‘Data Quality’

TG2: TOOLS AND SERVICES

Lee Belbin

Every time I give a presentation about the Atlas of Living Australia, ‘data quality’ is questioned.

Oddly, the questioners know about ‘fitness for use’, but have not examined it in the context of the Data Publishers

It took me one year to convince the Atlas of Living Australia to publish all data

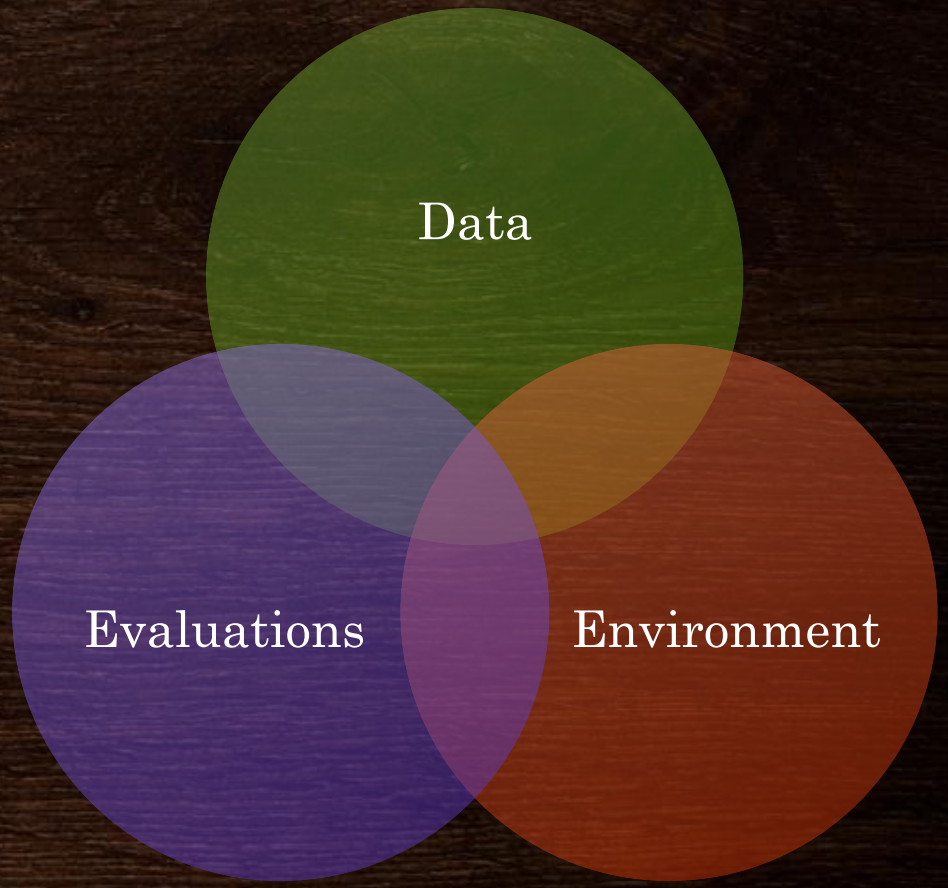
’Publish and be damned’

My case…

“How would you feel about the Atlas deciding what records were appropriate?”

Data Publishers must provide

1. All the data*
2. Record and dataset-level evaluations
3. An environment that makes it efficient to determine ‘fitness for use’



I have collected 152 available tests and assertions from Data Publishers, e.g....

Variable	Specification (Brief [User] Description)	Data resolution	OUTPUT TYPE	Darwin Core	INPUT: Darwin Core Fields (Elements)	Severity	Owner
	Number of assertions = TRUE. An indication of the record issues	Single Record	Measure	All	All	Warning	Lee Belbin
	Number of supplied + inferred Darwin Core fields. An indication of record completeness	Single Record	Measure	All	All	Warning	Lee Belbin
	Completeness checks – calculating ratio of null DwC values in a table?	Dataset	Measure	All	All	Warning	Tania Laity
MISSING_COLLECTION_DATE	Collection date field is missing or null	Single Record	Validation	Event	eventDate	Error	ALA
INCOMPLETE_COLLECTION_DATE	The supplied collection date is missing a day and/or month component. This is used to differentiate non error conditions for an event date.	Single Record	Validation	Event	eventDate	Warning	ALA
MODIFIED_DATE_INVALID	A (partial) invalid date is given for dc:modified, such as a non existing date, invalid zero month, etc.	Single Record	Validation	All	dcterms:modified		GBIF
INVALID_COLLECTION_DATE	The collecting event date was given as pre 1700, or is otherwise invalid. This is used as a general date issue	Single Record	Validation	Event	eventDate	Error	ALA, GBIF
MODIFIED_DATE_UNLIKELY	The date given for dc:modified is in the future or predates unix time (1970).	Single Record	Validation	All	dcterms:modified		GBIF
datecollected_bounds	Date Collected out of bounds (1700-01-02, Date of Indexing).	Single Record	Validation	Event	eventDate		iDigBio
RECORDED_DATE_UNLIKELY	The recording date is highly unlikely, falling either into the future or represents a very old date before 1600 that predates modern taxonomy.	Single Record	Validation	Event	eventDate		GBIF
RECORDED_DATE_MISMATCH	The recording date specified as the eventDate string and the individual year, month, day are contradicting.	Single Record	Validation	Event	eventDate, day, month, year		GBIF
DAY_MONTH_TRANSPOSED	Supplied day and month fields appear to be transposed. if month > 12 and day <12 we can infer the fields have been incorrectly mapped	Single Record	Validation and Improvement	Event	eventDate	Warning	ALA
FIRST_OF_MONTH	May indicate the date is only known or recorded to the Month. Flag if there is no precision data. datePrecision is not a curent DwC field	Single Record	Validation	Event	eventDate, datePrecision(nonDwC)	Warning	ALA
FIRST_OF_YEAR	May indicate the date is only known or recorded to the Year. Flag if there is no precision data. datePrecision is not a curent DwC field	Single Record	Validation	Event	eventDate, datePrecision(nonDwC)	Warning	ALA
FIRST_OF_CENTURY	May indicate the date is only known or recorded to the Century. Flag if there is no precision data. datePrecision is not a curent DwC field	Single Record	Validation	Event	eventDate, datePrecision(nonDwC)	Warning	ALA
DATE_PRECISION_MISMATCH	Date precision does not match the data. datePrecision is not a curent DwC field	Single Record	Validation	Event	eventDate, datePrecision(nonDwC)	Error	ALA

There are overlaps between the tests but rationalization is relatively easy

Recommendations

- A (TDWG) standard suite of tests be finalized
- (TDWG) Darwin Core fields used as a foundation
- Code for the tests/assertions to be openly available
- Any records viewed or downloaded report all test fails (assertions)
- All Data Publishers should be encouraged to adopt the standard

Why?

- Comprehensive assertions can help users determine fitness for use
- Consistency will build trust in the community
- Assertions are far more stable than tools and workflows that use them
- Tests/assertions are easy to implement and maintain
- Re-use is better than re-invention
- Data Publishers will be supporting best current practice
- Enhance collaboration between Data Publishers and TDWG

An Example

Assertion: ALTITUDE_OUT_OF_RANGE

Description: Altitude greater than 10000m, or less than -100m

Applies to: Single Record

Output type: Validation

Darwin Core Class: Location

Darwin Core Field: verbatimElevation

Severity: Warning

Data Publisher: GBIF

Additional

- I have updated the (40) Tools listed in the GBIF Resources (http://www.gbif.org/resources?sort_by=gr_date_of_publication&sort_order=DESC&items_per_page=All&f[0]=gr_resource_type%3A1010&f[1]=gr_purpose%3A938&searched=1)
  - Help would be appreciated
- I have collated 29 references that address ‘data quality’
  - Help would be appreciated

Thank You