B2SHARE on Invenio 3

Storing Heterogeneous Metadata

Nicolas Harraudeau
CERN

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECT e-Infrastructures. Contract No. 654065

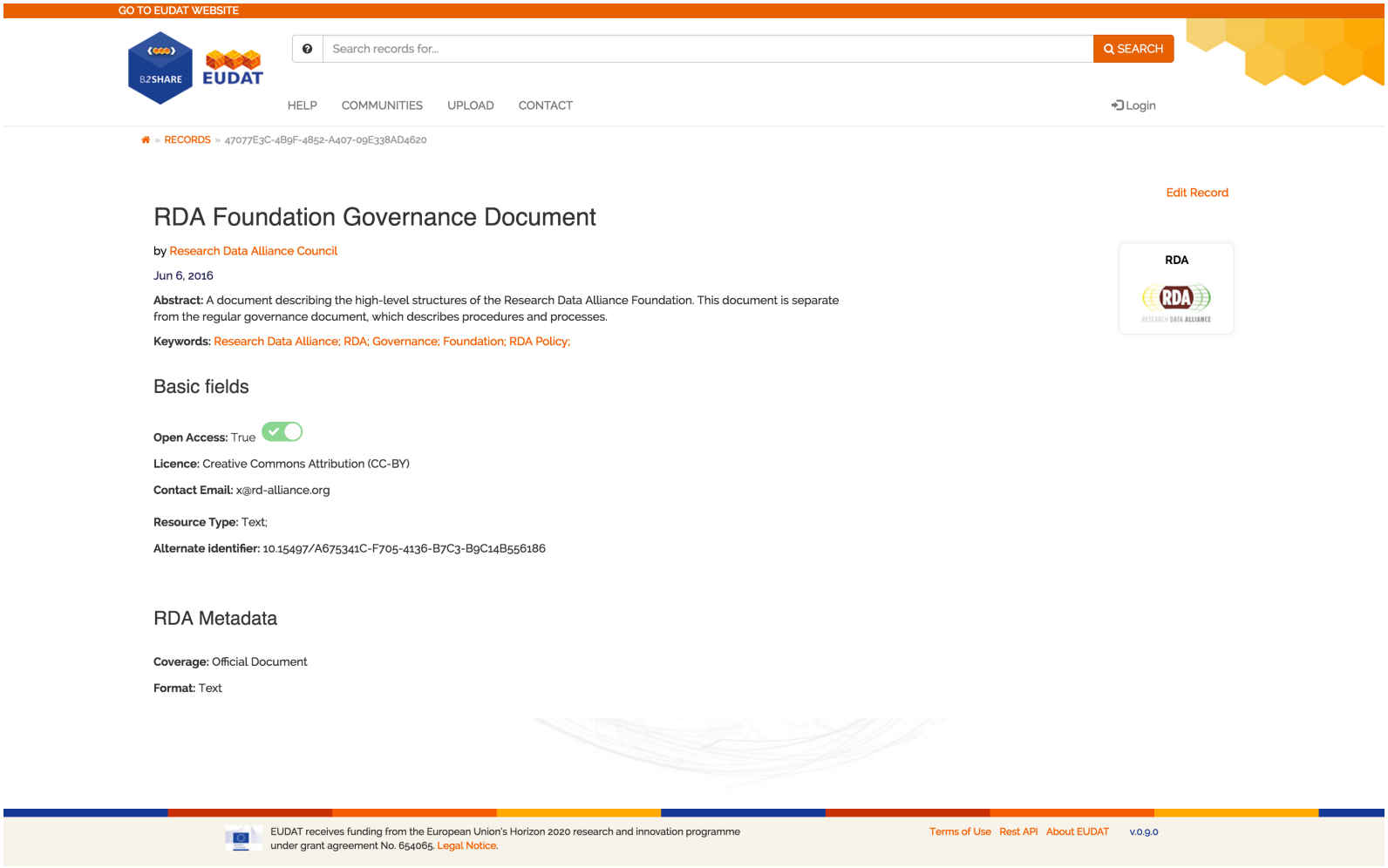
How to lose research data 101

"The odds of a data set being reported as extant fell by **17% per year**"

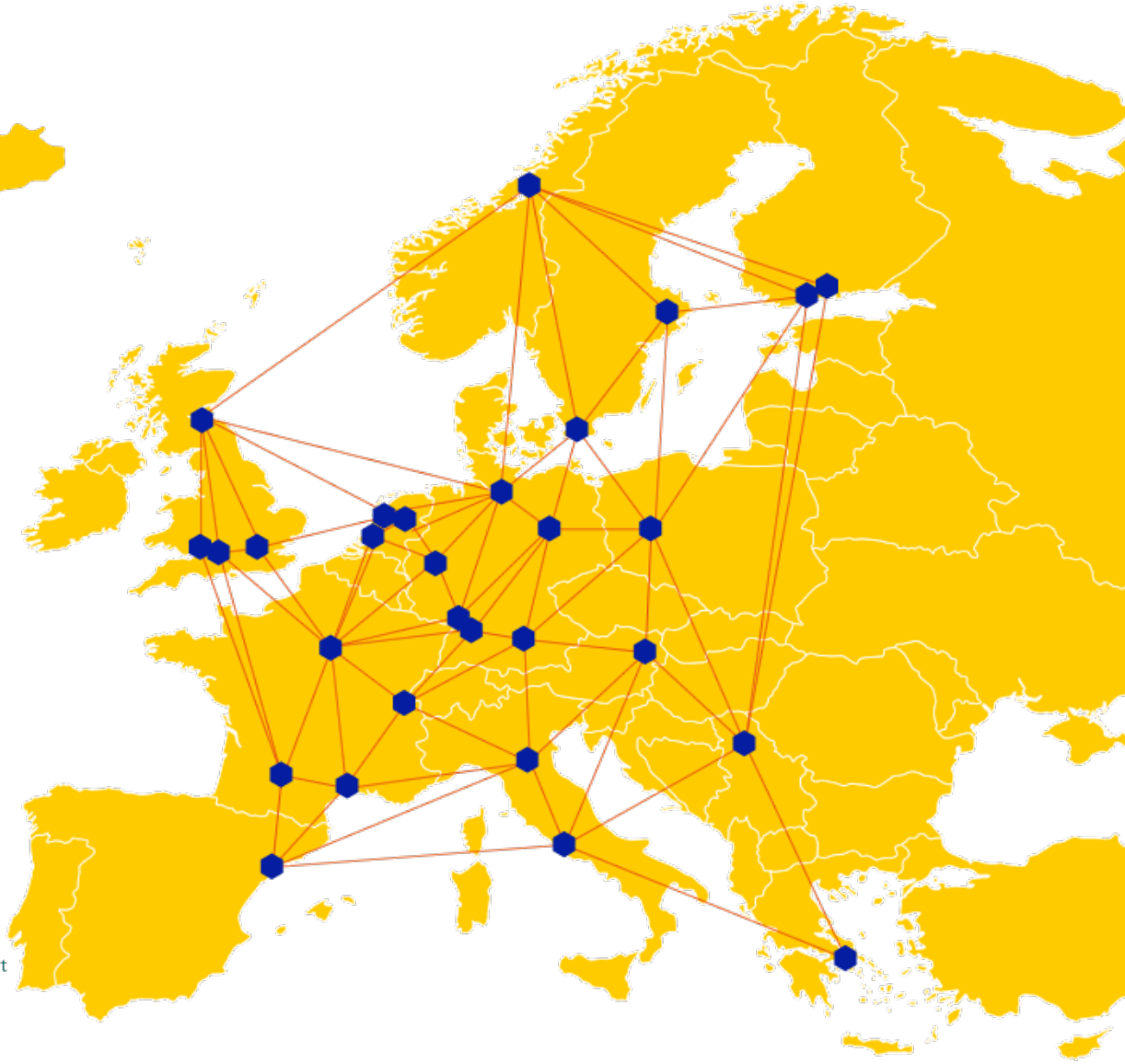
The Availability of Research Data Declines Rapidly with Article Age
October 22, 2013

B2Share

GOAL: Publish Long tail research datasets



Pan-European Data Infrastructure



Different metadata for each scientific Domain

Common Metadata

(Title, Authors, License, Open Access…)

+

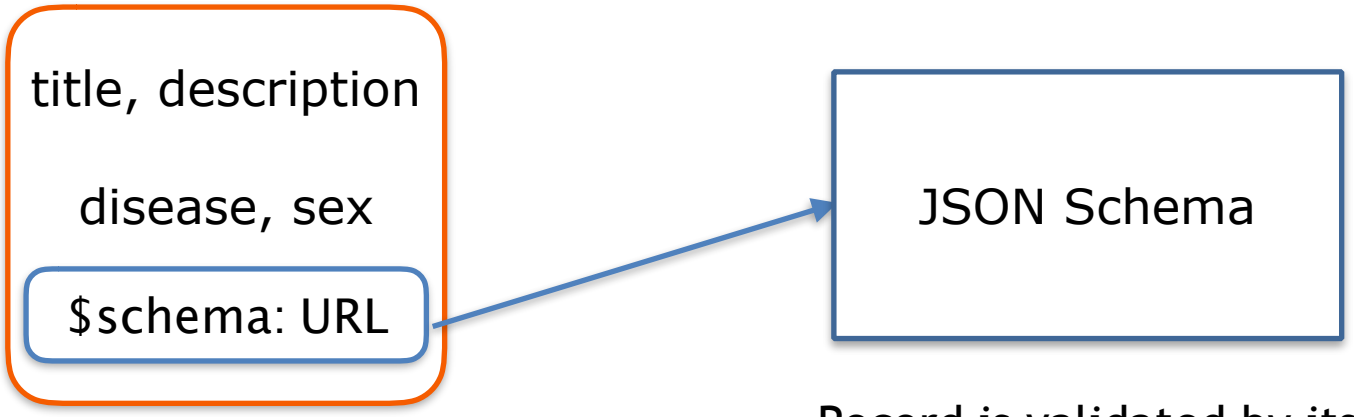
Domain specific Metadata

Language, Country/Region…	Geographic Location, Spatial Resolution…	Disease, Sex, Age, Study design…
CLARIN	**DRIHM**	**BBMRI**
Linguistic	Meteorology and Climate	Biomedical research

Invenio 3 metadata storage

- Metadata is stored as a **JSON,** which **enables any structure**.
- **Efficient storage:** PostgreSQL JSONB, Elasticsearch (Invenio 3 can still be used with MySQL)
- No conversion as the **REST API is also using JSON**
- **Conversion to/from Marc 21** is still possible by using the dojson library.

Invenio 3 metadata validation



Record

Record stores the reference to its JSON Schema

Record is validated by its Schema when it is CREATED and UPDATED

JSON Schemas 101 - Vehicles

```
1 {
2     "vehicle type": "motorcycle",
3     "number of wheels": 3
4 }
```

author: Steindy (Wikipedia)

```
1 {
2     "vehicle type": "car",
3     "number of wheels": 3
4 }
```

author: Brian Snelson (Wikipedia)

```
1 {
2     "vehicle type": "car",
3     "number of wheels": 8
4 }
```

author: Anetode (Wikipedia)

JSON Schemas 101 - Vehicles

```
{
    "allOf": [
        {
            "properties": {
                "vehicle type": { "type": "string" },
                "number of wheels": { "type": "integer" }
            }
        },
        {
            "oneOf": [
                {
                    "properties": {
                        "vehicle type": { "enum": ["motorcycle"] },
                        "number of wheels": { "minimum": 2, "maximum": 3 }
                    }
                },
                {
                    "properties": {
                        "vehicle type": { "enum": ["car"] },
                        "number of wheels": { "minimum": 3 }
                    }
                }
            ]
        }
    ]
}
```

JSON Schemas 101 - Vehicles

```
1  {
2      "allOf": [
3          {
4              "properties": {
5                  "vehicle type": { "type": "string" },
6                  "number of wheels": { "type": "integer" }
7              }
8          },
9          {
10             "oneOf": [
11                 {
12                     "properties": {
13                         "vehicle type": { "enum": ["motorcycle"] },
14                         "number of wheels": { "minimum": 2, "maximum": 3 }
15                     }
16                 },
17                 {
18                     "properties": {
19                         "vehicle type": { "enum": ["car"] },
20                         "number of wheels": { "minimum": 3 }
21                     }
22                 }
23             ]
24         }
25     ]
26 }
```

A

B

1

2

A and B => A and (1 xor 2)

People have to agree on a schema

POPULAR SCIENCE

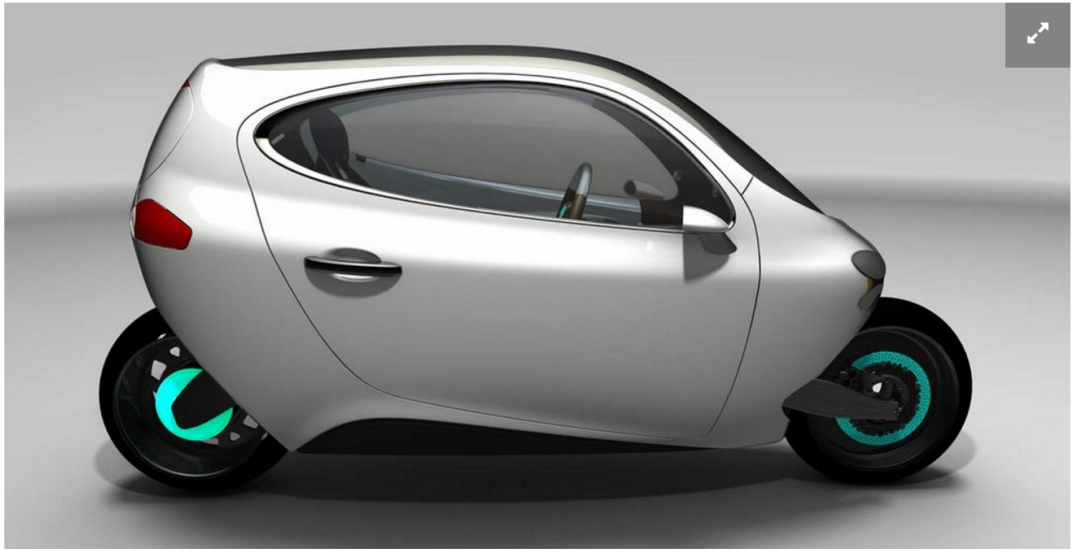
TRENDING: APPLE DRONES MARS NASA A.I. MORE SHOP SUBSCRIBE

CARS

A SELF-BALANCING, TWO-WHEELED CAR

ONE OF TEN BRILLIANT INNOVATIONS FROM OUR 2015 INVENTION AWARDS

By James Vlahos Posted May 13, 2015



WANT MORE NEWS LIKE THIS?

http://www.popsci.com/self-balancing-two-wheeled-car

People have to agree on a schema

POPULAR SCIENCE

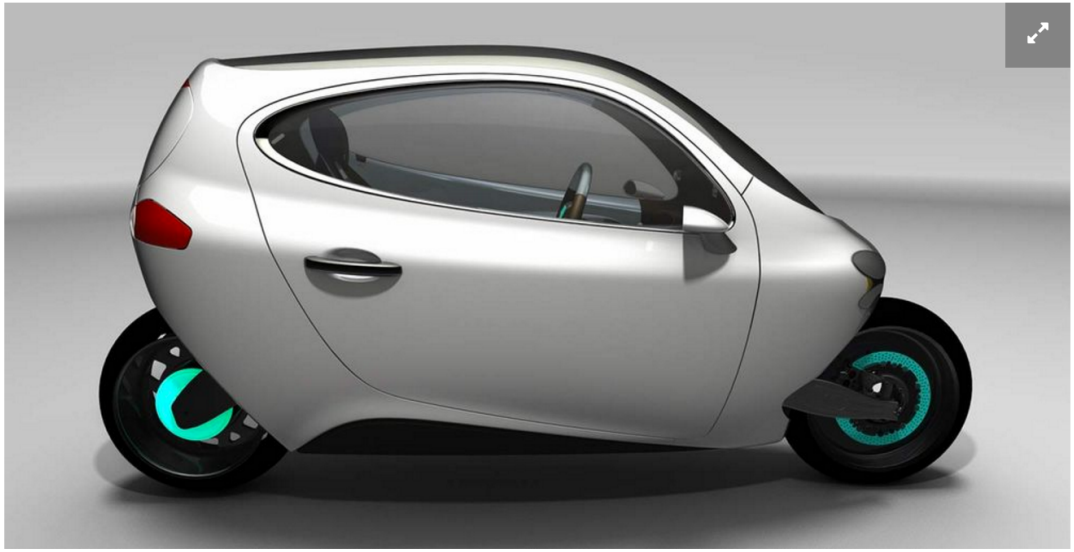
TRENDING: APPLE DRONES MARS NASA A.I. MORE SHOP SUBSCRIBE

CARS

A SELF-BALANCING, TWO-WHEELED CAR

ONE OF TEN BRILLIANT INNOVATIONS FROM OUR 2015 INVENTION AWARDS

By James Vlahos Posted May 13, 2015



```
...
"properties": {
    "vehicle type": {
        "enum": ["car"]
    },
    "number of wheels": {
        "minimum": 2
    }
}
...
```

WANT MORE NEWS LIKE THIS?

http://www.popsci.com/self-balancing-two-wheeled-car

People have to agree on a schema

POPULAR SCIENCE

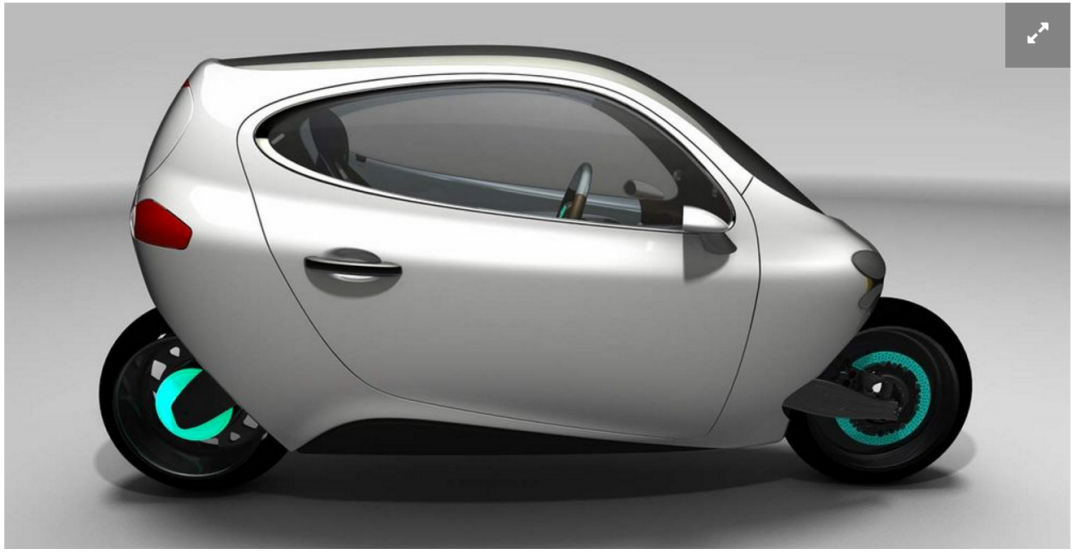
TRENDING: APPLE DRONES MARS NASA A.I. MORE SHOP SUBSCRIBE

CARS

A SELF-BALANCING, TWO-WHEELED CAR

ONE OF TEN BRILLIANT INNOVATIONS FROM OUR 2015 INVENTION AWARDS

By James Vlahos Posted May 13, 2015



```
...
"properties": {
    "vehicle type": {
        "enum": ["car"]
    },
    "number of wheels": {
        "minimum": 2
    }
}
...
```

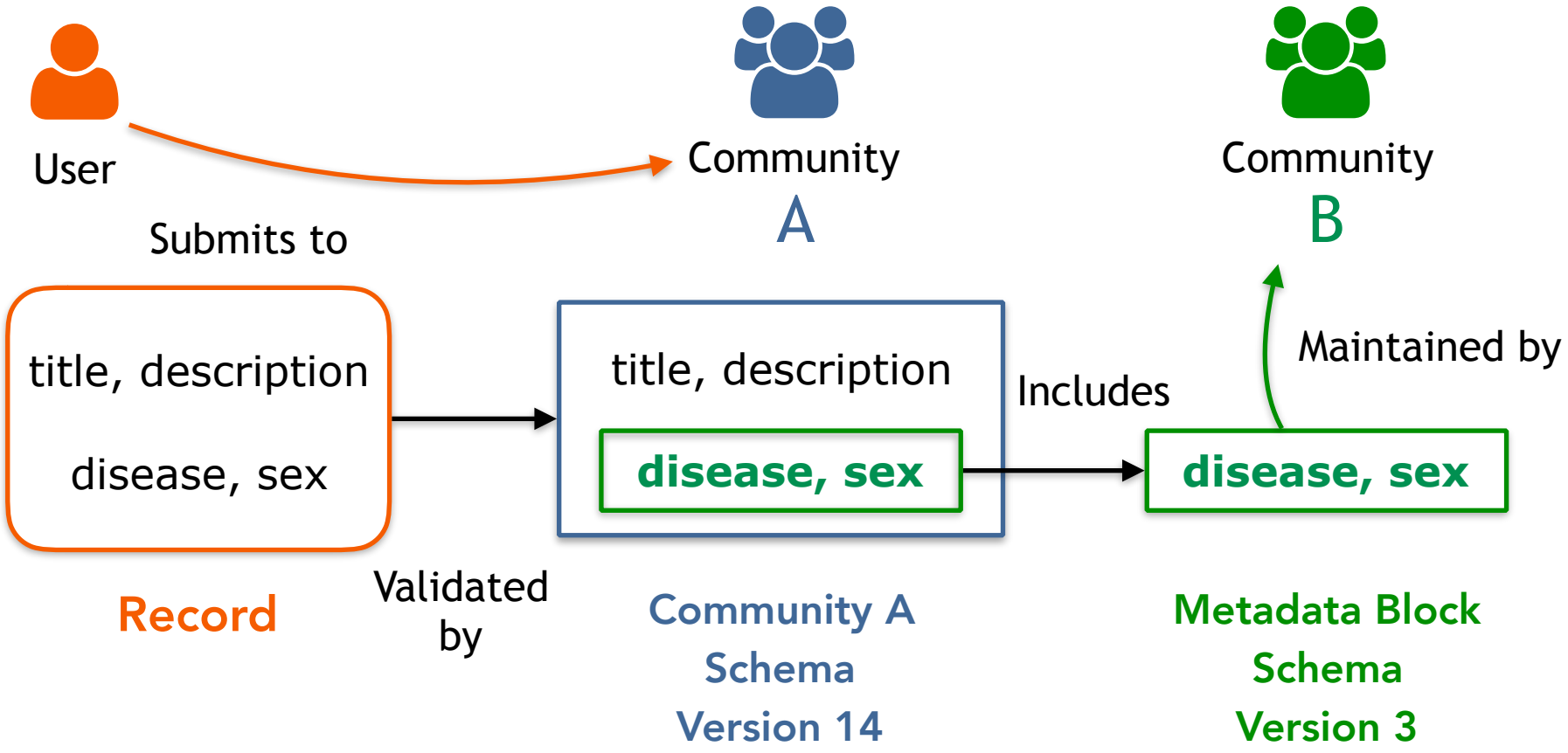
For now...

WANT MORE NEWS LIKE THIS?

http://www.popsci.com/self-balancing-two-wheeled-car

B2Share's composite schemas

Every Community has **one versioned schema**. It is composed of a **common fields** and a set of **metadata block schemas** which are maintained by other (or the same) communities.



B2Share's composite schemas

GET "http://.../api/communities/<community_ID>/<version_nb>"

```
1  {
2      "properties": {
3
4          "title": "...",
5          "description": "...",
6          ...
7
8          "community_specific": {
9
10             "<schema_ID1>": {
11                 "$ref": "http://.../api/schemas/<schema_ID1>/versions/5#/json_schema"
12             },
13 Metadata
14   Blocks    "<schema_ID2>": {
15                 "$ref": "http://.../api/schemas/<schema_ID2>/versions/3#/json_schema"
16             },
17             ...
18
19         }
20     }
21 }
```

Fields common to every dataset

Schema backward compatibility

- Community Schemas and Metadata Block schemas have to be **backward compatible**.

  i.e: Any **JSON Pointer** (path) should always resolve the **same field type** whatever the schema version.

  Consequence: If one wants to **change the structure**, he has to either **rename** the field and **deprecate** the old field name, or **create a new schema**.

- doschema library enables to check schemas backward compatibility

Schema backward compatibility

"http://.../api/schemas/<schema_ID1>/versions/1#/json_schema"

```
1 {
2     "properties": {
3
4         "disease": "...",
5         "sex": "...",
6         ...
7     }
8 }
```

type: "string"

"http://.../api/schemas/<schema_ID1>/versions/2#/json_schema"

```
1 {
2     "properties": {
3
4         "disease": {
5             "name": "..."
6         },
7         "sex": "...",
8         ...
9     }
10 }
```

type: "object" => INCOMPATIBLE

Possible next steps

- **Support better mapping to existing standards** by using **JSON-LD** in schemas and/or metadata.
- Enable easy **schema bootstrap** from existing standards when they use JSON-LD or RDF.
- **Publish the schemas** on EUDAT **Data Type Registry**.

What B2Share will NOT do

- Catalogue all existing standards (Other tools exist for that, ex: biosharing.org, Data Type Registry…)
- Use JSON Schemas as a replacement for existing standards.
- Use JSON Schema for everything. Ex: linking data is better done with JSON-LD.

Links and bibliography

- EUDAT eudat.org
- B2Share b2share.eudat.eu
- Invenio http://invenio-software.org/
- JSON Schema http://json-schema.org/
- JSON-LD's JSON Schema http://json-ld.org/schemas/jsonld-schema.json

- *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project* (Nature Biotechnology) doi: http://dx.doi.org/10.1038/nbt.1411
- *The Availability of Research Data Declines Rapidly with Article Age* (Cell) doi: http://dx.doi.org/10.1016/j.cub.2013.11.014
- *Medical science experiments and research lost in University of Leicester blaze* (Leicester Mercury) http://www.leicestermercury.co.uk/Medical-science-experiments-research-lost/story-29163384-detail/story.html