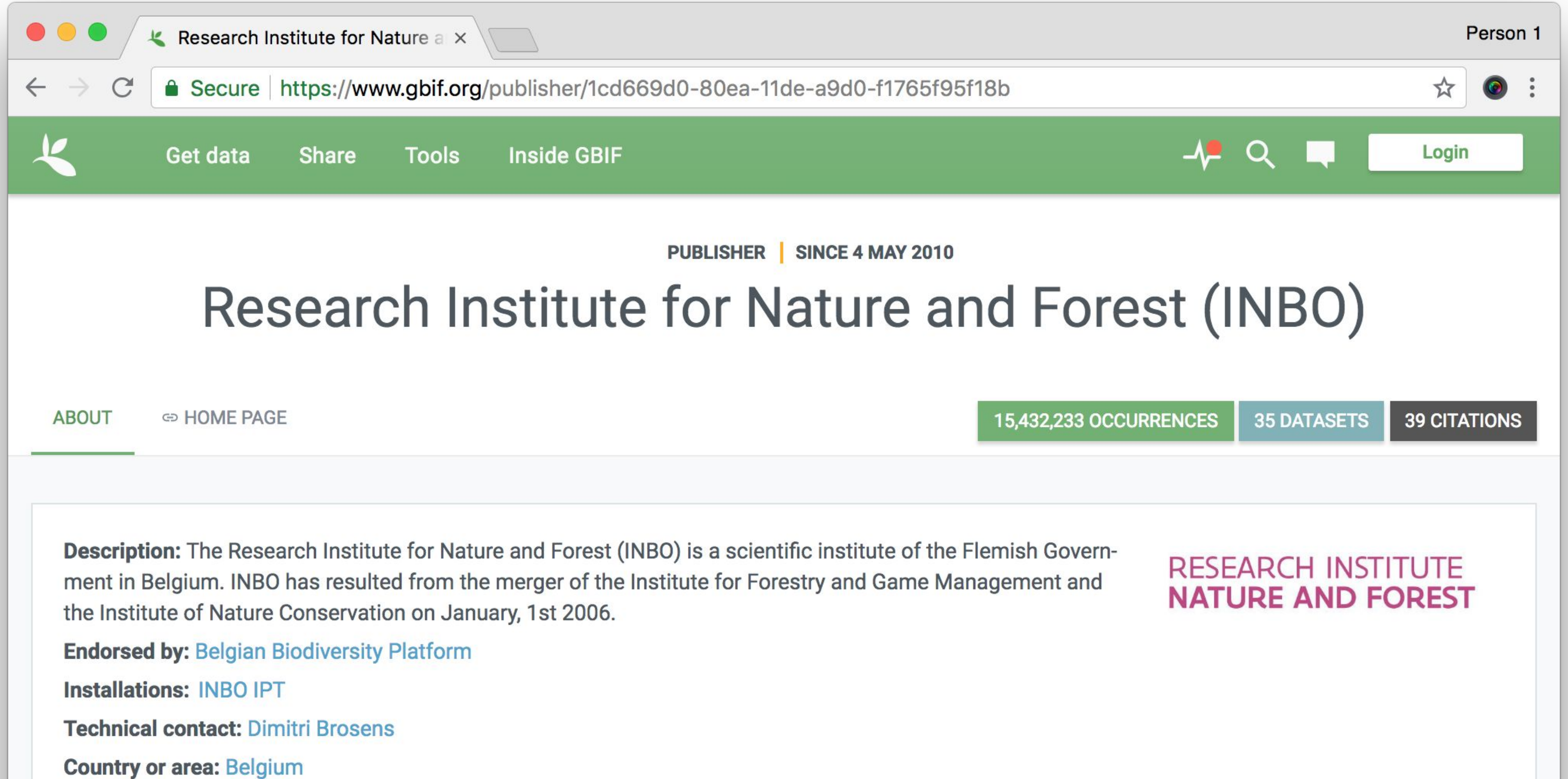
Whip

Human and machine-readable specifications for data

Stijn Van Hoey & Peter Desmet

We are a data publisher



Research Institute for Nature a ×
Person 1
Secure
https://www.gbif.org/publisher/1cd669d0-80ea-11de-a9d0-f1765f95f18b
Get data
Share
Tools
Inside GBIF
Login

PUBLISHER | SINCE 4 MAY 2010

Research Institute for Nature and Forest (INBO)

ABOUT
HOME PAGE

15,432,233 OCCURRENCES | 35 DATASETS | 39 CITATIONS

Description: The Research Institute for Nature and Forest (INBO) is a scientific institute of the Flemish Government in Belgium. INBO has resulted from the merger of the Institute for Forestry and Game Management and the Institute of Nature Conservation on January, 1st 2006.

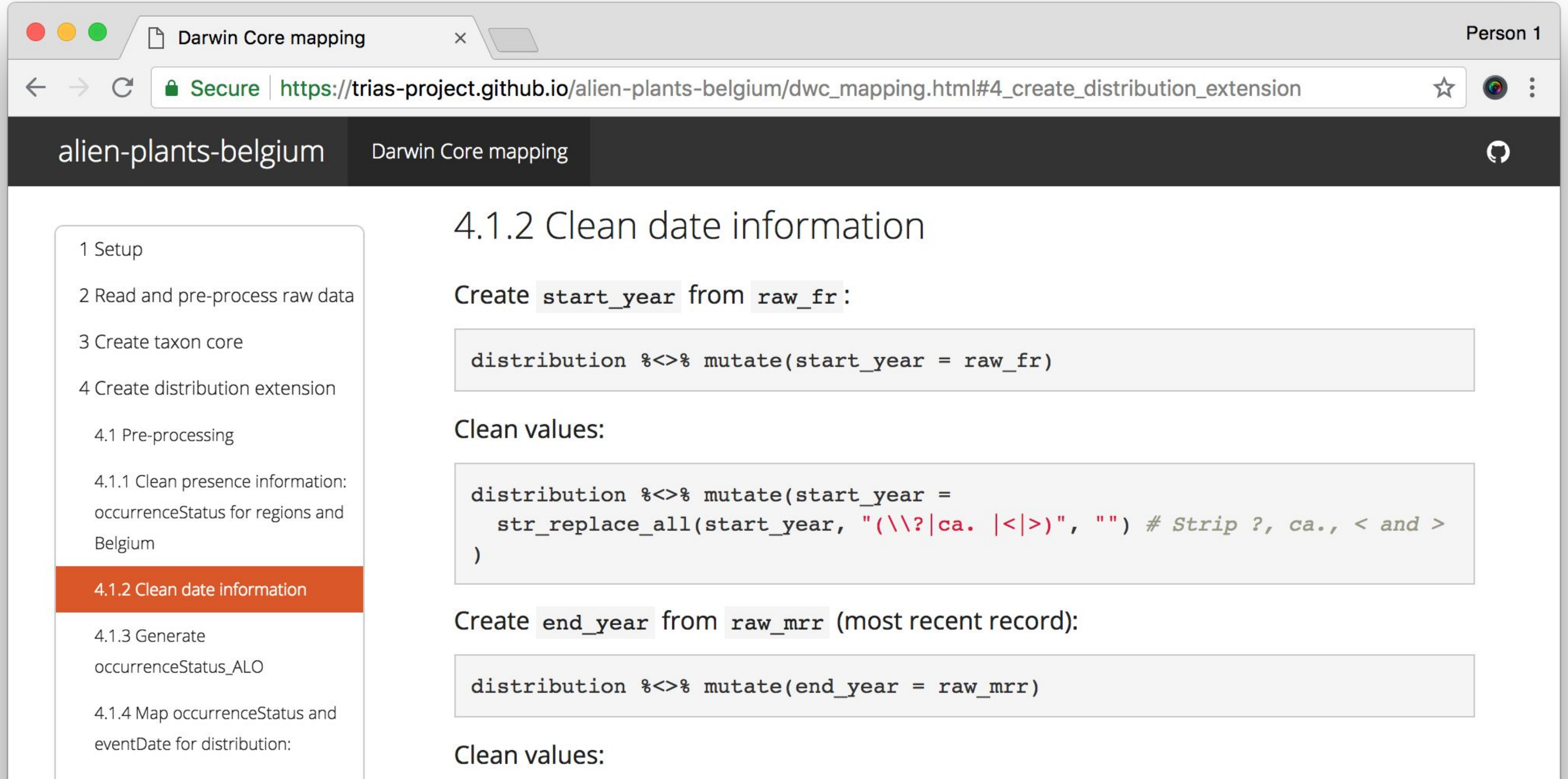
Endorsed by: Belgian Biodiversity Platform

Installations: INBO IPT

Technical contact: Dimitri Brosens

Country or area: Belgium

We care


Darwin Core mapping
Person 1
Secure
https://trias-project.github.io/alien-plants-belgium/dwc_mapping.html#4_create_distribution_extension
alien-plants-belgium
Darwin Core mapping

1 Setup
2 Read and pre-process raw data
3 Create taxon core
4 Create distribution extension
4.1 Pre-processing
4.1.1 Clean presence information: occurrenceStatus for regions and Belgium
4.1.2 Clean date information
4.1.3 Generate occurrenceStatus_ALO
4.1.4 Map occurrenceStatus and eventDate for distribution:

4.1.2 Clean date information

Create `start_year` from `raw_fr`:

```
distribution %<>% mutate(start_year = raw_fr)
```

Clean values:

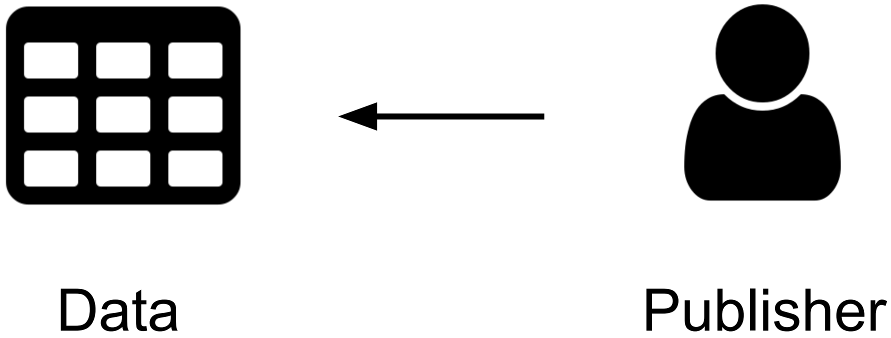
```
distribution %<>% mutate(start_year =
  str_replace_all(start_year, "(\\?|ca. |<|>)", "") # Strip ?, ca., < and >
)
```

Create `end_year` from `raw_mrr` (most recent record):

```
distribution %<>% mutate(end_year = raw_mrr)
```

Clean values:

What to expect



Data

Publisher

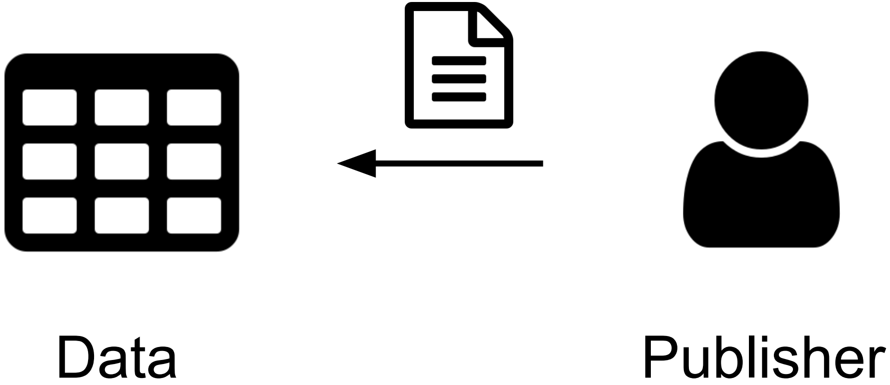
What to expect

Data quality

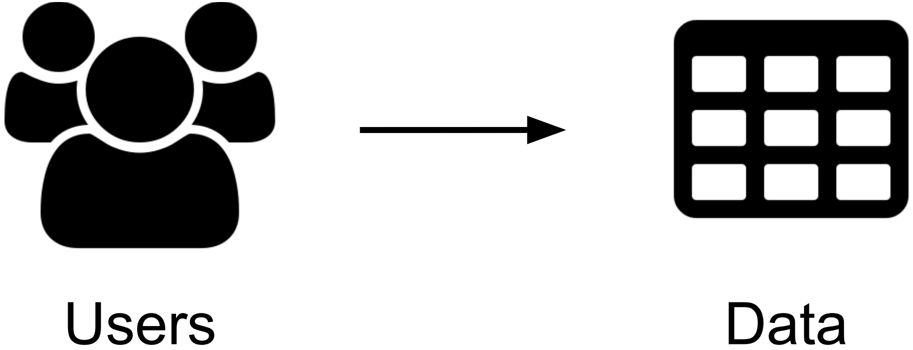
Standardization

Community recommendations

Dataset characteristics



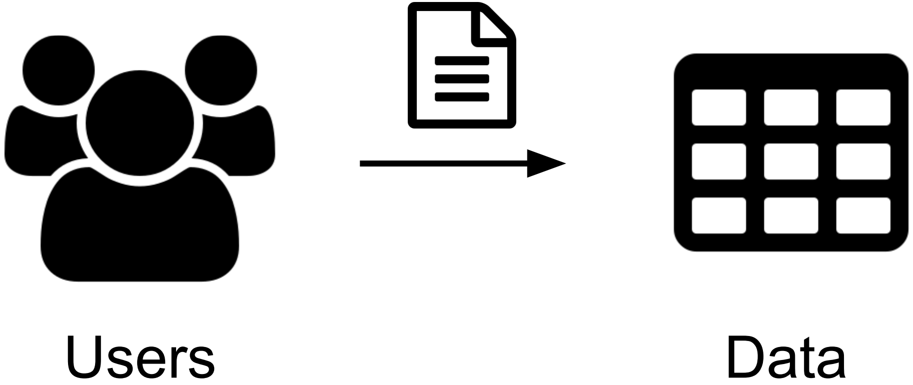
Expectations



Users

Data

Expectations



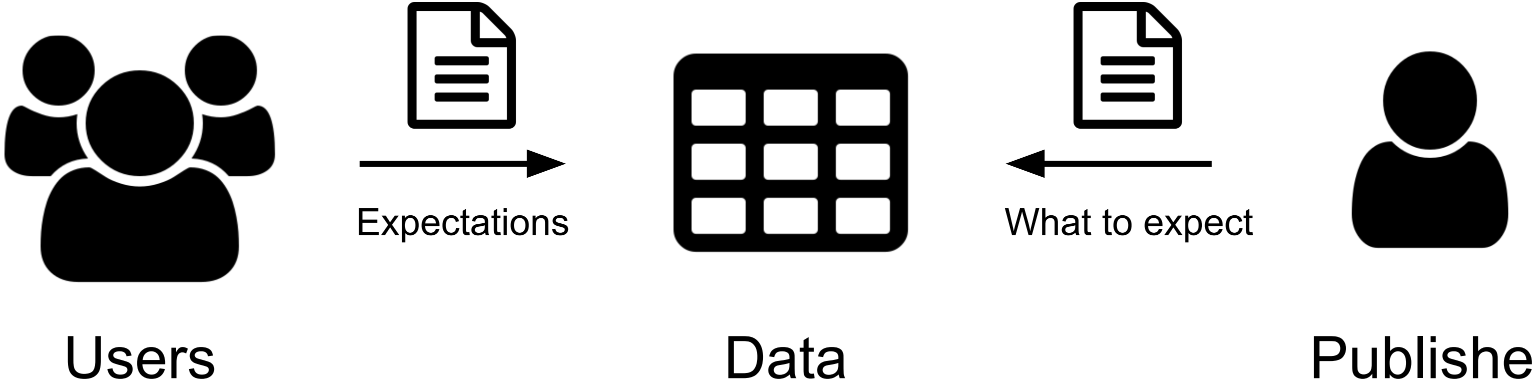
Users

Data

Fit for my research?

Fit for specific user community?

Expectations / What to expect



Expectations

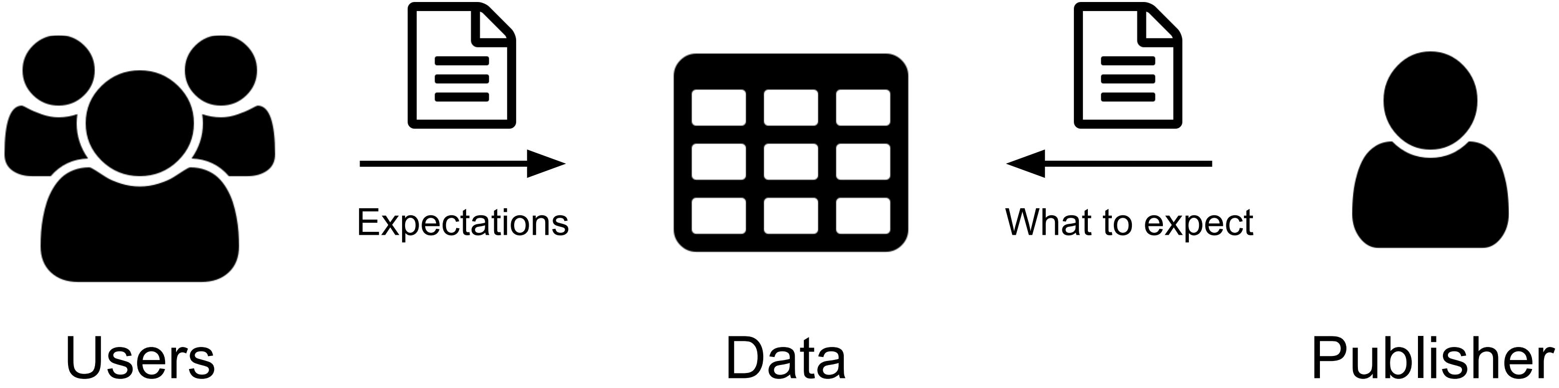
What to expect

Users

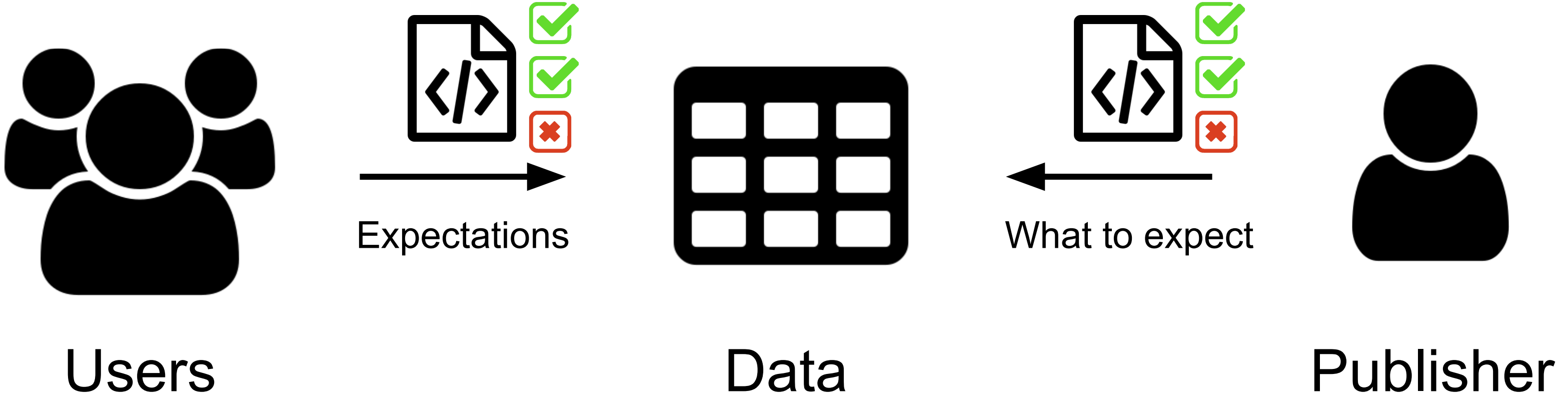
Data

Publisher

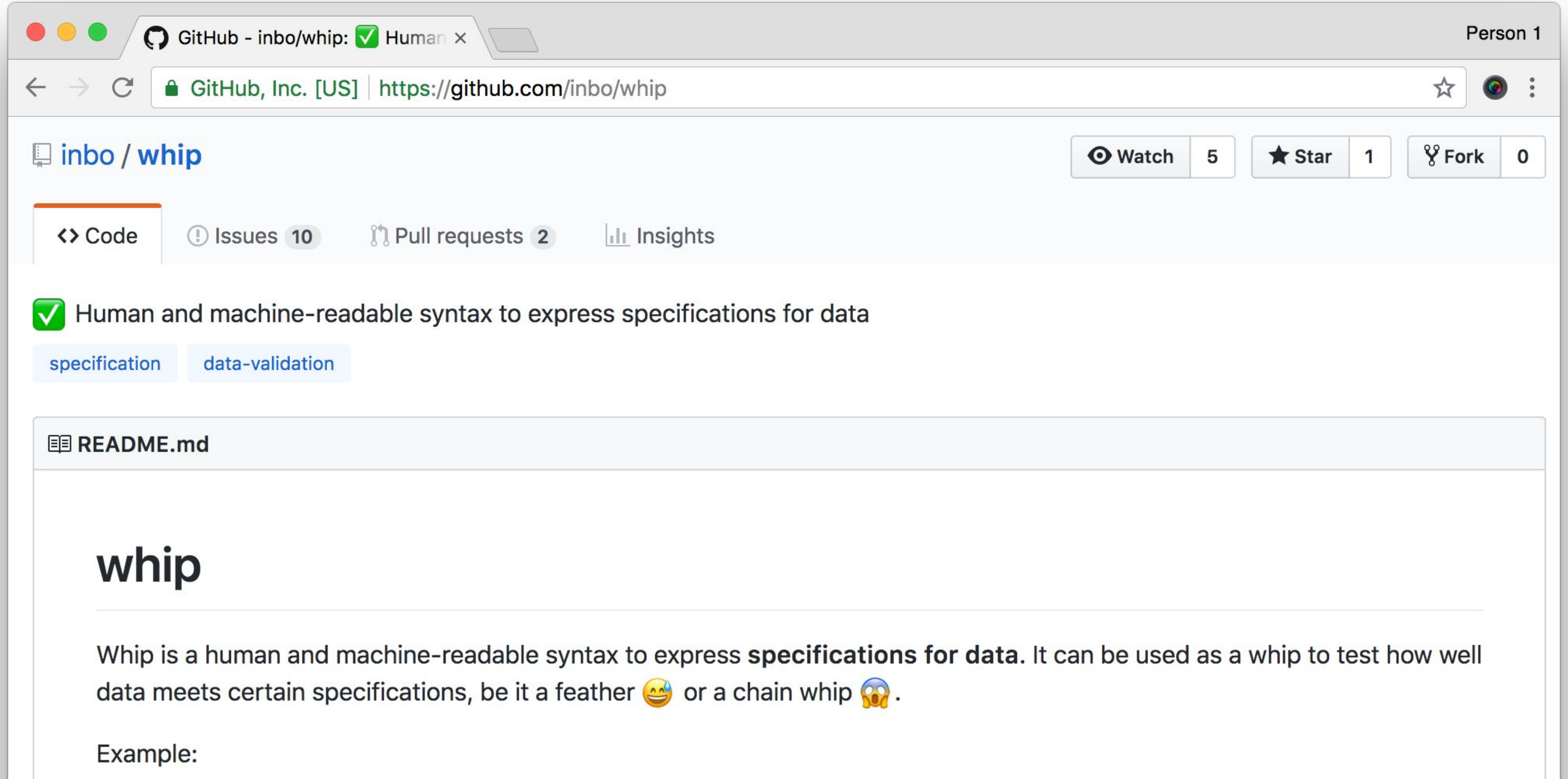
How to communicate expectations?



How to test expectations?



Whip



✅ Human and machine-readable syntax to express specifications for data

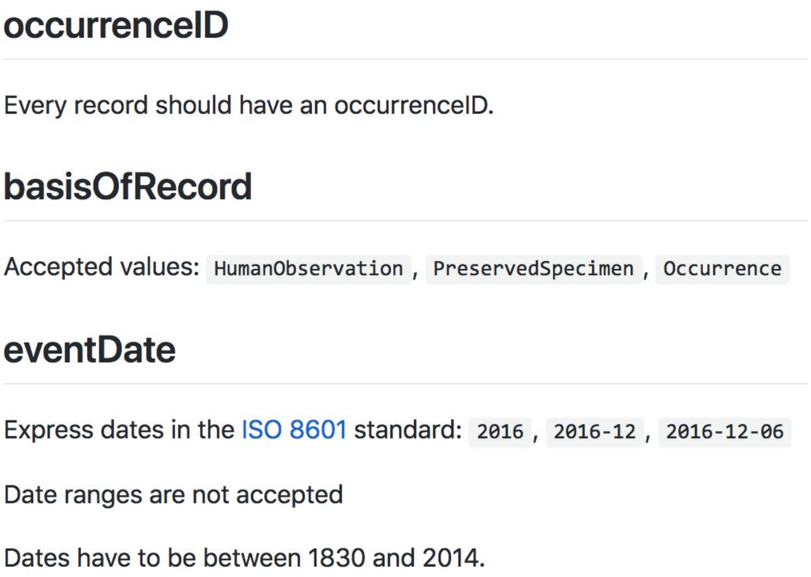
specification data-validation

whip

Whip is a human and machine-readable syntax to express **specifications for data.** It can be used as a whip to test how well data meets certain specifications, be it a feather 😅 or a chain whip 😱.

Example:

Whip syntax



occurrenceID

Every record should have an occurrenceID.

basisOfRecord

Accepted values: `HumanObservation`, `PreservedSpecimen`, `Occurrence`

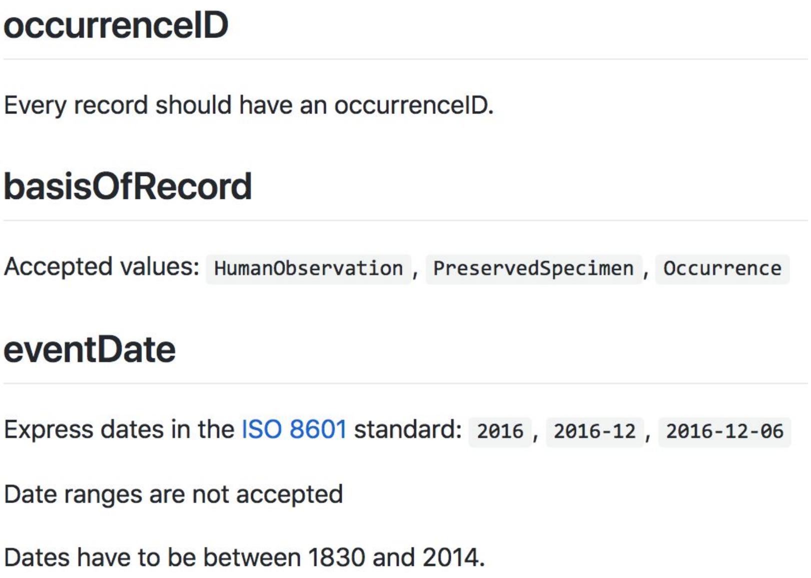
eventDate

Express dates in the ISO 8601 standard: `2016`, `2016-12`, `2016-12-06`

Date ranges are not accepted

Dates have to be between 1830 and 2014.

Whip syntax



occurrenceID

Every record should have an occurrenceID.

basisOfRecord

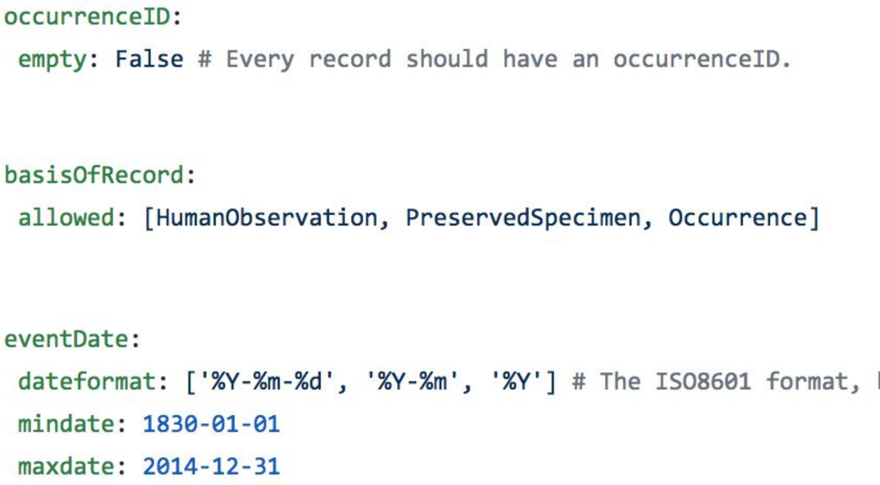
Accepted values: `HumanObservation` , `PreservedSpecimen` , `Occurrence`

eventDate

Express dates in the ISO 8601 standard: `2016` , `2016-12` , `2016-12-06`

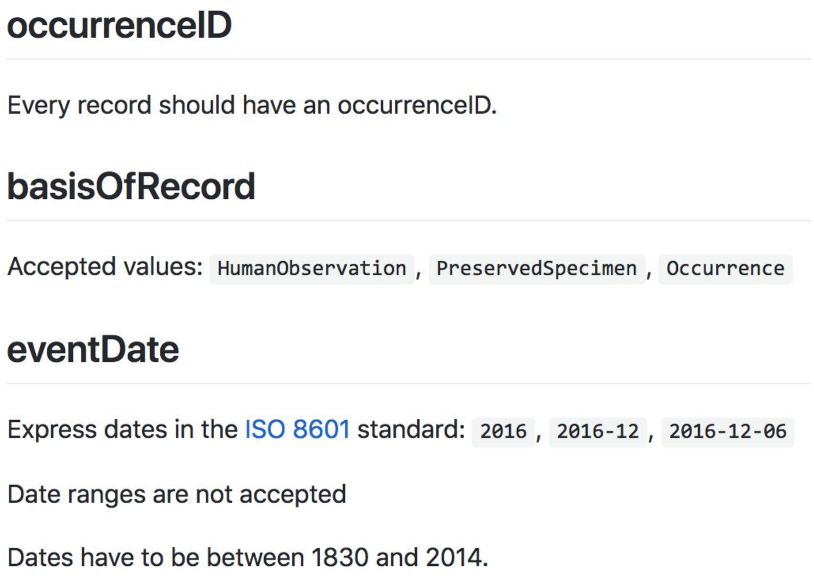
Date ranges are not accepted

Dates have to be between 1830 and 2014.



```
occurrenceID:
 empty: False # Every record should have an occurrenceID.


basisOfRecord:
 allowed: [HumanObservation, PreservedSpecimen, Occurrence]


eventDate:
 dateformat: ['%Y-%m-%d', '%Y-%m', '%Y'] # The ISO8601 format,
 mindate: 1830-01-01
 maxdate: 2014-12-31
```

Whip syntax



occurrenceID

Every record should have an occurrenceID.

basisOfRecord

Accepted values: `HumanObservation` , `PreservedSpecimen` , `Occurrence`

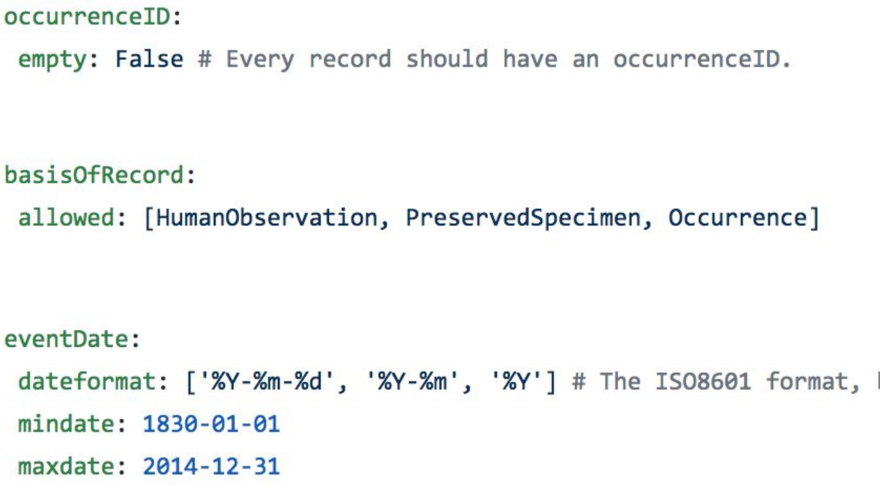
eventDate

Express dates in the ISO 8601 standard: `2016` , `2016-12` , `2016-12-06`

Date ranges are not accepted

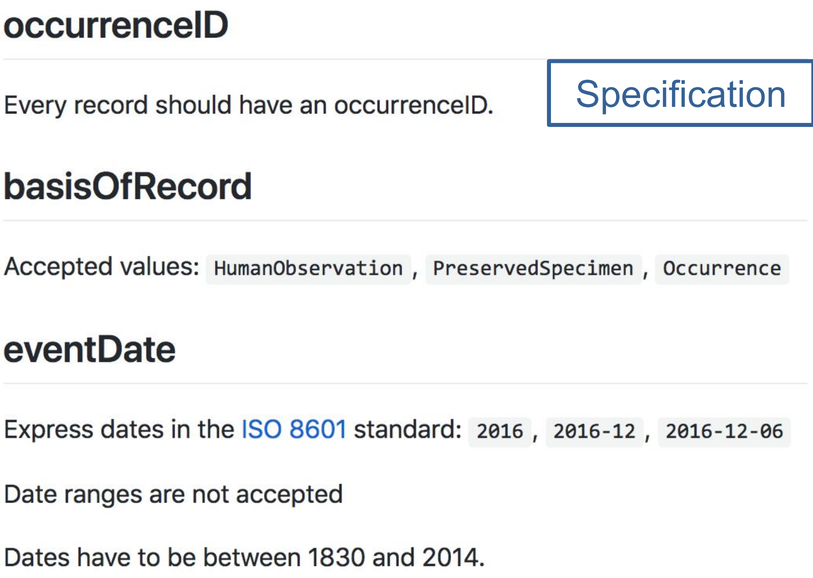
Dates have to be between 1830 and 2014.

Field



```
occurrenceID:
 empty: False # Every record should have an occurrenceID.


basisOfRecord:
 allowed: [HumanObservation, PreservedSpecimen, Occurrence]


eventDate:
 dateformat: ['%Y-%m-%d', '%Y-%m', '%Y'] # The ISO8601 format,
 mindate: 1830-01-01
 maxdate: 2014-12-31
```

Whip syntax



occurrenceID

Every record should have an occurrenceID.

basisOfRecord

Accepted values: `HumanObservation`, `PreservedSpecimen`, `Occurrence`

eventDate

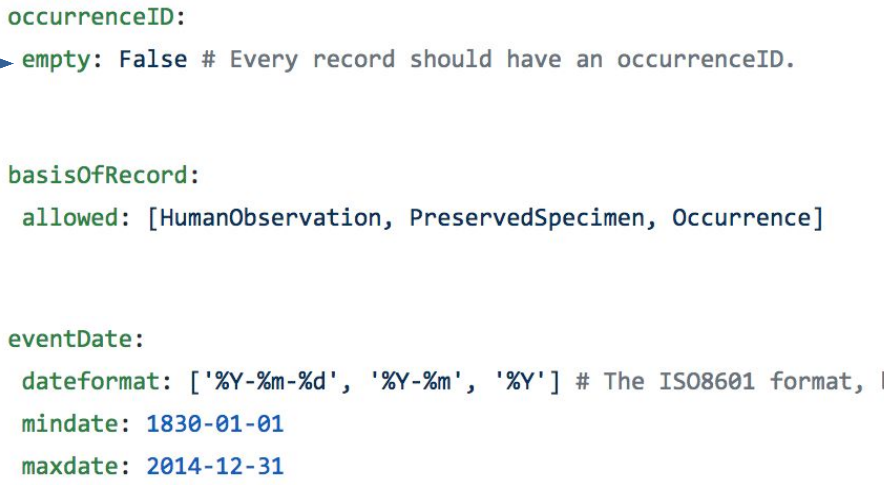
Express dates in the ISO 8601 standard: `2016`, `2016-12`, `2016-12-06`

Date ranges are not accepted

Dates have to be between 1830 and 2014.

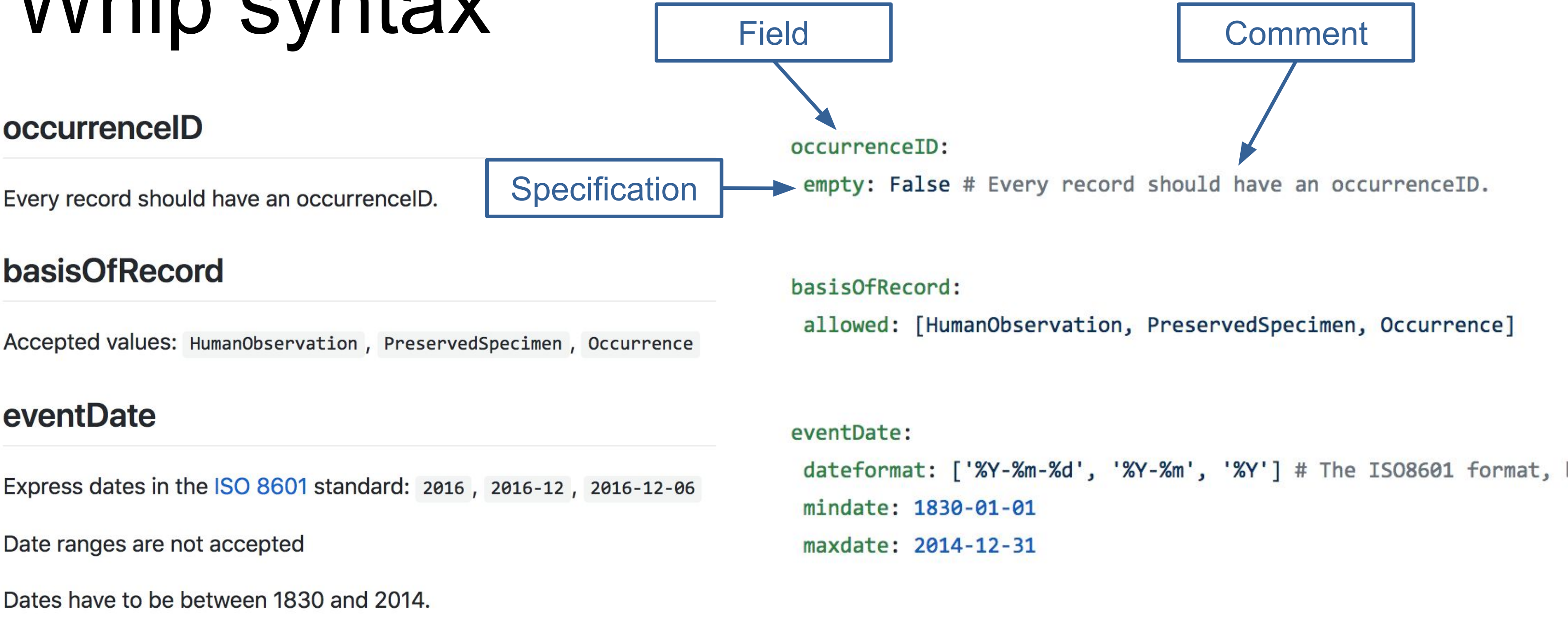
Field

Specification



```
occurrenceID:
  empty: False # Every record should have an occurrenceID.


basisOfRecord:
  allowed: [HumanObservation, PreservedSpecimen, Occurrence]


eventDate:
  dateformat: ['%Y-%m-%d', '%Y-%m', '%Y'] # The ISO8601 format,
  mindate: 1830-01-01
  maxdate: 2014-12-31
```

Whip syntax



occurrenceID

Every record should have an occurrenceID.

basisOfRecord

Accepted values: `HumanObservation`, `PreservedSpecimen`, `Occurrence`

eventDate

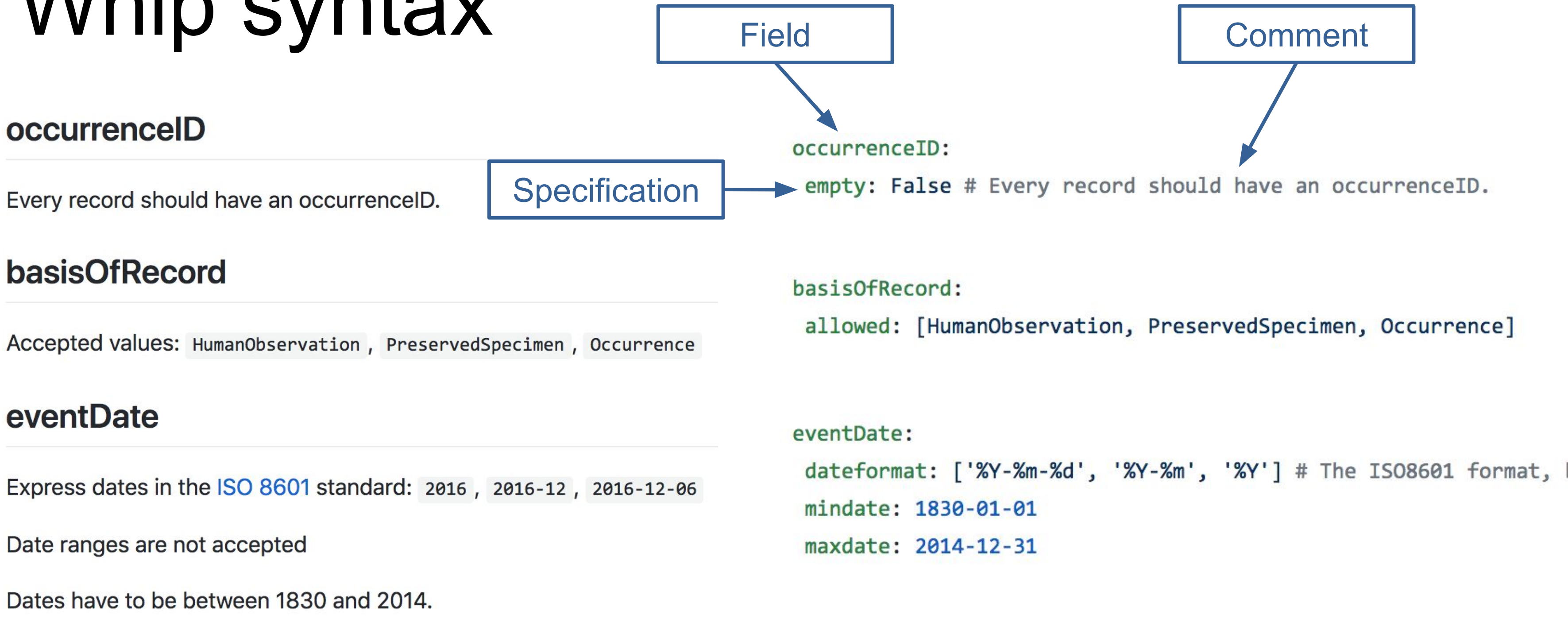
Express dates in the ISO 8601 standard: `2016`, `2016-12`, `2016-12-06`

Date ranges are not accepted

Dates have to be between 1830 and 2014.

```
occurrenceID:
 empty: False # Every record should have an occurrenceID.


basisOfRecord:
 allowed: [HumanObservation, PreservedSpecimen, Occurrence]


eventDate:
 dateformat: ['%Y-%m-%d', '%Y-%m', '%Y'] # The ISO8601 format,
 mindate: 1830-01-01
 maxdate: 2014-12-31
```

Whip syntax



occurrenceID

Every record should have an occurrenceID.

basisOfRecord

Accepted values: `HumanObservation`, `PreservedSpecimen`, `Occurrence`

eventDate

Express dates in the ISO 8601 standard: `2016`, `2016-12`, `2016-12-06`

Date ranges are not accepted

Dates have to be between 1830 and 2014.

```
occurrenceID:
 empty: False # Every record should have an occurrenceID.


basisOfRecord:
 allowed: [HumanObservation, PreservedSpecimen, Occurrence]


eventDate:
 dateformat: ['%Y-%m-%d', '%Y-%m', '%Y'] # The ISO8601 format,
 mindate: 1830-01-01
 maxdate: 2014-12-31
```

Whip specifications

allowed

minlength / maxlength

stringformat

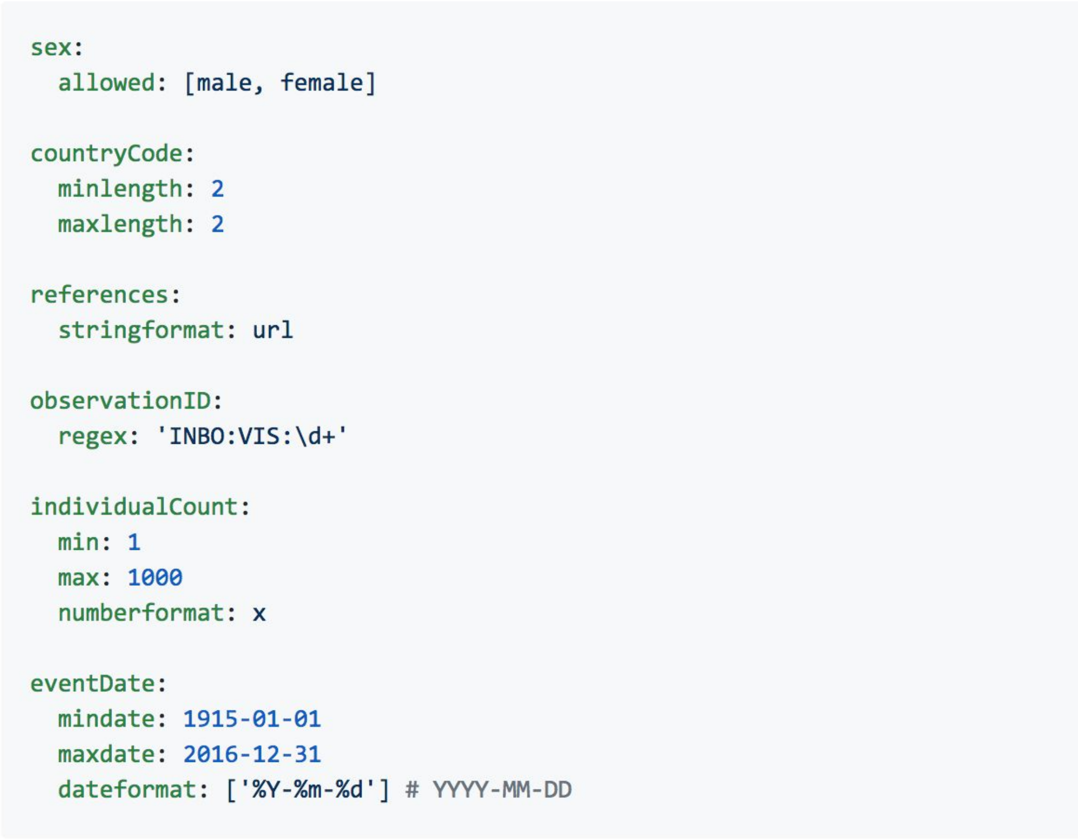
regex

min / max

numberformat

mindate / maxdate

dateformat



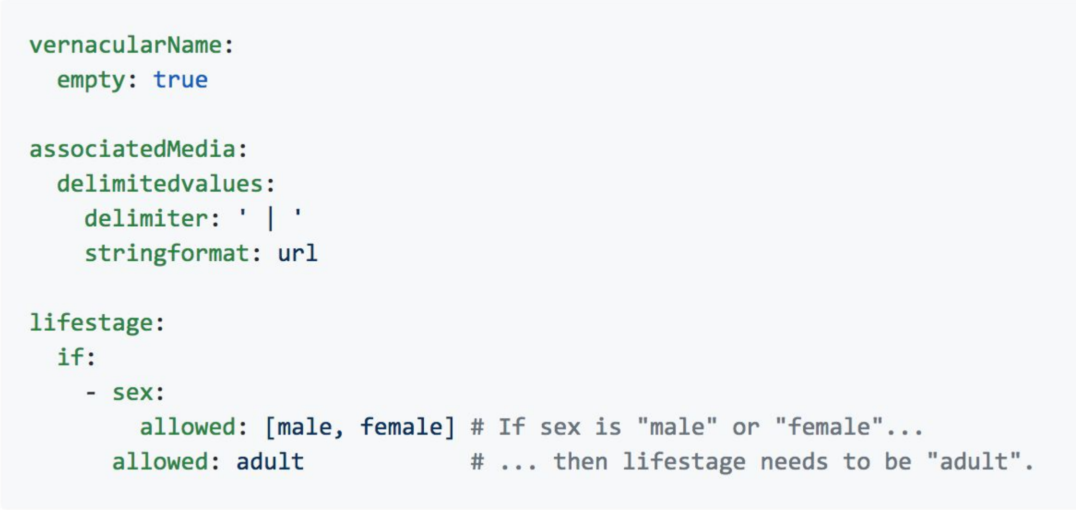
```
sex:
  allowed: [male, female]

countryCode:
  minlength: 2
  maxlength: 2

references:
  stringformat: url

observationID:
  regex: 'INBO:VIS:\d+'

individualCount:
  min: 1
  max: 1000
  numberformat: x

eventDate:
  mindate: 1915-01-01
  maxdate: 2016-12-31
  dateformat: ['%Y-%m-%d'] # YYYY-MM-DD
```

Whip scope specifications

empty

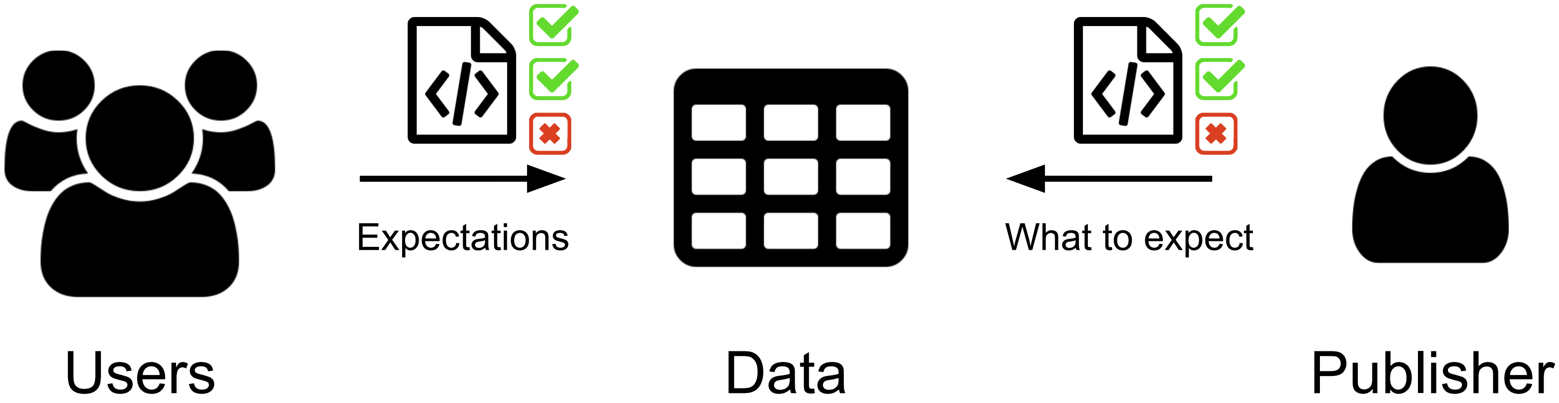
delimitedvalues

if



```
vernacularName:
  empty: true

associatedMedia:
  delimitedvalues:
    delimiter: ' | '
    stringformat: url

lifestage:
  if:
    - sex:
        allowed: [male, female] # If sex is "male" or "female"...
      allowed: adult            # ... then lifestage needs to be "adult".
```

How to test expectations?



Users

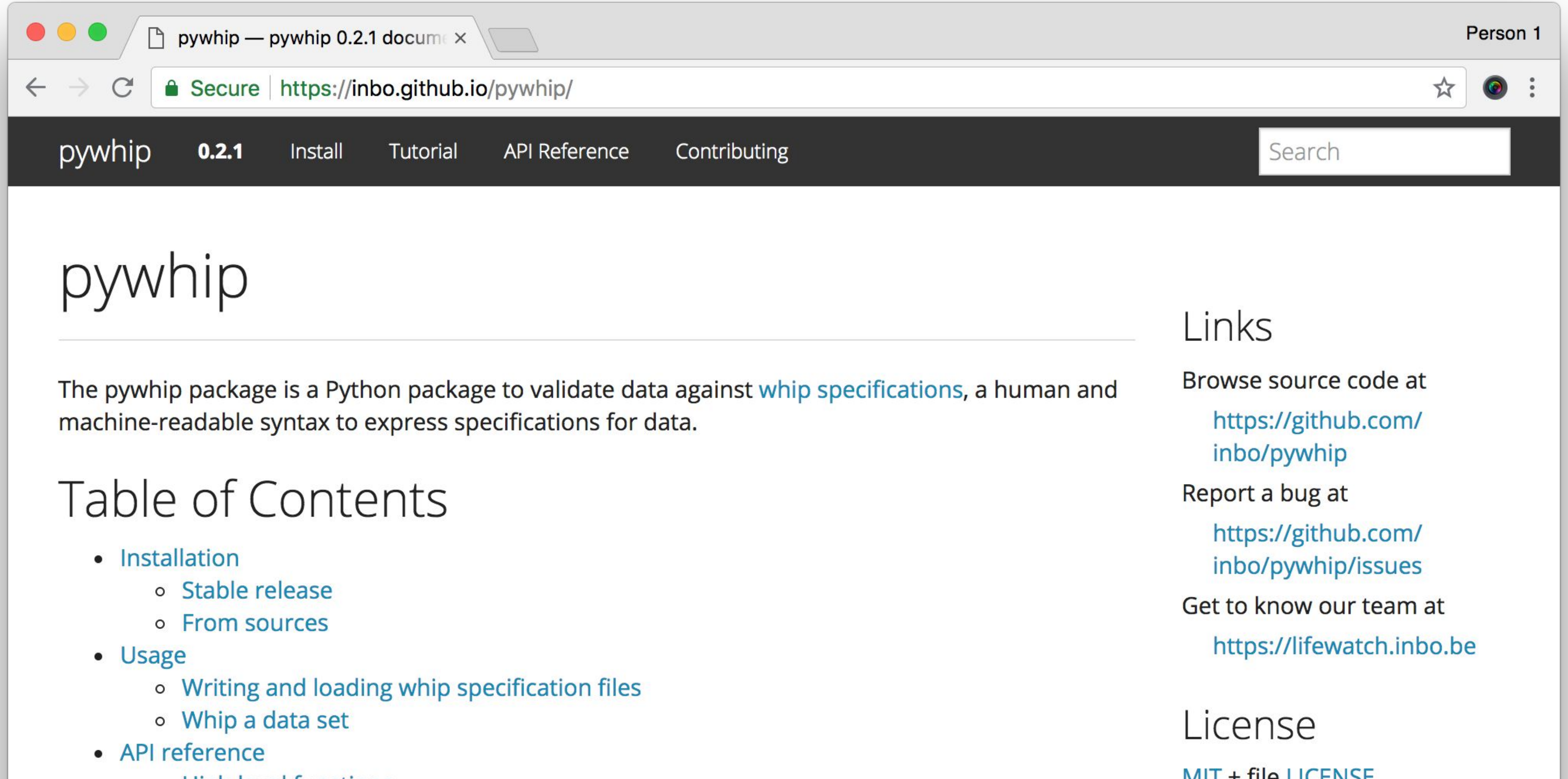
Expectations

Data

What to expect

Publisher

pywhip: a whip implementation



pywhip

The pywhip package is a Python package to validate data against whip specifications, a human and machine-readable syntax to express specifications for data.

Table of Contents

- Installation
  - Stable release
  - From sources
- Usage
  - Writing and loading whip specification files
  - Whip a data set
- API reference

Links

Browse source code at

https://github.com/inbo/pywhip

Report a bug at

https://github.com/inbo/pywhip/issues

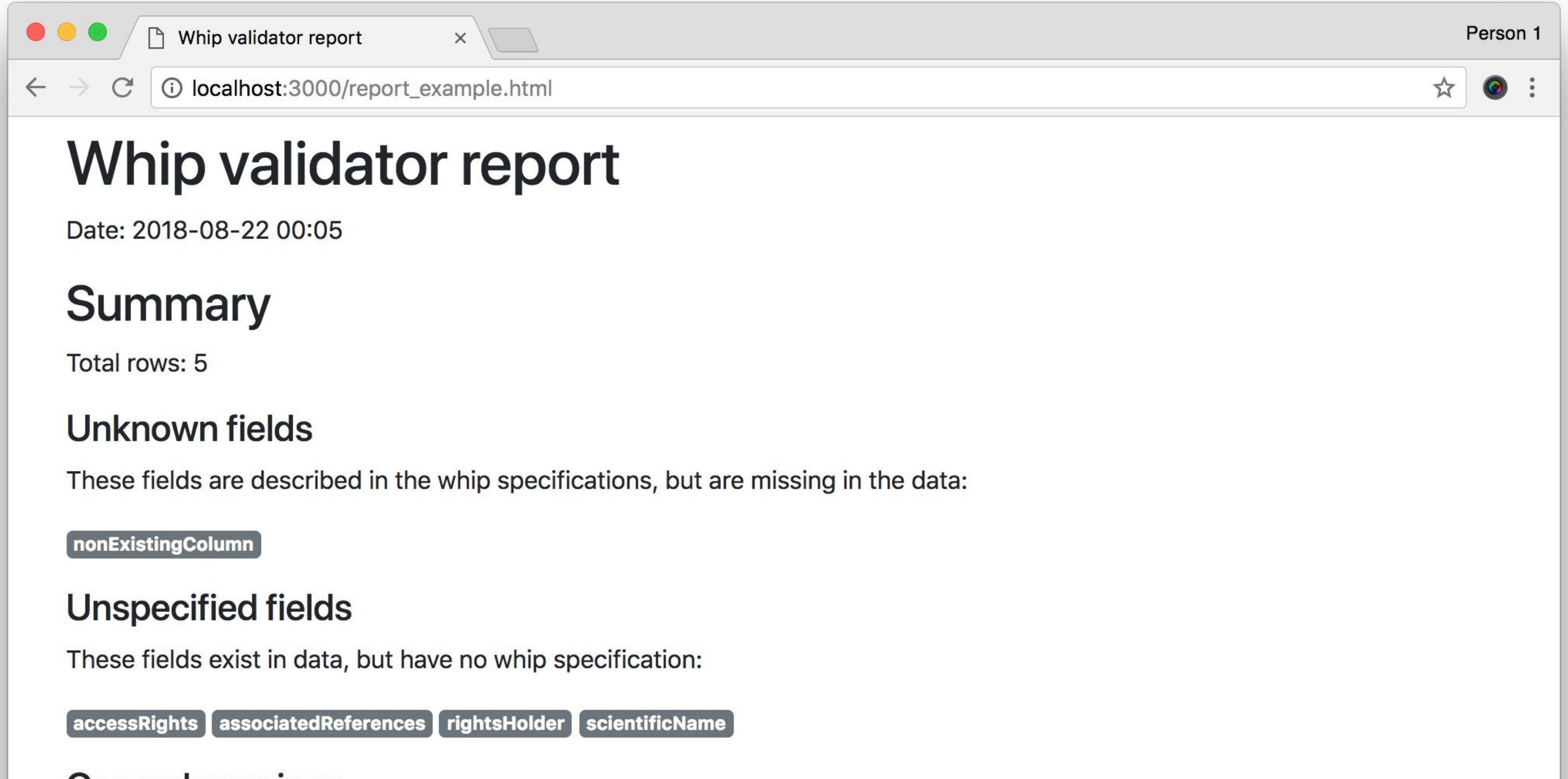
Get to know our team at

https://lifewatch.inbo.be

License

MIT + file LICENSE

pywhip



Whip validator report

Date: 2018-08-22 00:05

Summary

Total rows: 5

Unknown fields

These fields are described in the whip specifications, but are missing in the data:

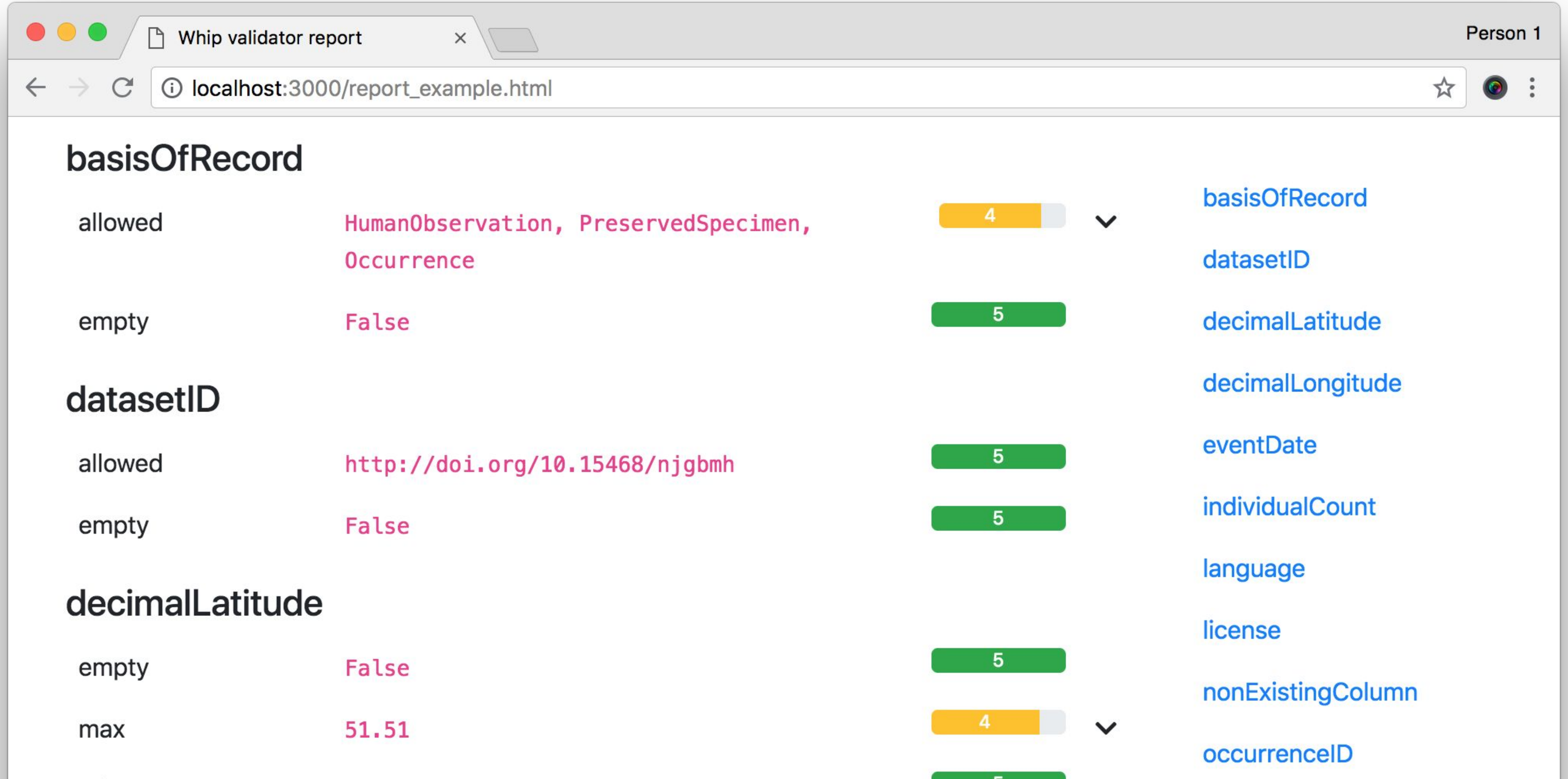
nonExistingColumn

Unspecified fields

These fields exist in data, but have no whip specification:

accessRights **associatedReferences** **rightsHolder** **scientificName**

pywhip



basisOfRecord

allowed	HumanObservation, PreservedSpecimen, Occurrence	4
empty	False	5

datasetID

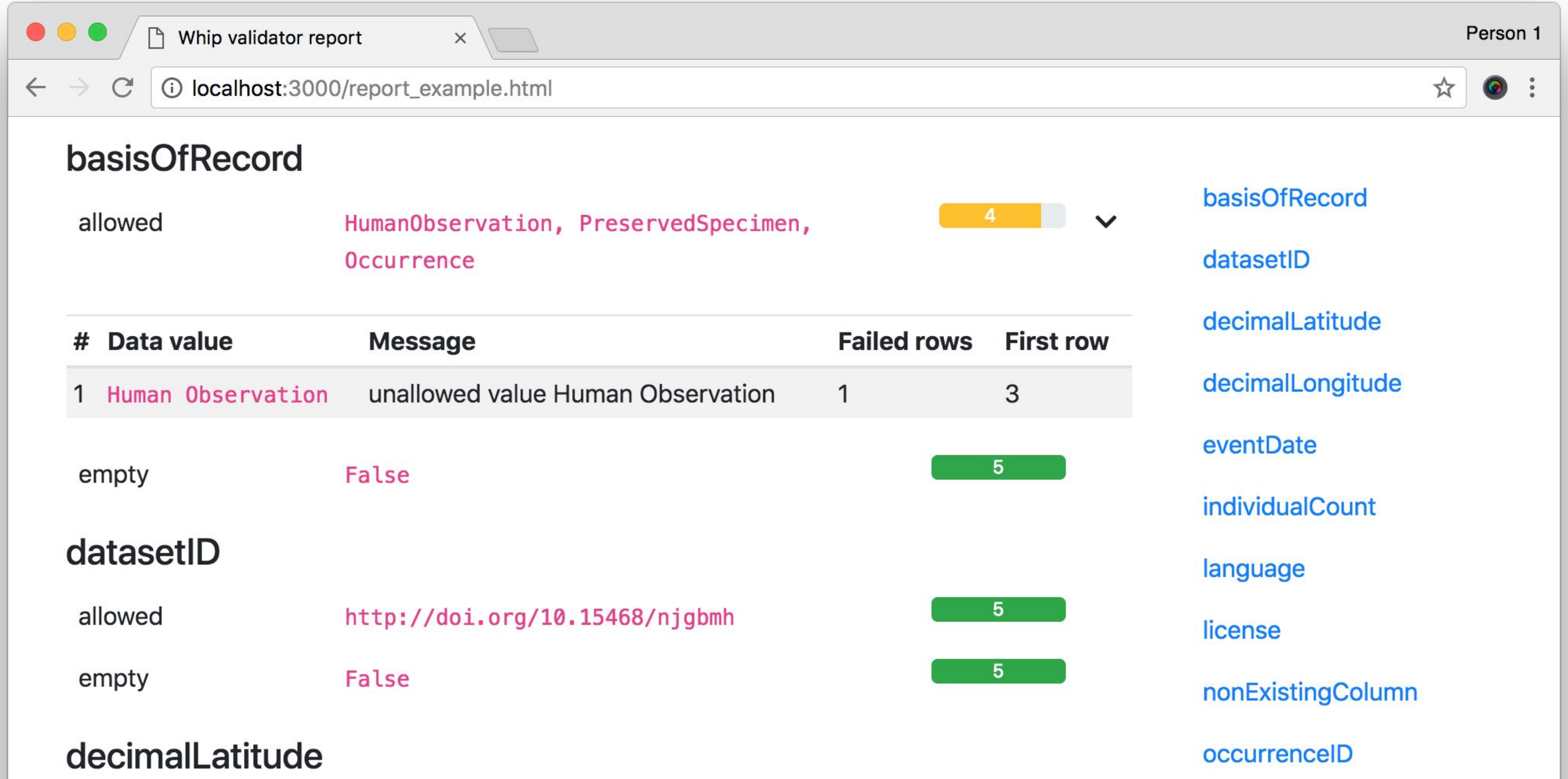
allowed	http://doi.org/10.15468/njgbmh	5
empty	False	5

decimalLatitude

empty	False	5
max	51.51	4

- basisOfRecord
- datasetID
- decimalLatitude
- decimalLongitude
- eventDate
- individualCount
- language
- license
- nonExistingColumn
- occurrenceID

pywhip



basisOfRecord

allowed `HumanObservation, PreservedSpecimen, Occurrence` 4

#	Data value	Message	Failed rows	First row
1	Human Observation	unallowed value Human Observation	1	3

empty `False` 5

datasetID

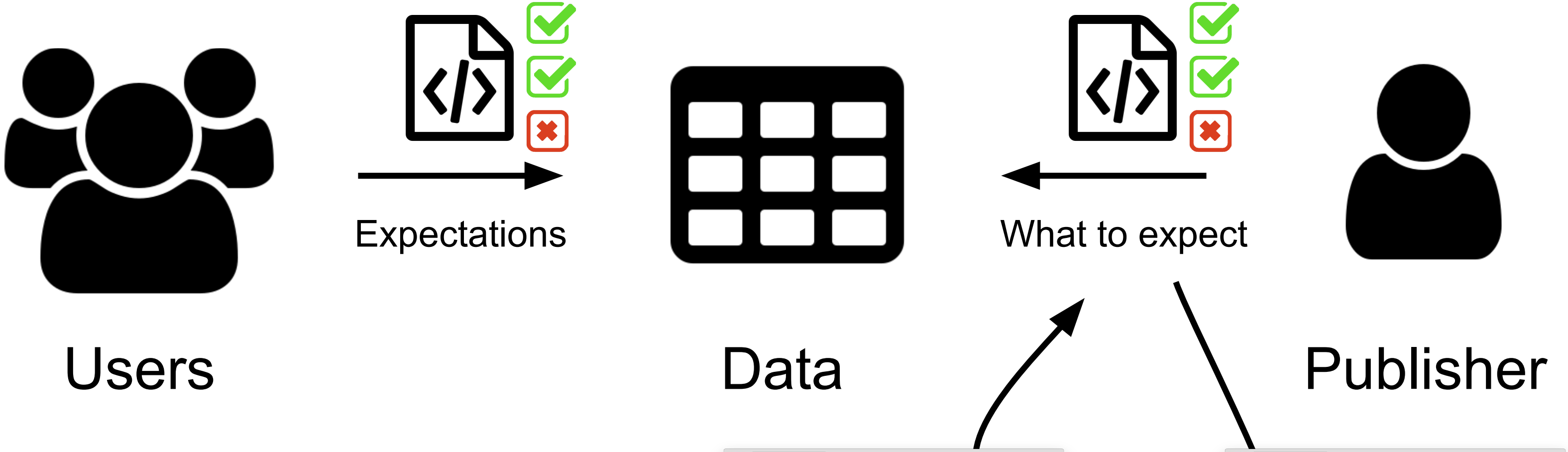
allowed `http://doi.org/10.15468/njgbmh` 5

empty `False` 5

decimalLatitude

- basisOfRecord
- datasetID
- decimalLatitude
- decimalLongitude
- eventDate
- individualCount
- language
- license
- nonExistingColumn
- occurrenceID

whip as unit tests to improve data quality



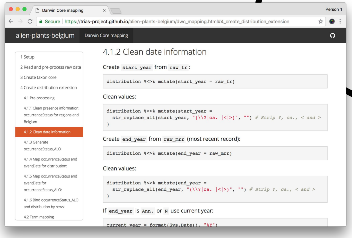
Expectations

What to expect

Users

Data

Publisher


Darwin Core mapping
Person 1
Secure
https://trias-project.github.io/alien-plants-belgium/dwc_mapping.html#4_create_distribution_extension
alien-plants-belgium
Darwin Core mapping

1 Setup
2 Read and pre-process raw data
3 Create taxon core
4 Create distribution extension
4.1 Pre-processing
4.1.1 Clean presence information: occurrenceStatus for regions and Belgium
4.1.2 Clean date information
4.1.3 Generate occurrenceStatus_ALO
4.1.4 Map occurrenceStatus and eventDate for distribution:
4.1.5 Map occurrenceStatus and eventDate for occurrenceStatus_ALO:
4.1.6 Bind occurrenceStatus_ALO and distribution by rows:
4.2 Term mapping

4.1.2 Clean date information

Create `start_year` from `raw_fr`:

```
distribution %<>% mutate(start_year = raw_fr)
```

Clean values:

```
distribution %<>% mutate(start_year =
  str_replace_all(start_year, "(\\?|ca. |<|>)", "") # Strip ?, ca., < and >
)
```

Create `end_year` from `raw_mrr` (most recent record):

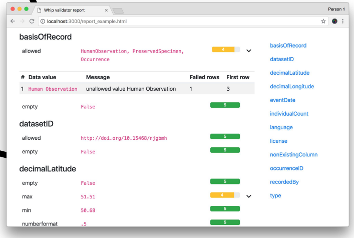
```
distribution %<>% mutate(end_year = raw_mrr)
```

Clean values:

```
distribution %<>% mutate(end_year =
  str_replace_all(end_year, "(\\?|ca. |<|>)", "") # Strip ?, ca., < and >
)
```

If `end_year` is `Ann.` or `N` use current year:

```
current_year = format(Sys.Date(), "%Y")
```



basisOfRecord

allowed HumanObservation, PreservedSpecimen, Occurrence 4

#	Data value	Message	Failed rows	First row
1	Human Observation	unallowed value Human Observation	1	3

empty False 5

datasetID

allowed http://doi.org/10.15468/njgbmh 5

empty False 5

decimalLatitude

empty False 5

max 51.51 4

min 50.68 5

numberformat .5 5

basisOfRecord
datasetID
decimalLatitude
decimalLongitude
eventDate
individualCount
language
license
nonExistingColumn
occurrenceID
recordedBy
type

Conclusion

Human and machine-readable **syntax to express specifications** for data

Not specific to biodiversity data (but we plan to use it for that)

Can be adopted by **users** (expectations) and **publishers** (what to expect)

Can be included with dataset as testable **metadata**

pywhip: first **implementation** for testing whip specifications

Thank you!

github.com/inbo/whip

github.com/inbo/pywhip

bit.ly/pywhip_binder

▦ ♡ 🗎

Data Specifications