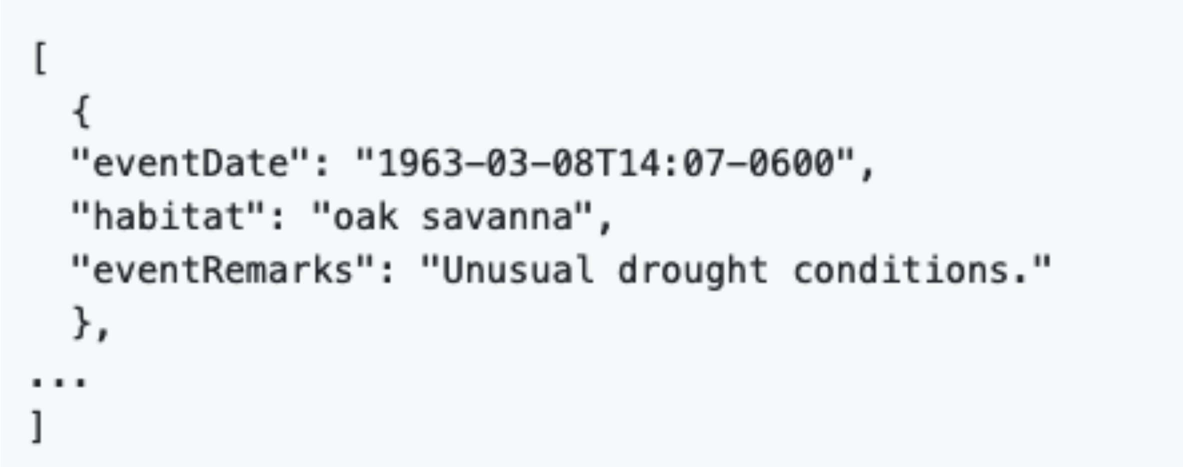
Complex values in JSON and CSV

Steve Baskauf – TDWG TAG meeting – 2023-05-08

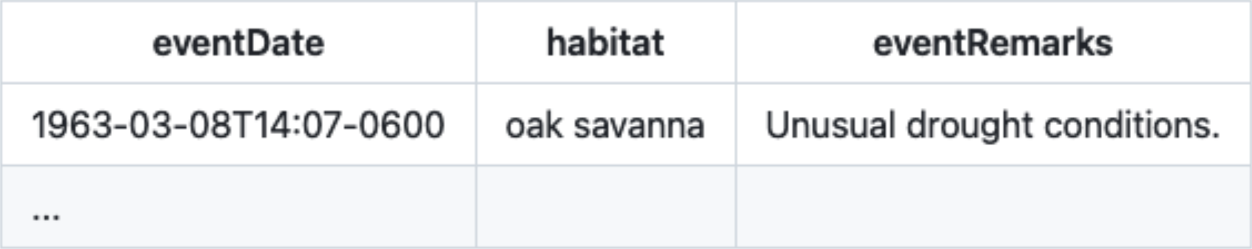


```
[
  {
  "eventDate": "1963-03-08T14:07-0600",
  "habitat": "oak savanna",
  "eventRemarks": "Unusual drought conditions."
  },
...
]
```

JSON

We want all of these representations to "mean" the same thing (identical structure and semantics).

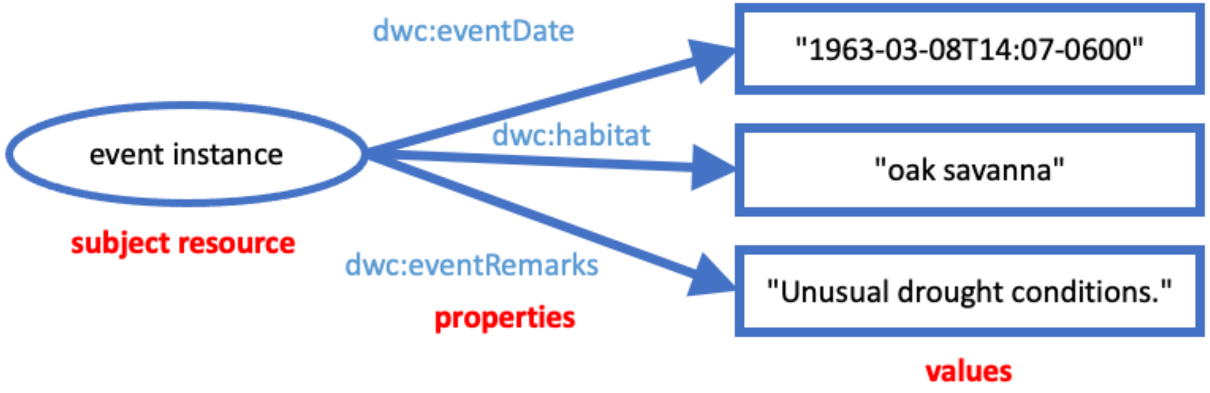
=



eventDate	habitat	eventRemarks
1963-03-08T14:07-0600	oak savanna	Unusual drought conditions.
...		

CSV (delineated text)

=



Event class example

```
@prefix dwc: <http://rs.tdwg.org/dwc/terms/> .
[] dwc:eventDate "1963-03-08T14:07-0600"^^xsd:dateTime;
   dwc:habitat "oak savanna";
   dwc:eventRemarks "Unusual drought conditions.".
```

Linked Data graph (RDF)

JSON-LD strategy for vanilla JSON

```
{
    "@context": {
        "dwc": "http://rs.tdwg.org/dwc/terms/"
        },
    "dwc:preparations": [
        "skin",
        "skull",
        "skeleton"
    ]
}
```

JSON-LD with CURIEs

- Semantics are clarified by using globally unique TDWG IRIs for vocabulary terms.
- Graph structure specified by hierarchical JSON structure.

```
{
    "@context": {
        "preparations": "http://rs.tdwg.org/dwc/terms/preparations"
        },
    "preparations": [
        "skin",
        "skull",
        "skeleton"
    ]
}
```

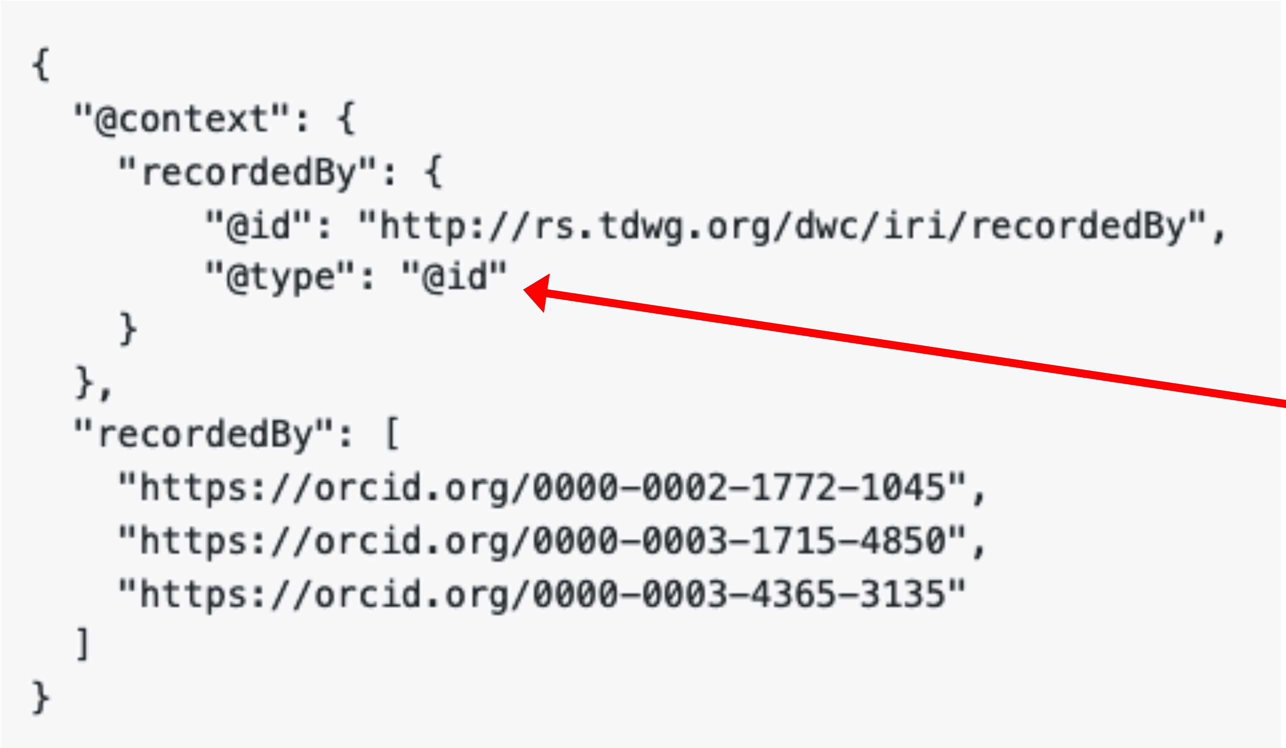
JSON-LD without CURIEs (locally defined context)

```
{
    "@context":"http://rs.tdwg.org/contexts/dwc.json",
    "preparations": [
        "skin",
        "skull",
        "skeleton"
    ]
}
```

JSON-LD without CURIEs (externally defined context)

```
{
    "@context": {
        "dwc": "http://rs.tdwg.org/dwc/terms/"
        },
    "preparations": {
        "@id": "dwc:preparations"
        }
}
```

external file http://rs.tdwg.org/contexts/dwc.json



```
{
  "@context": {
    "recordedBy": {
        "@id": "http://rs.tdwg.org/dwc/iri/recordedBy",
        "@type": "@id"
    }
  },
  "recordedBy": [
    "https://orcid.org/0000-0002-1772-1045",
    "https://orcid.org/0000-0003-1715-4850",
    "https://orcid.org/0000-0003-4365-3135"
  ]
}
```

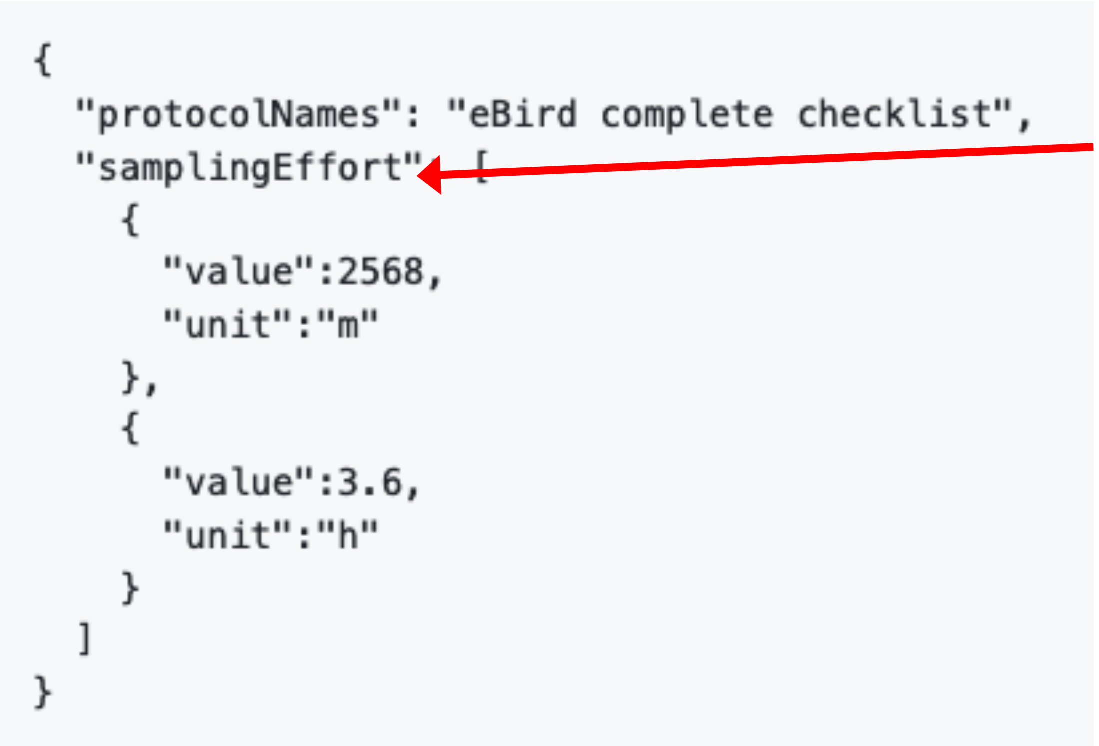
Context can clarify IRI-valued terms, datatypes, and languages of literal strings.

```
{
    "@context": {
        "providerLiteral": "http://rs.tdwg.org/ac/terms/providerLiteral",
        "provider": {
            "@id": "http://rs.tdwg.org/ac/terms/provider",
            "@type": "@id"
        },
        "recordedBy": "http://rs.tdwg.org/dwc/terms/recordedBy",
        "recordedByIRI": {
            "@id": "http://rs.tdwg.org/dwc/iri/recordedBy",
            "@type": "@id"
        },
    },
    "recordedBy": "Carol J. Baskauf",
    "recordedByIRI": "https://orcid.org/0000-0003-1715-4850",
    "providerLiteral": "Bioimages",
    "provider": "http://bioimages.vanderbilt.edu/"
}
```

DwC dual namespace design precludes a "local name" default for JSON names

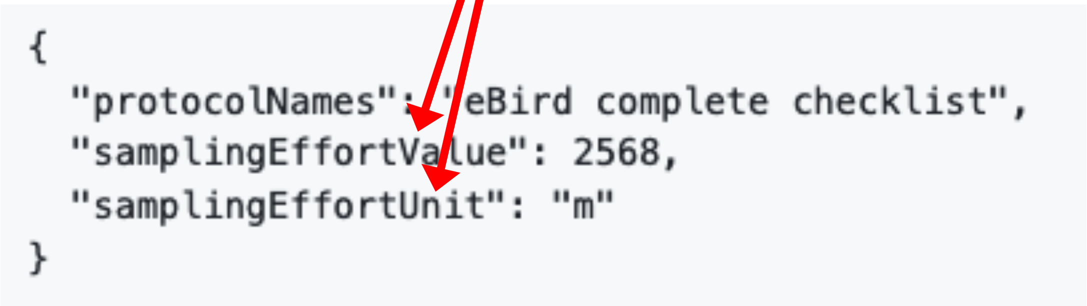
Issues arising with multiple values

- Do we always require array values any time a property might have multiple values?
- Do we define term variants if we allow either single or multiple values for a complex value type?



```
{
  "protocolNames": "eBird complete checklist",
  "samplingEffort": [
    {
      "value":2568,
      "unit":"m"
    },
    {
      "value":3.6,
      "unit":"h"
    }
  ]
}
```

term variants for single and multiple values



```
{
  "protocolNames": "eBird complete checklist",
  "samplingEffortValue": 2568,
  "samplingEffortUnit": "m"
}
```

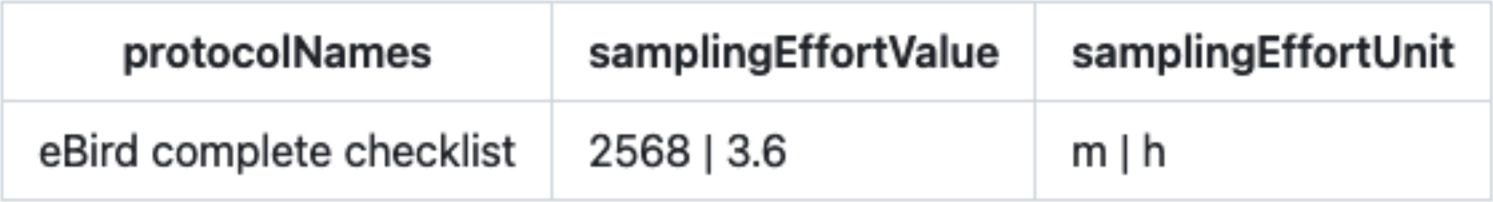
Fielded text issues

Problems

- Differences in semantics of multiple IRI- and non-IRI-valued properties (is "space pipe space" appropriate for dwciri: namespace terms?)
- How do we facilitate multiple values for complex values that are actually instances of another class? (Humbold Extension issue)

Multiple values for complex values problem

- Specifically, https://github.com/tdwg/tag/issues/43

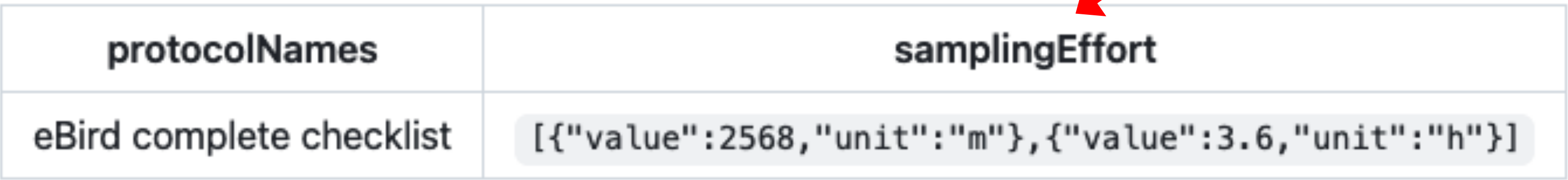


protocolNames	samplingEffortValue	samplingEffortUnit
eBird complete checklist	2568 | 3.6	m | h

unclear relationship between multiple values that are paired

"JSON in a box" solution

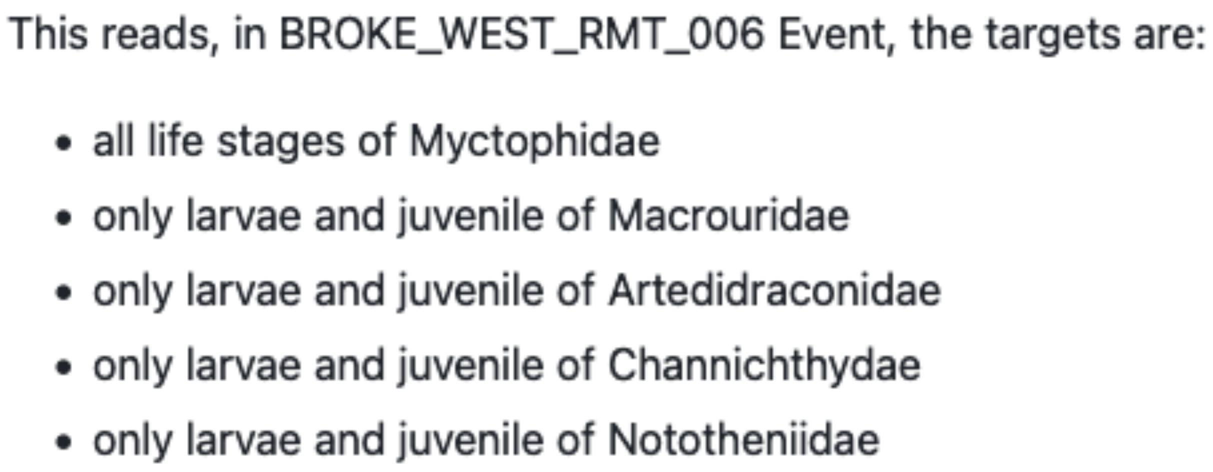
Note term variant for multiple values



protocolNames	samplingEffort
eBird complete checklist	`[{"value":2568,"unit":"m"},{"value":3.6,"unit":"h"}]`

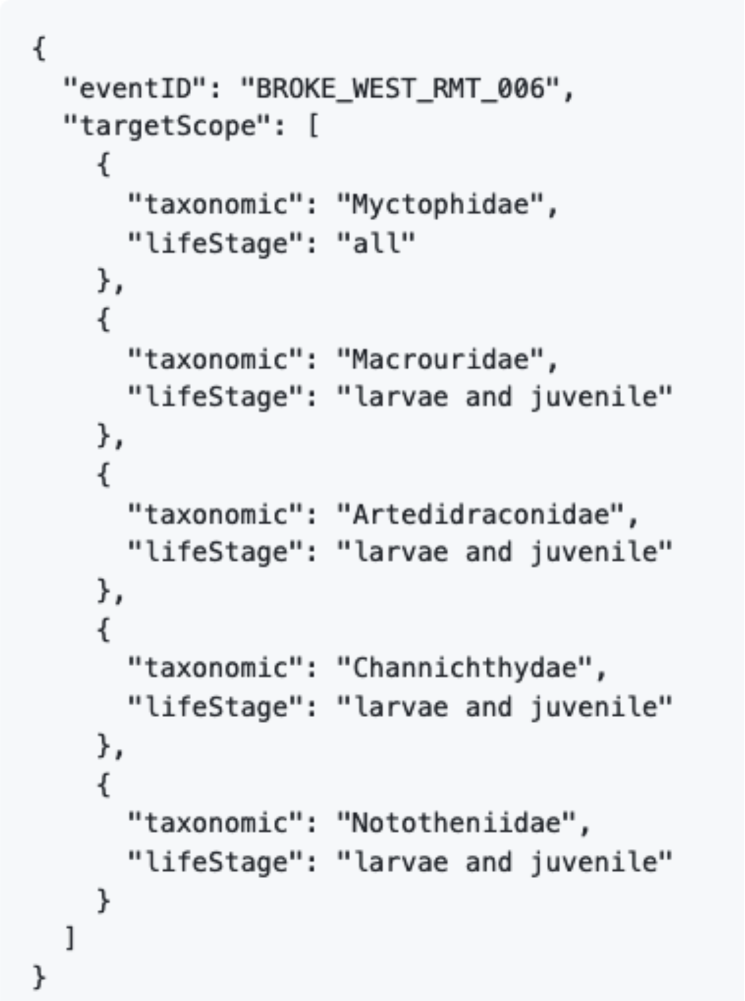
similar to dwc:dynamicProperties values

Messy if many values or complex values



This reads, in BROKE_WEST_RMT_006 Event, the targets are:

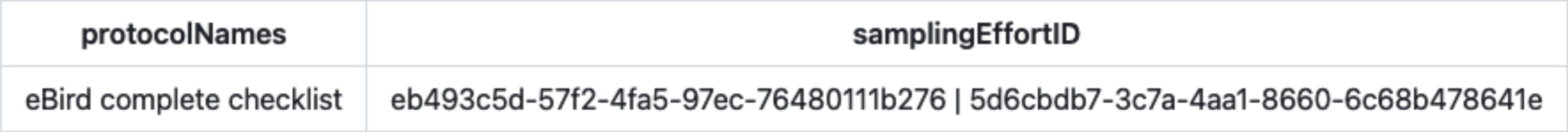
- all life stages of Myctophidae
- only larvae and juvenile of Macrouridae
- only larvae and juvenile of Artedidraconidae
- only larvae and juvenile of Channichthydae
- only larvae and juvenile of Nototheniidae



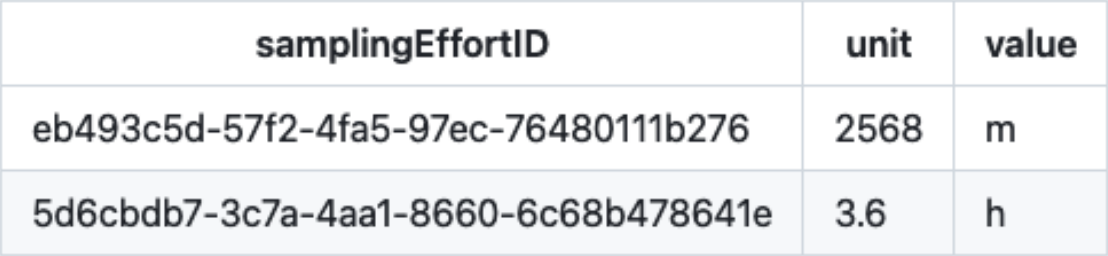
```
{
  "eventID": "BROKE_WEST_RMT_006",
  "targetScope": [
    {
      "taxonomic": "Myctophidae",
      "lifeStage": "all"
    },
    {
      "taxonomic": "Macrouridae",
      "lifeStage": "larvae and juvenile"
    },
    {
      "taxonomic": "Artedidraconidae",
      "lifeStage": "larvae and juvenile"
    },
    {
      "taxonomic": "Channichthydae",
      "lifeStage": "larvae and juvenile"
    },
    {
      "taxonomic": "Nototheniidae",
      "lifeStage": "larvae and juvenile"
    }
  ]
}
```

Would we really cram all of this into a cell? Humans couldn't do it. Scripts could read and write no matter how large and complex.

"ID terms" solution



protocolNames	samplingEffortID
eBird complete checklist	eb493c5d-57f2-4fa5-97ec-76480111b276 | 5d6cbdb7-3c7a-4aa1-8660-6c68b478641e

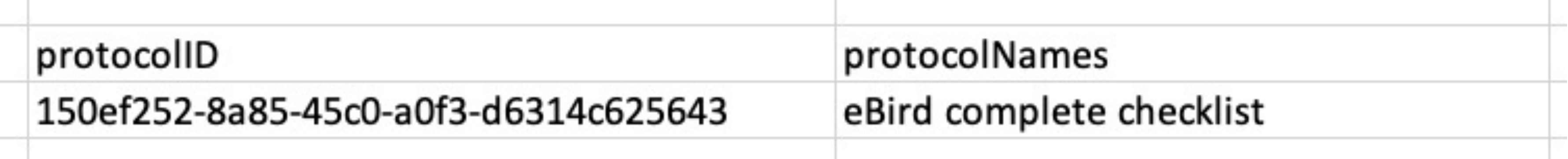


samplingEffortID	unit	value
eb493c5d-57f2-4fa5-97ec-76480111b276	2568	m
5d6cbdb7-3c7a-4aa1-8660-6c68b478641e	3.6	h

term variant required for multiple values (`samplingEffordID` vs. `samplingEffortValue` and `samplingEffortUnit`)

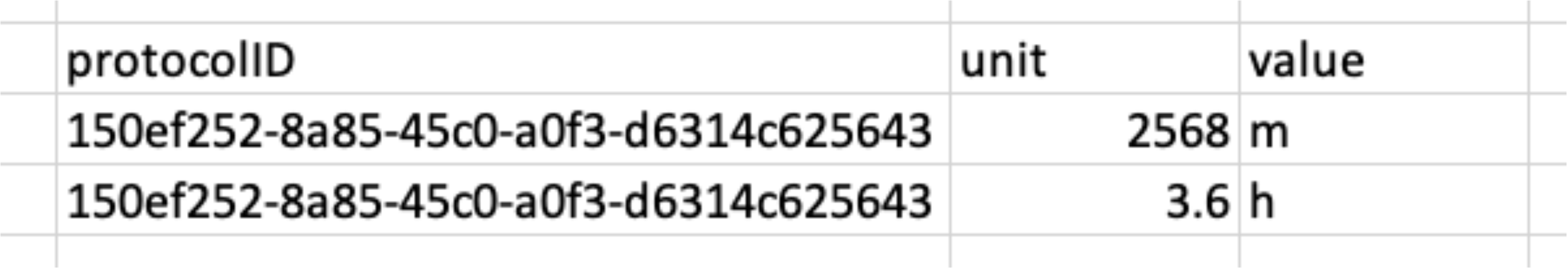
"star schema" solution

In this case, the `samplingEffort` term(s) doesn't exist – the link is implicit in the `meta.xml` file or equivalent.



protocolID	protocolNames
150ef252-8a85-45c0-a0f3-d6314c625643	eBird complete checklist

protocol "core" file



protocolID	unit	value
150ef252-8a85-45c0-a0f3-d6314c625643	2568	m
150ef252-8a85-45c0-a0f3-d6314c625643	3.6	h

measurement "extension" file