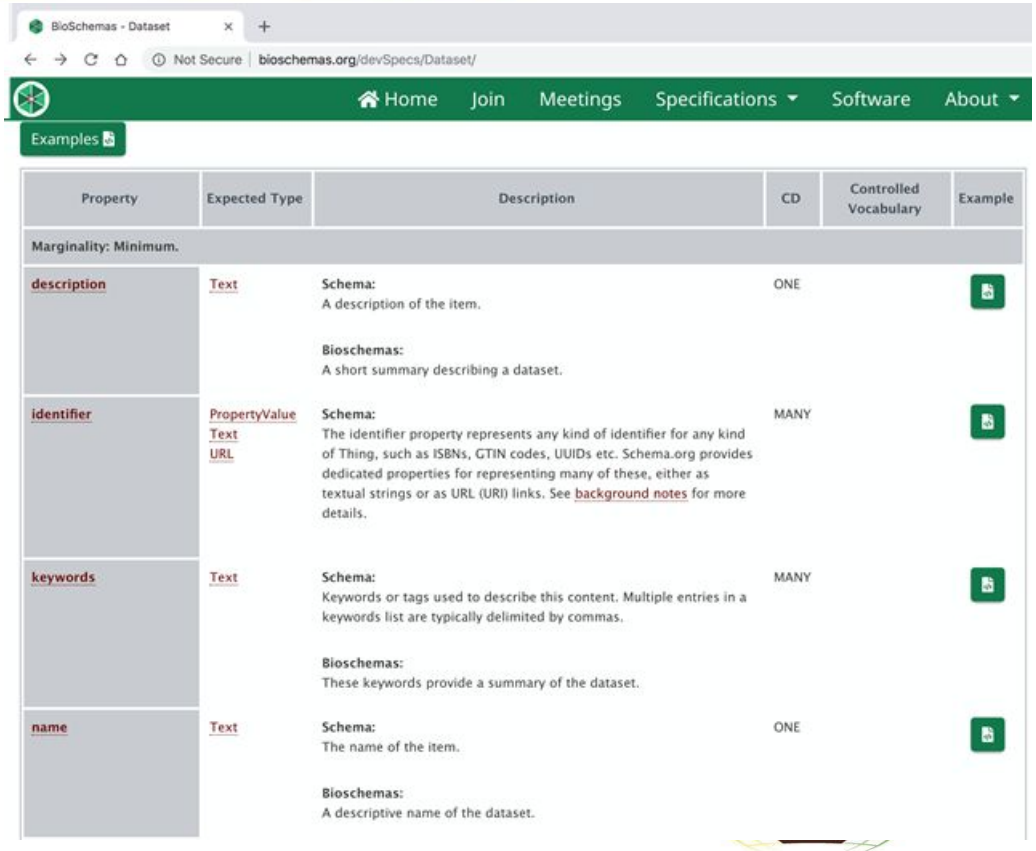
Schema.org hierarchy

This Profile fits into the schema.org hierarchy as follows:

Thing > CreativeWork > Dataset

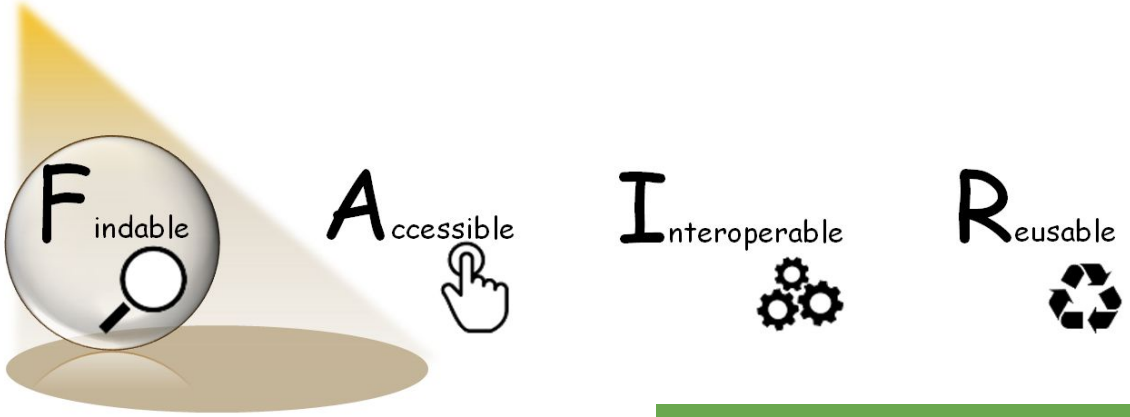
Description

A guide for how to describe datasets in the life-sciences using Schema.org-like annotation.



Property	Expected Type	Description	CD	Controlled Vocabulary	Example
Marginality: Minimum.					
description	Text	**Schema:** A description of the item. **Bioschemas:** A short summary describing a dataset.	ONE		
identifier	PropertyValue Text URL	**Schema:** The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.	MANY		
keywords	Text	**Schema:** Keywords or tags used to describe this content. Multiple entries in a keywords list are typically delimited by commas. **Bioschemas:** These keywords provide a summary of the dataset.	MANY		
name	Text	**Schema:** The name of the item. **Bioschemas:** A descriptive name of the dataset.	ONE		

Contribution to FAIRability



- Unique identifiers (included in metadata)
- Descriptive metadata
- Indexed and available

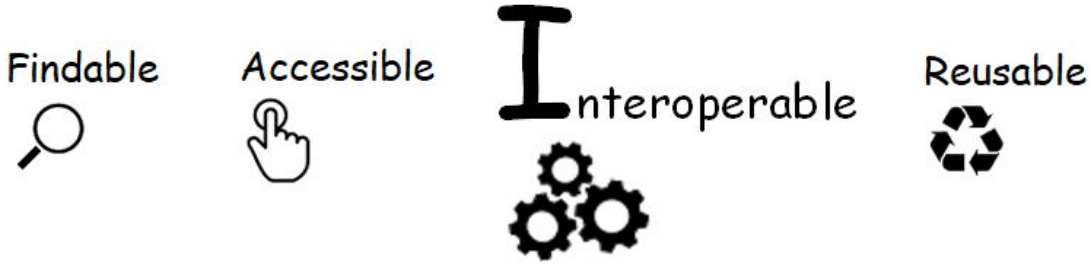
- Identifier → minimum → always part of the metadata
- Metadata focused on use cases
- Indexed by Google, some specialized resources (bio.tools, TeSS)

- Future → Bioschemas specialized indexes

Findable  Accessible  Interoperable  Reusable 

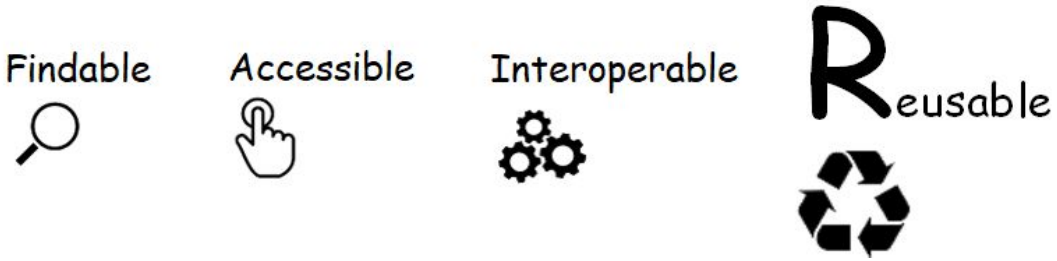
- Open protocol
- Authorization-based protocols
- Metadata always accessible

- Mainly HTTP & HTTPS
- Authorization protocols **could** be express via metadata (but not included yet)
- Longevity via specialized registries (not always and mostly no plan if registries disappear)



- Metadata on structured language
- Use of FAIR vocabularies
- Qualified references to other resources

- Schema.org + profiles → JSON-LD
- Recommendations on controlled vocabularies (but FAIR vocabularies are not always available/suitable)
- Not supported, mainly generic relations to others



- Rich metadata
- License
- Provenance
- Community standards

- Partially supported → Minimum, recommended and optional
- License (but not always as minimum)
- Some provenance covered → Dataset provider, publication authors, structure data creator
- Community involvement since conception

Summary

- Focus on key properties prioritized as Minimum, Recommended and Optional based on community agreements and common practices

•Additional recommendations regarding properties cardinality

- Customization on schema.org types to better supports needs on the life sciences community

•Terms reused from well-known ontologies thus avoiding reinventing the wheel

https://biohackathon-europe.org

BIOHACKATHON

-22nd Nov 2019 Paris

Themes Can include :

Text mining, Structured metadata, Identifiers, Data distribution, Data integration, Data Validation, Tools, Containers, Tools discovery, and Training materials

Opportunities for companies to submit hacking topics

Project submission closes 7th April

Contact: Jen.Harrow@elixir-europe.org

Science-on-schema

Adam Shepherd

research data sharing without barriers
rd-alliance.org

Science-on-schema.org - Origin

NSF EarthCube Repository Working Group

PROBLEM

- Lots of websites describing repositories
- Repository managers duplicating effort

SOLUTIONS

- repositories self-publish using schema.org + extensions
- based on re3data.org
- Produced Guidelines
- dozen NSF data repositories

Science-on-schema.org - Origin

NSF EarthCube Project 418 (P418)

- Registry of NSF-GEO funded datasets

- Expanded Repository Working Group:
  - **Guidelines** for publishing schema.org for Datasets
  - schema.org harvesting & indexing **software**
  - **prototype UI** and API

Google

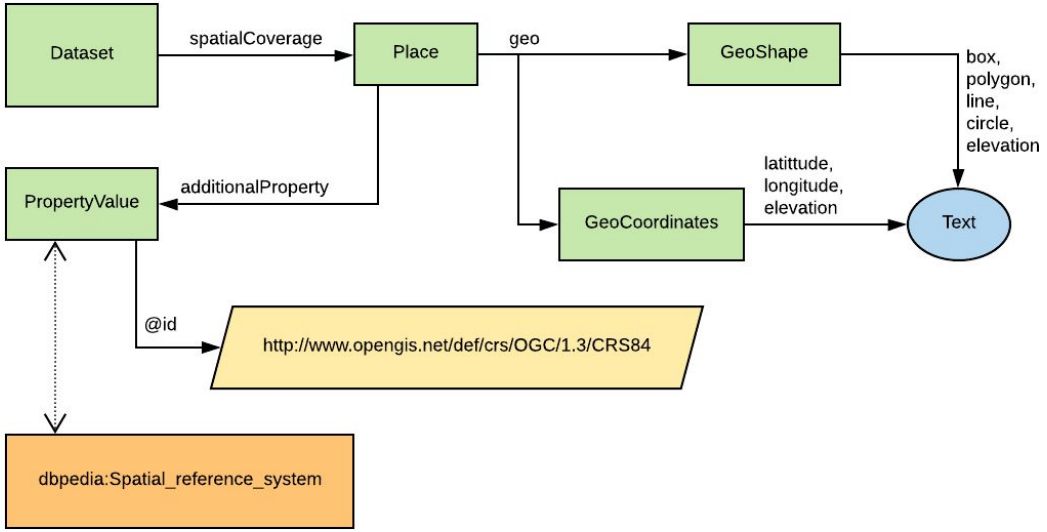
github: ESIPFed/science-on-schema.org

A point, or coordinate, would defined in this way:

```
{
  "@context": {
    "@vocab": "http://schema.org/",
    "datacite": "http://purl.org/spar/datacite/"
  },
  "@type": "Dataset",
  "name": "Removal of organic carbon by natural bacterioplan
  ...
  "spatialCoverage": {
    "@type": "Place",
    "geo": {
      "@type": "GeoCoordinates",
      "latitude": 39.3280
      "longitude": 120.1633
    }
  }
}
```

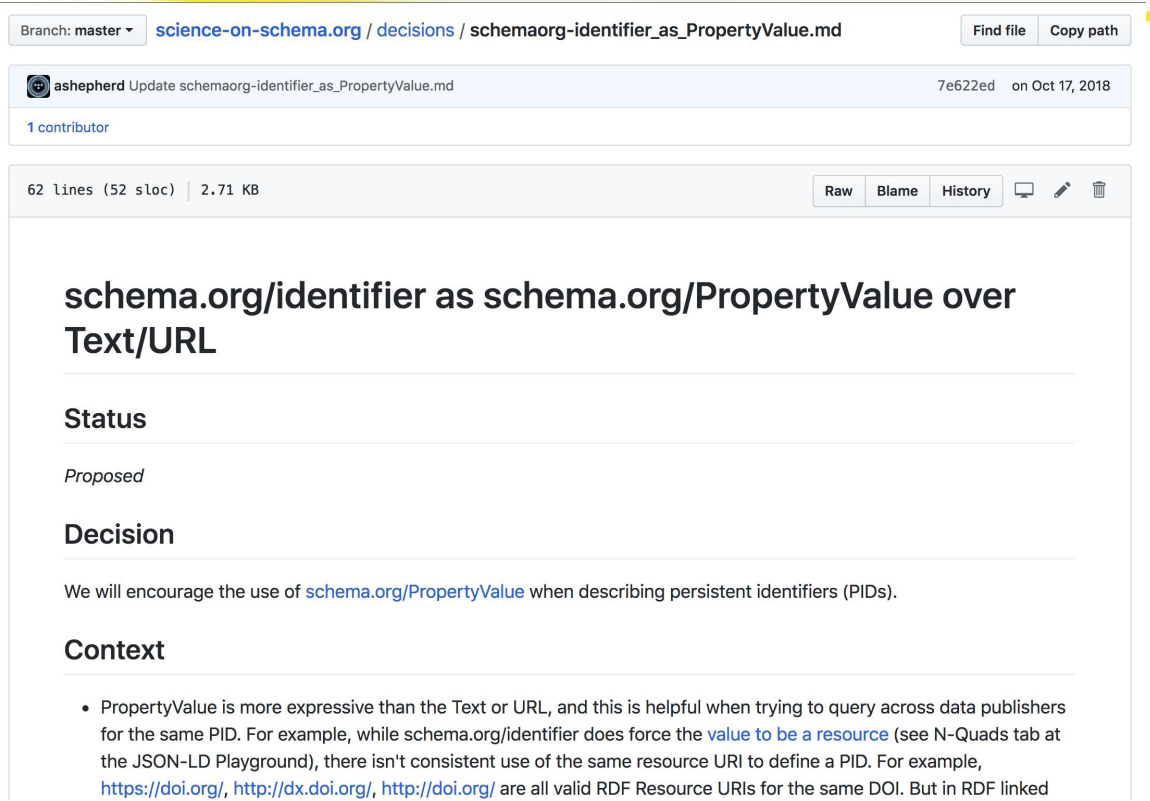
All other shapes, are defined using the schema:GeoShape:

```
  "spatialCoverage": {
    "@type": "Place",
    "geo": {
      "@type": "GeoShape",
      "line": "39.3280,120.1633 40.445,123.7878"
    }
  }
```



RDA
RESEARCH DATA ALLIANCE

github: ESIPFed/science-on-schema.org



Branch: master ▾ science-on-schema.org / decisions / schemaorg-identifier_as_PropertyValue.md Find file Copy path

ashepherd Update schemaorg-identifier_as_PropertyValue.md 7e622ed on Oct 17, 2018

1 contributor

62 lines (52 sloc) | 2.71 KB Raw Blame History

schema.org/identifier as schema.org/PropertyValue over Text/URL

Status

Proposed

Decision

We will encourage the use of schema.org/PropertyValue when describing persistent identifiers (PIDs).

Context

- PropertyValue is more expressive than the Text or URL, and this is helpful when trying to query across data publishers for the same PID. For example, while schema.org/identifier does force the value to be a resource (see N-Quads tab at the JSON-LD Playground), there isn't consistent use of the same resource URI to define a PID. For example, https://doi.org/, http://dx.doi.org/, http://doi.org/ are all valid RDF Resource URIs for the same DOI. But in RDF linked

Decisions

document:

- status
- context
- consequences

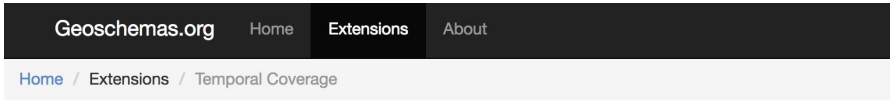
follows

Architectural Decision Records (ADRs)

geoschemas.org

Extending existing schema.org properties

- Technique blessed by Dan Brickley, Google



Temporal Coverage

The OWL Time vocabulary defines useful ways for describing the temporal coverage of some schema:CreativeWork such as a Dataset.

Example

```
{
               "@context": [
                {
                  "@vocab": "http://schema.org/",
                  "time": "http://www.w3.org/2006/time#"
                },
                "https://geoschemas.org/context/temporal.jsonld"
               ],
               "@type": "Dataset",
               "temporalConverage": {
                 "@type": "time:TemporalEntity",
                 "time:hasBeginning": {
                   "@type": "time:Instant",
                   "time:inXSDDateTimeStamp": "2015-11-01T17:58:16.102Z"
                 }
               }
             }
```

1. Make your `@context` an *array*.
2. Make sure the *first* context is schema.org
3. Add the temporal converage context file to the `@context`.
4. Now you can begin to use OWL Time classes as a valid `schema:temporalCoverage`

Adoption

BALTO
CHORDS
CDF RWG

MagIC

BCO-DMO
Biological & Chemical Oceanography Data Management Office

Neotoma

HYDROSHARE

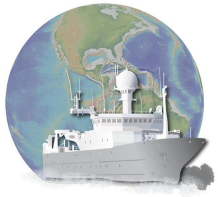
UNAVCO

IEDA
INTERDISCIPLINARY
EARTH DATA ALLIANCE

OpenTopography
High-Resolution Topography Data and Tools

LinkedEarth

IRIS



R2R

research data sharing without barriers
rd-alliance.org

Open
Core

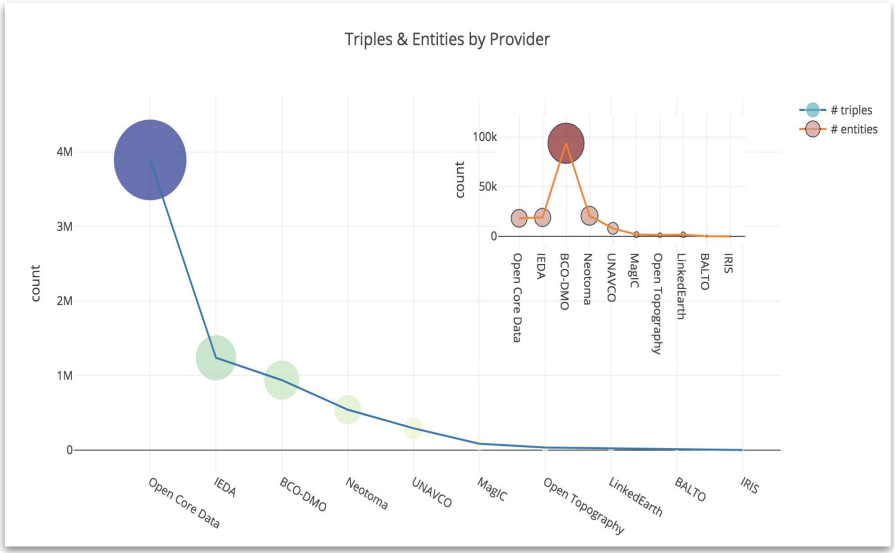
JOIDES Resolution
Science Operator

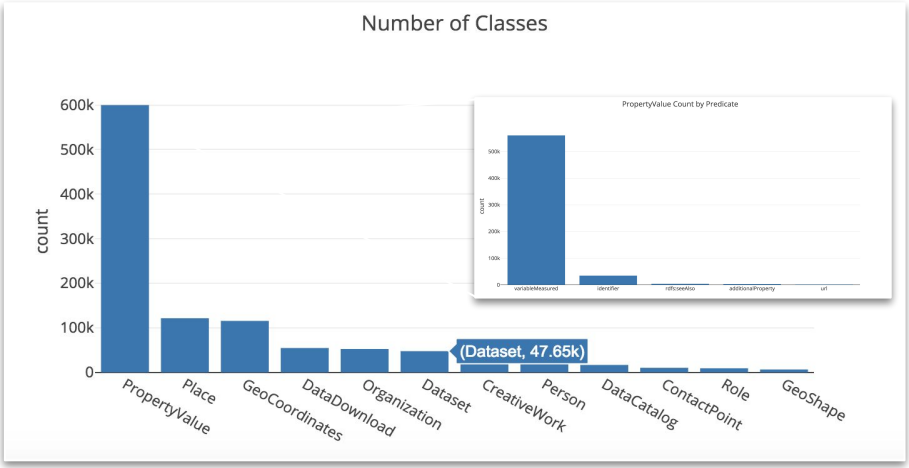
CSDCO
Continental Scientific Drilling
Coordination Office

RDA
RESEARCH DATA ALLIANCE

Summary Statistics

1,113,210 entities
7,087,380 triples





47,650 Dataset
54,665 DataDownload
599,960 PropertyValue
~ 35k Identifiers
~560k Dataset Variables

Use of Guidelines

Dataset Properties	Google Requires / Recommends	Provider Usage	Dataset Coverage	
			Implemented	*Overall*
@context	**Required**. Set `@context` to "http://schema.org/"	**80%**	omitted ending slash: 'http://schema.org'	
@type	**Required**. Set `@type` to "Dataset"	**100%**	**47,650 datasets**	*n/a*
name	**Required**. A descriptive name	**80%**	99.9%	73%
description	**Required**. A short summary	**70%**	97%	69%
url	*Recommended*.	**70%**	100%	62%
citation	*Recommended.*	**60%**	100%	36%
keywords	*Recommended.*	**70%**	99.9%	66%
spatialCoverage	*Recommended.*	**80%**	92%	91%
temporalCoverage	*Recommended.*	**10%**	15%	<1%
variableMeasured	*Recommended.*	**30%**	83%	40%
version	*Recommended.*	**40%**	95%	25%
sameAs	*Recommended.* Same data, different URL.	**10%**	100%	<1%

https://developers.google.com/search/docs/data-types/dataset

80 - 100%	50 - 79%	0 - 49%

Use of Guidelines *cont'd*

Dataset Properties	Provider Usage		Dataset Coverage	
			Implemented	*Overall*
identifier	**30%**	10,556 datasets	100%	22%
author/creator/contributor	**80%**	28,765 datasets	98%	69%
funder *(not awards)*	**30%**	4,069 datasets	78%	9%
distribution	**60%**	45,221 datasets	100%	95%
license	**70%**	42,523 datasets	98%	89%
hasPart *ex: linking PhysicalSamples to Datasets*	**10%**	122 datasets	2%	<1%

80 - 100%	50 - 79%	0 - 49%

Project 418 - Tools

Project 418 Tools: Four main tools developments

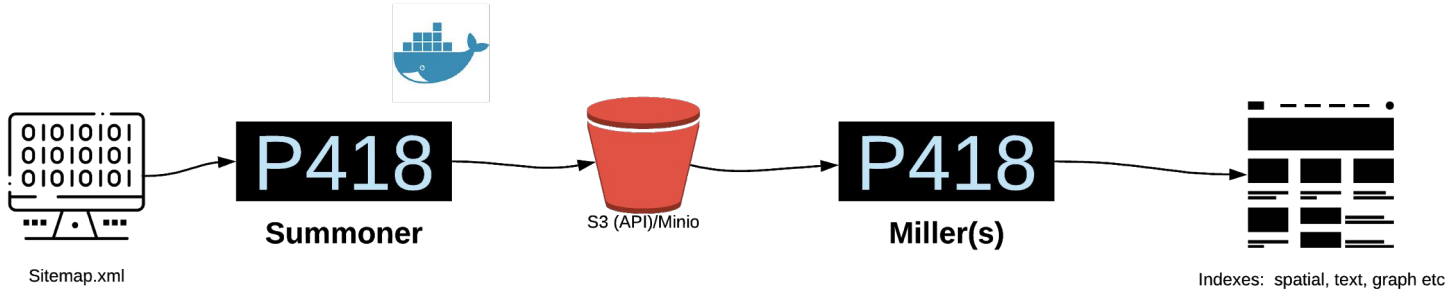
- Gleaner: The access and process tool.
  - Go based, deployed as a Docker image
  - A Docker Compose/Stack file can deploy Gleaner and all its dependencies
  - Accesses SDO data graphs and processes them into indexes
- Geodex
  - A simple web site used to provide a test UI for the generated indexes from Gleaner. NOT a production site.. Only for testing.
  - Also a set of test APIs to explore leveraging the index
- SHACL
  - Not a service, though we have developed a RESTful service around the TopQuadrant library. This service allows easy access to SHACL processing.
- Fence
  - In development, a inspection site/tool along the lines of JSON-LD playground or Google SDTT but more focused on the needs of the NSF data facilities

Science-on-schema.org

Project 418 Tools: Gleaner

Two components

- Summoner
  Accesses files via sitemap.xml and extracts JSON-LD (validates JSON-LD form)
- Miller
  A simple framework to support multiple processing pipelines on the summoned JSON-LD These include; Graph, Spatial, Prov, Full text indexing (multiple options), SHACL validation, more
- Go based, Dockerized, GitHub: https://github.com/earthcubearchitecture-project418/gleaner



Science-on-schema.org

Project 418 Tools: Geodex, SHACL, Fence

Geodex.org

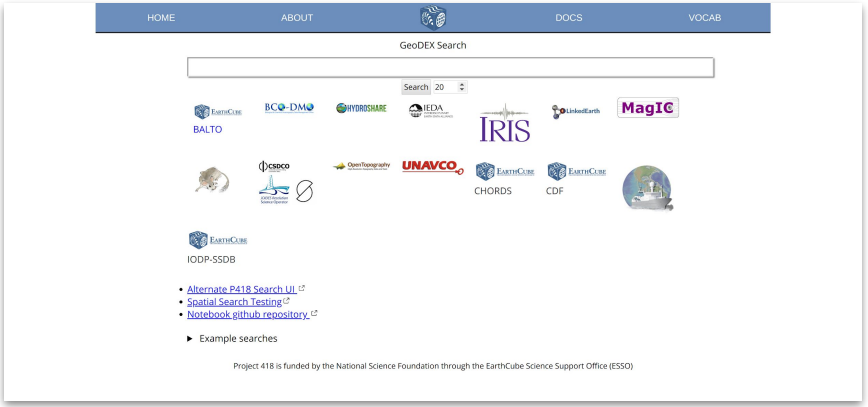
A simple testing site with web interface and services (RESTful). Allows us to leverage the indexes to see how they perform. (other clients too)

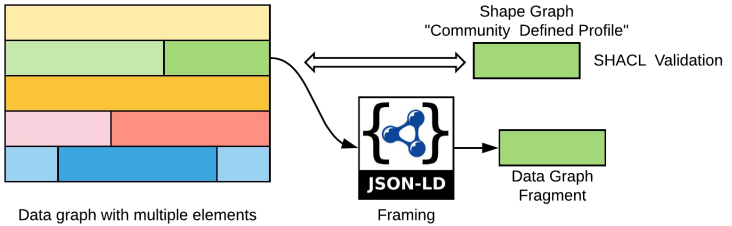
SHACL

Developing a set of “shape files” to test data graphs against. Developing as services to use in Gleaner and other tool chains. Allows providers to validate their data graphs.

Fence

Under development. A website to allow inspection of data graphs. Test their use via web components, validate with SHACL, route to other tools.





A mapping tool from metadata to schema.org

Josef Hardi, John Graybeal
Stanford BMIR

research data sharing without barriers
rd-alliance.org

About the tool

https://schemaorg.metadatacenter.org/playground

- We have developed an ETL tool that uses a *system-independent mapping language* that works well with source data in open-standard formats to produce schema.org-compliant instances.
- The idea of using a high-level language for data mapping is not new and many off-the-shelf ETL tools have it. We re-implemented the idea as an open source library where we then built a tool that could make the whole data transformation processes automatic and transparent to users.
- Currently, our tool supports XML and RDF source formats.
- We have transformed over 550,000 public scientific metadata from ClinicalTrials.gov, PubMed and DrugBank by using the tool.

How the tool works https://schemaorg.metadatacenter.org/playground

The tool requires an input of a **data source** document and a **data mapping**. In the batch mode, the source documents can be many and a single mapping will process them.

Other than that, the ETL processing details are hidden from the users.

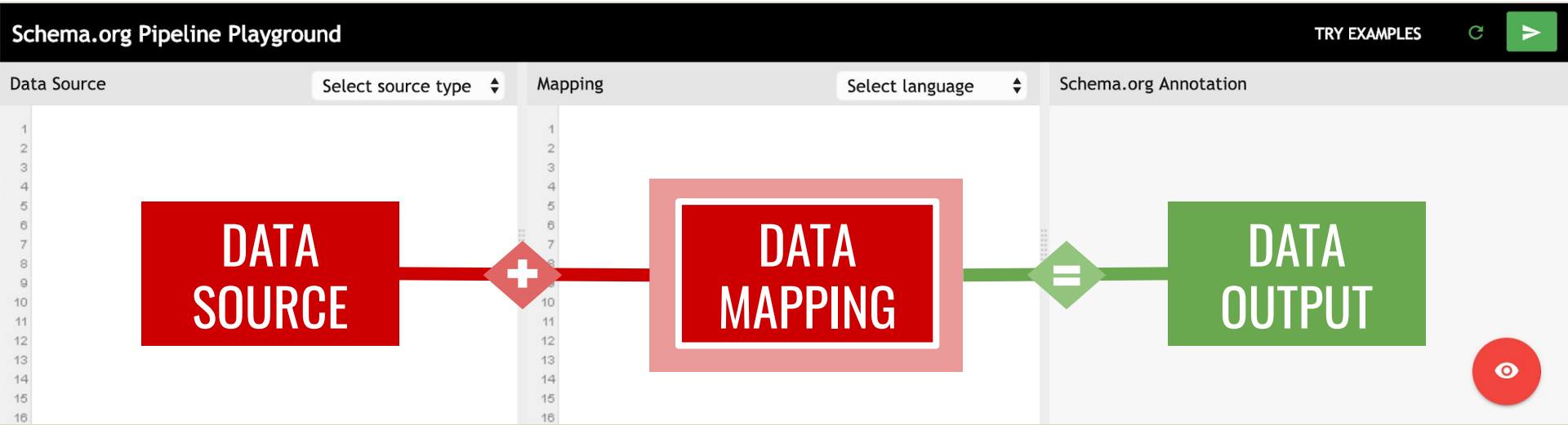


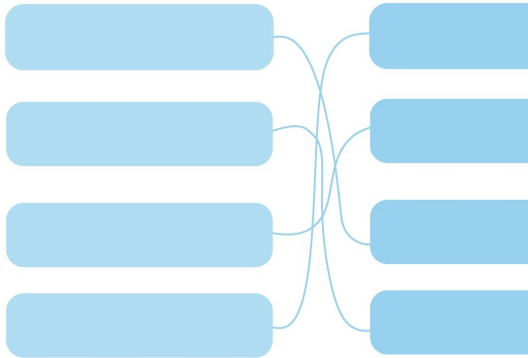
Figure 1: The tool playground

Choosing the mapping language

Users can choose from two language options. Both work equally well but they have pros and cons as described below:

1) **RML** / *"RDF Mapping Language"*
   - ➕ A well-known mapping language in the semantic web community
   - ➕ Rich language features
   - ➖ Too verbose

2) **CAML** / *"CAML is Another Mapping Language"*
   - ➕ Less verbose
   - ➖ An experimental language developed for the tool
   - ➖ Less rich language features



RML documentation: http://rml.io | CAML documentation: https://github.com/metadatacenter/schemaorg-pipeline/wiki/CAML

Writing a data mapping specification

A **map** requires a **target field** and a **source field** as a reference for extracting the value.

Target:

SCHEMA.ORG FIELDS

```
@type:                  'MedicalTrial'
name:                   /clinical_study/official_title
identifier:             /clinical_study/id_info/nct_id
status:                 /clinical_study/overall_status
description:            /clinical_study/detailed_description/textblock
studySubject:           /clinical_study/condition
phase:                  /clinical_study/phase
code:                   /clinical_study/condition_browse/mesh_term
    @type:              'MedicalCode'
    codeValue:          /.
studyLocation:          /clinical_study/location/facility
    @type:              'AdministrativeArea'
    name:               /name
    address:            /address
        @type:          'PostalAddress'
        addressLocality: /city
        addressRegion:  /state
        postalCode:     /zip
        addressCountry: /country
```

Source:

VALUE FIELDS

Figure 2: Data mapping in CAML

Automatic and transparent processing

The tool will automatically create a *data transformation program* based on the chosen data format. At the end of the pipeline, it will produce a schema.org instance in JSON-LD format. Two possible paths are available:

- If XML is chosen then a data transformation program based on XSLT will be auto-created,
- If RDF is chosen then a data transformation program based on SPARQL will be auto-created.

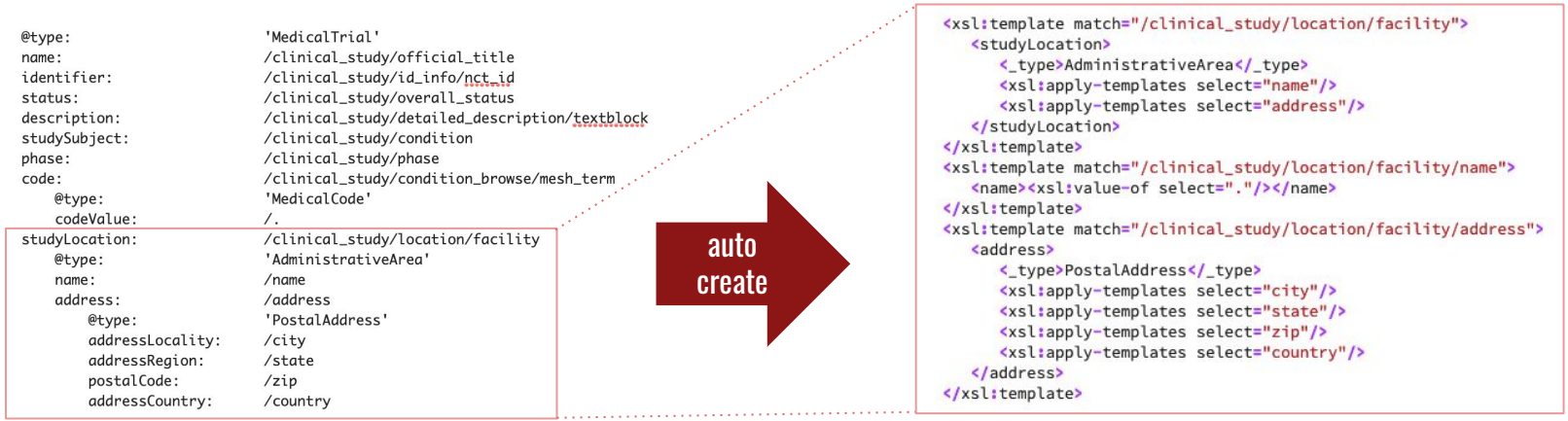
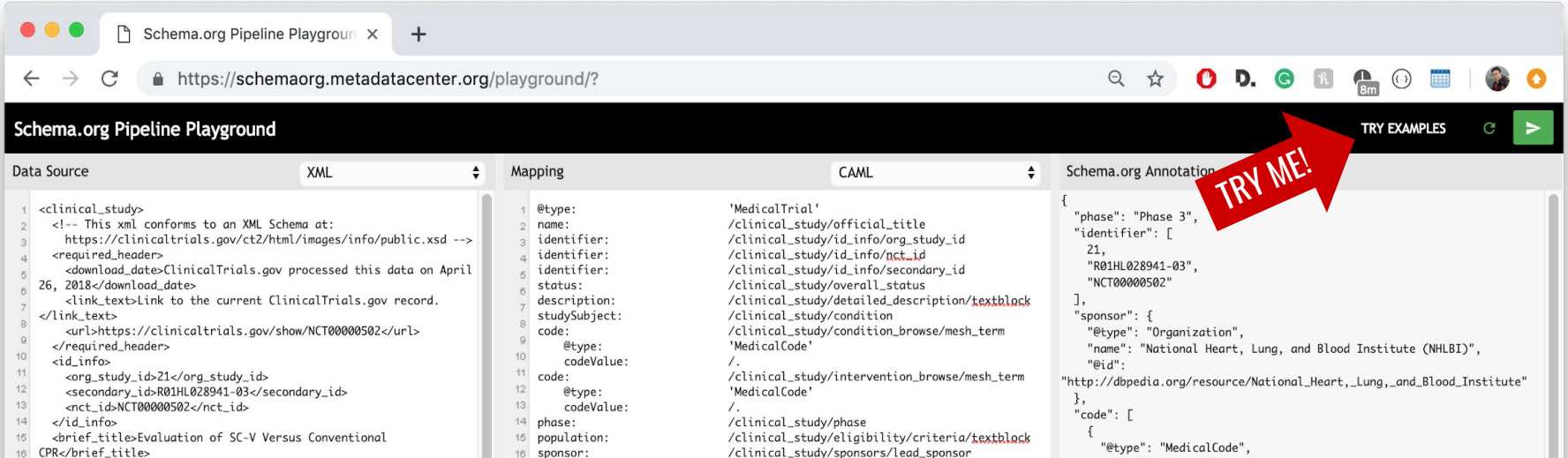


Figure 3: The auto-creation from a data mapping to an XSLT transformation specification

Take home messages

Please visit https://schemaorg.metadatacenter.org/playground and try some of the data transformation examples to get a better sense about the mapping and the tool itself.

All code resources are available freely at https://github.com/metadatacenter/schemaorg-pipeline

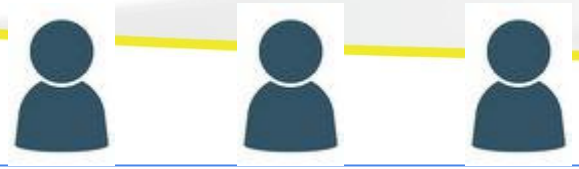


Data Discovery Paradigms IG
Using schema.org for research data discovery TF

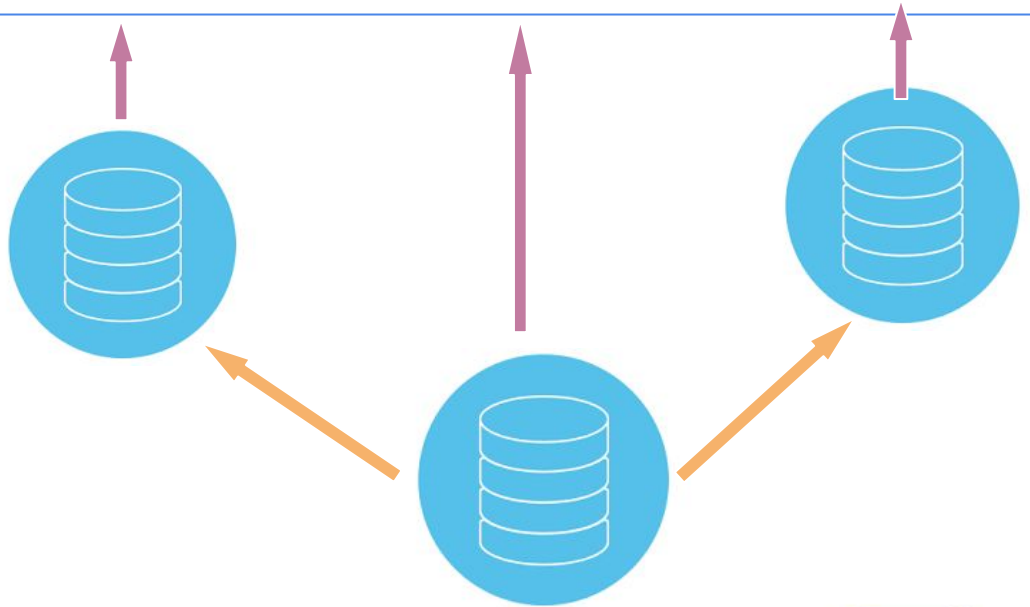
Mingfang Wu / Australia Research Data Commons

research data sharing without barriers
rd-alliance.org

Data discoverability via web search engines



Web search engines



- Common way to describe metadata across many resources
- Structured metadata even in resources without an API

Extensions to schema.org

https://schema.org/docs/schemas.html

Extensions

As schema.org has grown, we have developed mechanisms for community extension as a way of adding more detailed descriptive vocabulary that builds on the schema.org core.

Hosted extensions are managed and published as part of the schema.org project, with their design often led by one of more dedicated community groups.

External extensions live elsewhere in the Web, typically managed by other organizations with their own processes and collaboration mechanisms. Please consult external documentation for full details of their vocabulary, versioning system and release history.

Hosted Extensions

Specialized terms from hosted extensions can be used alongside core schema.org terms like Event and Person. For example in the auto extension there is a property for emissionsCO2, and in the bibliographic extension we have a property publisherImprint.

Using the extension mechanism the core vocabulary is extended by the following hosted extensions:

- auto.schema.org
- bib.schema.org
- health-lifesci.schema.org

Task force objectives

- **Common elements across research domains**
  - Objective 1 - Define research schemas types and minimum information guidelines for discoverability and accessibility
  - Objective 2 - Crosswalk and gap analysis evaluating existing standards and guidelines
- **Domain specific elements**
  - Objective 3 - Review existing efforts working on Schemas to describe research types
  - Objective 4 - Engagement and communication strategy; collaboration and with existing efforts

Survey: Objective

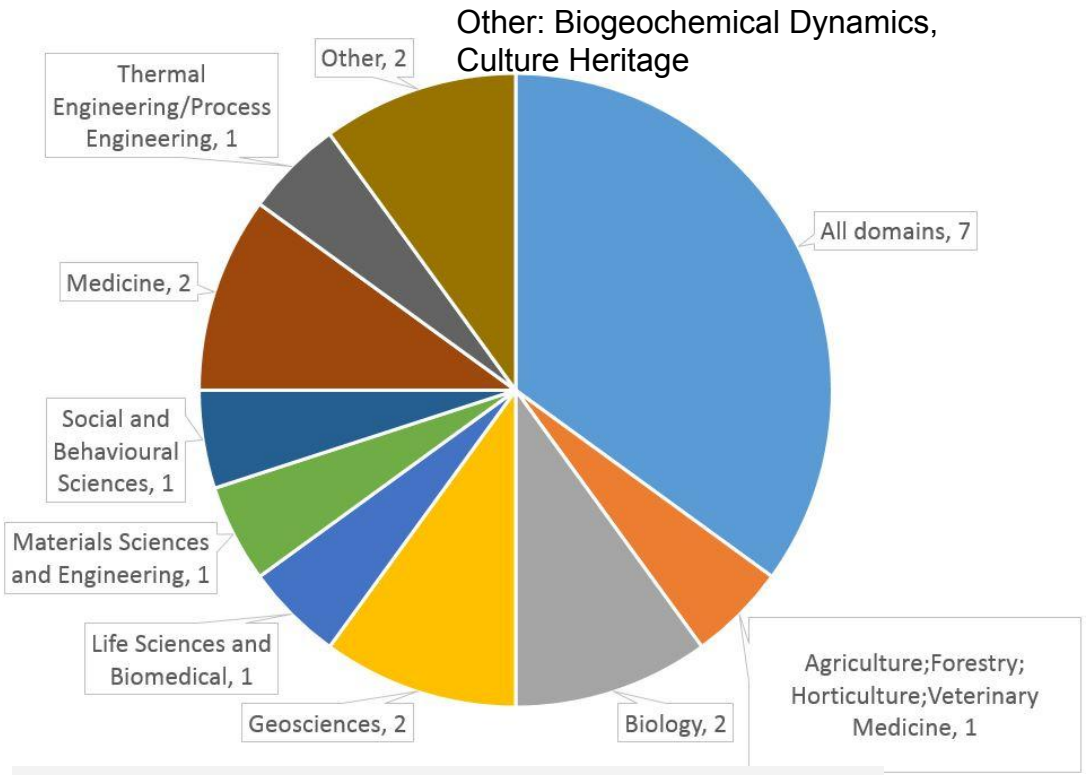
- This survey will gather information on existing work involving schemas to describe research data and related resources.
- Analysis of the survey results will help repositories and the proposed working group understand current practices, identify commonalities, gaps and barriers in using schemas for describing and discovering research datasets.
- It is envisaged that the survey results can inform the work group in planning its objectives and deliverables, along with sharing practices between data repositories.

Survey: current practices in using Schemas to describe research datasets

- Repository/Catalogue profile
  - Organisation name, URL of catalogue, domain covered, metadata schema(s)
- Current status of applying schema.org
  - Mapping from/to schema.org, mapping between other (non schema.org) schemas, the way schema.org is being applied
- Issue identification
  - Missing resource type, property, or relation property
- Suggestions to the research schemas ‘working group’

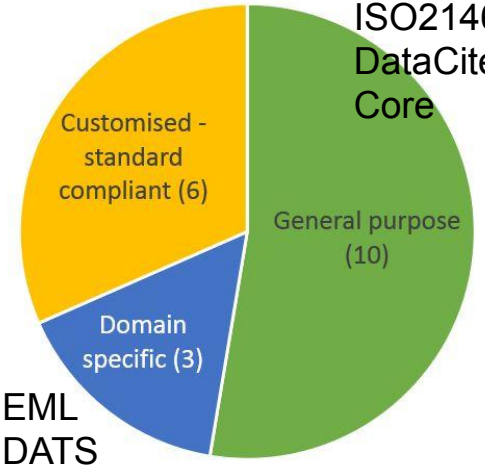
Survey: Catalogue profile

20 participations, dated on 25/03/2019



Domains covered by participating catalogues (20/20)

DCAT-AP, DCAT
DataCite
Schema.org
ISO2146
DataCite + Dublin Core



EML
DATS
DDI

Types of metadata schemas (19/20)

RDA
RESEARCH DATA ALLIANCE

Survey: Current status of applying schema.org

Mapping from/to schema.org (15/19)

- EML -> schema.org
- B2FIND <-> schema.org
- ISO2146/RIF-CS -> schema.org
- Dataverse -> schema.org
- CATS -> schema.org
- HCLS -> schema.org
- DCAT-AP -> schema.org
-

Ways of applying schema.org (16/19)

- Mark up of landing page in JSON-LD (8)
- Metadata schema (6)
- Possibly complementing (3)

Survey: Missing resource type/relation property/property

Resource type

- Scientific measurement
- Environmental entities
- Data services / APIs
- Tissue samples
- Data access arrangements
- Data reuse conditions/consent
- Data Controller (legal frameworks)
- Performances
- Digital artefacts
- Some from DataCite ResourceTypeGeneral, e.g. DataPaper, Model, Workflow

Relation property

- Dataset -> FundingAward
- Dataset -> Cruise (Event)
- Dataset -> Funder
- Study -> Study design
- Many from DataCite <relationType>, e.g. IsCitedBy, HasVersion, IsNewVersionOf, ...

Issues:

- Mapping multiple relation types into one
- Not sure if predicates (e.g. in the OBO Foundry Relation Ontology (RO), EnvO, and SWEET) are expressible

Property

- Keyword -> external vocabulary (e.g DefinedTerm, CategoryCode)
- Controlled vocabuary from DataCite <dateType> , e.g. Accepted, Available, Copyrighted, Updated, etc.
- Some semantic difference, e.g. schema:Dataset:name, DataCite:Author:name
- Specific term to generic term, e.g. dct:provenance to schema:description

Survey: Feedback (1)

? Why is schema.org meanwhile the one and only target metadata schema?

? Unfettered schema heterogeneity will hinder interoperability, so some mention of harmonisation strategies to be provided in the (still-to-be-developed) RDA Guidelines would be useful if we are hoping for: close data compatibility if not integratability.

? Schema.org (for metadata) rarely uses or allows for the use of controlled terminology (i.e. semantics) in the data values, and it does not make any effort to establish constraints, which result in a large set of metadata not actionable, validatable, or interoperable.

? Schema.org is not supporting some basic metadata information which is common across domains. This includes controlled vocabularies/thesauri/code lists

Survey: Feedback (2)

Developing/describing use cases and examples

One valuable contribution would be to
either provide a list of well-tested software for “collecting”
semantic assets (vocabularies, schemas, ontologies …) or host
such a repository themselves

Being able to document consistently “semantic assets” would be fundamental for enabling interoperability via re-use.

Would like to add Arts & Humanities in addition to scientific types.

The ‘problem’ of sensitive data and managed access datasets needs to be tackled by any group attempting to provide a comprehensive metadata framework.

Survey: Participants and Organisations - Thank you!

- NSF DataONE, NSF Arctic Data Center, university of California, Santa Barbara
- Scientific Computing Department, STFC Daresbury Laboratory
- Heinrich Widmann (European Datainfrastructure (EUDAT), hosted at and maintained by Deutsches Klimarechenzentrum)
- Joel Benn (Australian Research Data Commons)
- Douglas Fils (EarthCube Science Support Office)
- Adam Shepherd (Biological and Chemical Oceanography Office (BCO-DMO))
- Matt Styles (UK Clinical Research Council - Tissue Directory and Coordination Centre)
- Julian Gautier (Harvard Dataverse)
- Europeana
- Josef Hardi, John Graybeal (Stanford Center for Biomedical Informatics Research, Stanford University)
- Bruce Wilson (ORNL Distributed Active Archive Center for Biogeochemical Dynamics)
- Steve Canham, Christian Ohmann (European Clinical Research Infrastructure Network)
- Leopold Talirz (Swiss National Centre of Competence in Research for the Computational Design and Discovery of New Materials))
- Andrea Perego (European Commission, Joint Research Centre (JRC))
- Philippe Rocca-Serra, Susanna-Assunta Sansone (Oxford e-Research Centre, University of Oxford, UK)
- Chris Hunter (GigaScience Database, China National Gene Bank (CNGB))
- DataDryad (Version re-launch this summer)
- Kerrin Borschewski (Consortium of European Social Science Data Archives, European Research Infrastructure Consortium)
- Institut national de la recherche agronomique (INRA)
- The British Library

Survey: Acknowledgement

Survey instrument design

- Leyla Garcia (Elixir Hub)
- Nick Juty (ELIXIR-UK)
- Fotis Psomopoulos (INAB|CERTH)
- Siri Jodha Khalsa (NSIDC)
- Mingfang Wu (ARDC)

Provide feedback to the instrument

- Doug Fils (Ocean Leadership)
- Joel Benn (ARDC)
- Simon Cox (CSIRO)
- Kathleen Gregory (DANS)

And to those who helped to promote the survey!!

CORBEL
Horizon 2020
654248

Survey: https://goo.gl/iLunug

RESEARCH DATA ALLIANCE

Setting up a RDA WG

Mingfang Wu

research data sharing without barriers

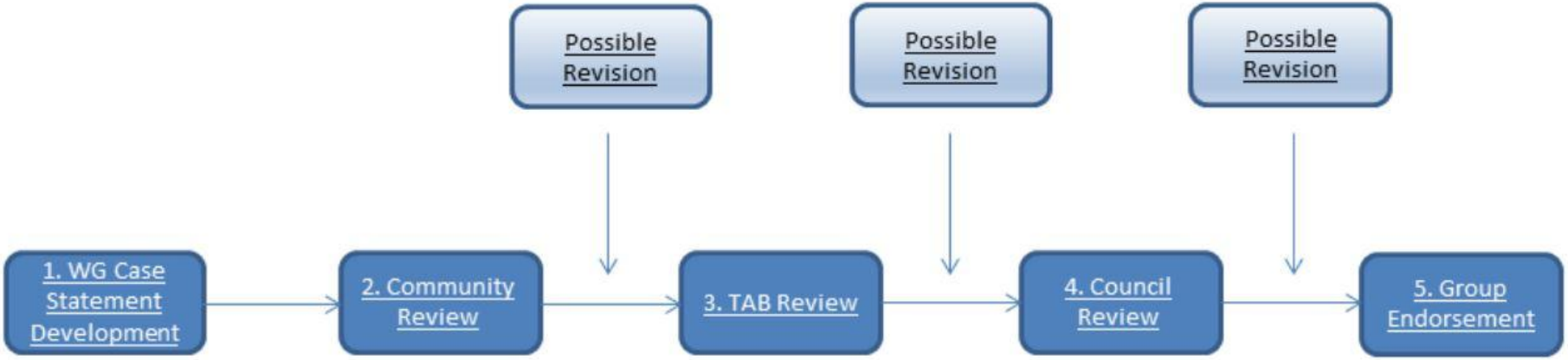
rd-alliance.org

Case statement

- Case statement content (https://goo.gl/jCin6h)
  - WG Charter
  - Value Proposition
  - Engagement with editing work in the area
  - Work plan
  - Adoption plan
  - Initial membership

The group page: https://www.rd-alliance.org/groups/research-schemas-wg

RDA WG case statement review process



https://rd-alliance.org/groups/creating-and-managing-rda-groups/creating-or-joining-rda-working-group.html

WG review criteria

- Fit with the overall RDA vision and mission
- International membership spanning, ideally, three or more continents
- 2-4 co-chairs leading the initiative
- Measurable outcomes
- Outcomes will foster data sharing and/or exchange, and be taken up by the intended community
- Proposed work, outcomes /deliverables, and Action Plan described in the Case Statement can be accomplished in 12-18 months
- Appropriate scope of the WG
- The effort adds value over and above what is currently being done within the community.

https://rd-alliance.org/groups/creating-and-managing-rda-groups/creating-or-joining-rda-working-group.html

Group Discussion

research data sharing without barriers
rd-alliance.org

Discussion

- WG Case statement
- Schemas and extensions
- Tooling (to consume, to generate)
- Guidelines
- FAIR alignment in research schemas
- ...

RESEARCH DATA ALLIANCE

Action items and next steps

research data sharing without barriers

rd-alliance.org

Closing and next step

1. Review of Actions coming out of this meeting
   - Action 1 (responsible person)
   - Action 2 (responsible person)
2. Next Steps