Data Quality
Whose Responsibility Is It?

Arthur Chapman

Australian Biodiversity Information Services

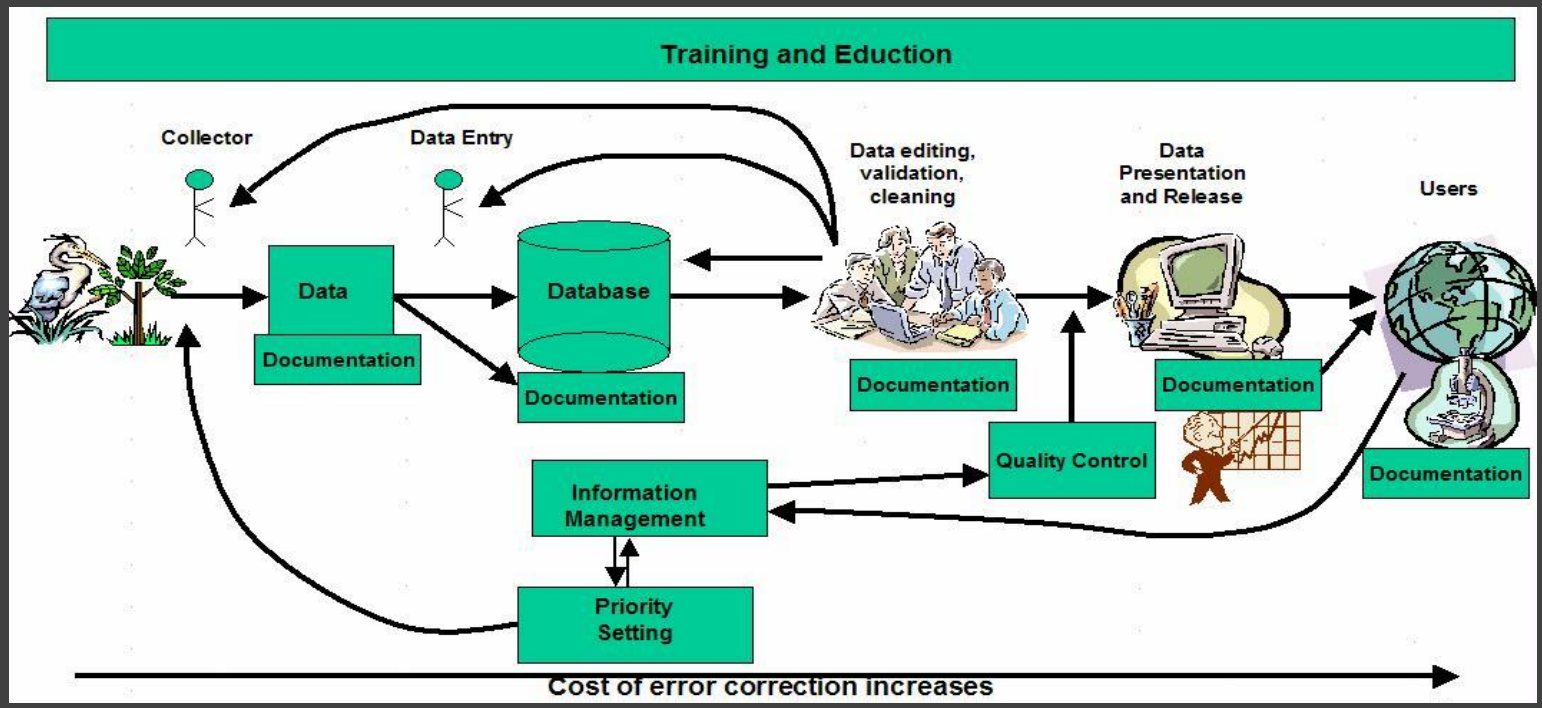
TDWG Data Quality Interest Group

Male Pied Kingfisher (Ceryle rudis)
Upper Zambezi River, Namibia

Simple Answer

TDWG DQIG

It is Everybodys!



From: Chapman 2005, Principles of Data Quality

Responsibilities

- Collectors of the specimens
- Database designers and builders
- Data entry operators
- Data curators and managers
- Those responsible for exporting/exchanging data
- Data aggregators
- Data publishers
- Data users
- Funding bodies

Why, why, why?

Why do we still have databases that allow:

- A latitude of 95°
- A month of 13
- A day of 32 (is it 31 or 23?)
- A year of 2020
- A year of 36 (is it 1936 or 1836?)
- Default of “0” in place of “Null” “”

You get the picture!
At least 25% of our tests are of this nature

So what do we mean by 'Data Quality'?

An essential or distinguishing characteristic necessary for [spatial] data to be fit for use.

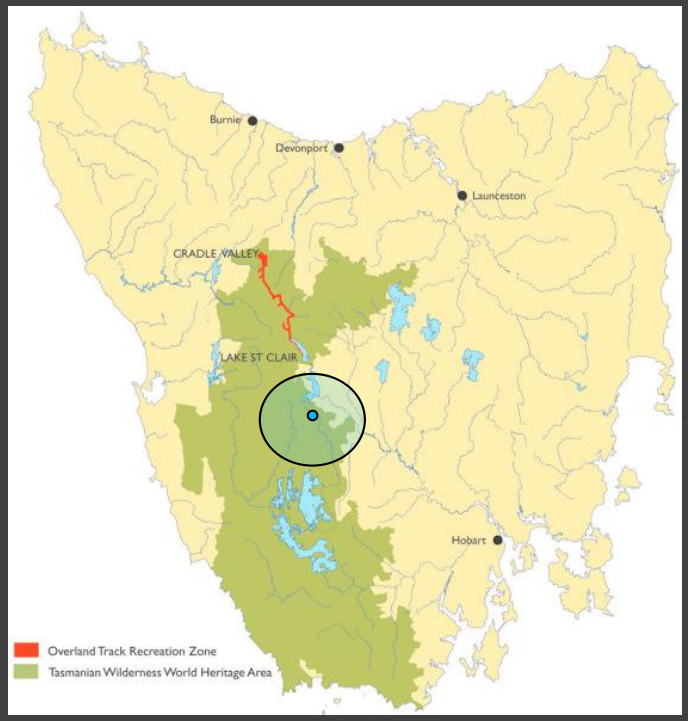
SDTS 02/92

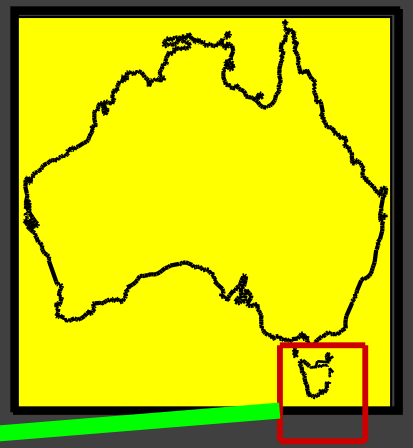
The general intent of describing the quality of a particular dataset or record is to describe the fitness of that dataset or record for a particular use that one may have in mind for the data. (Chrisman 1991)

Data quality - fitness for use?

Fitness for use

- Does species 'A' occur in Tasmania?
- Does species 'A' occur in National Park 'y





dwc:geocodeUncertainty=50,000

TDWG Data Quality Interest Group

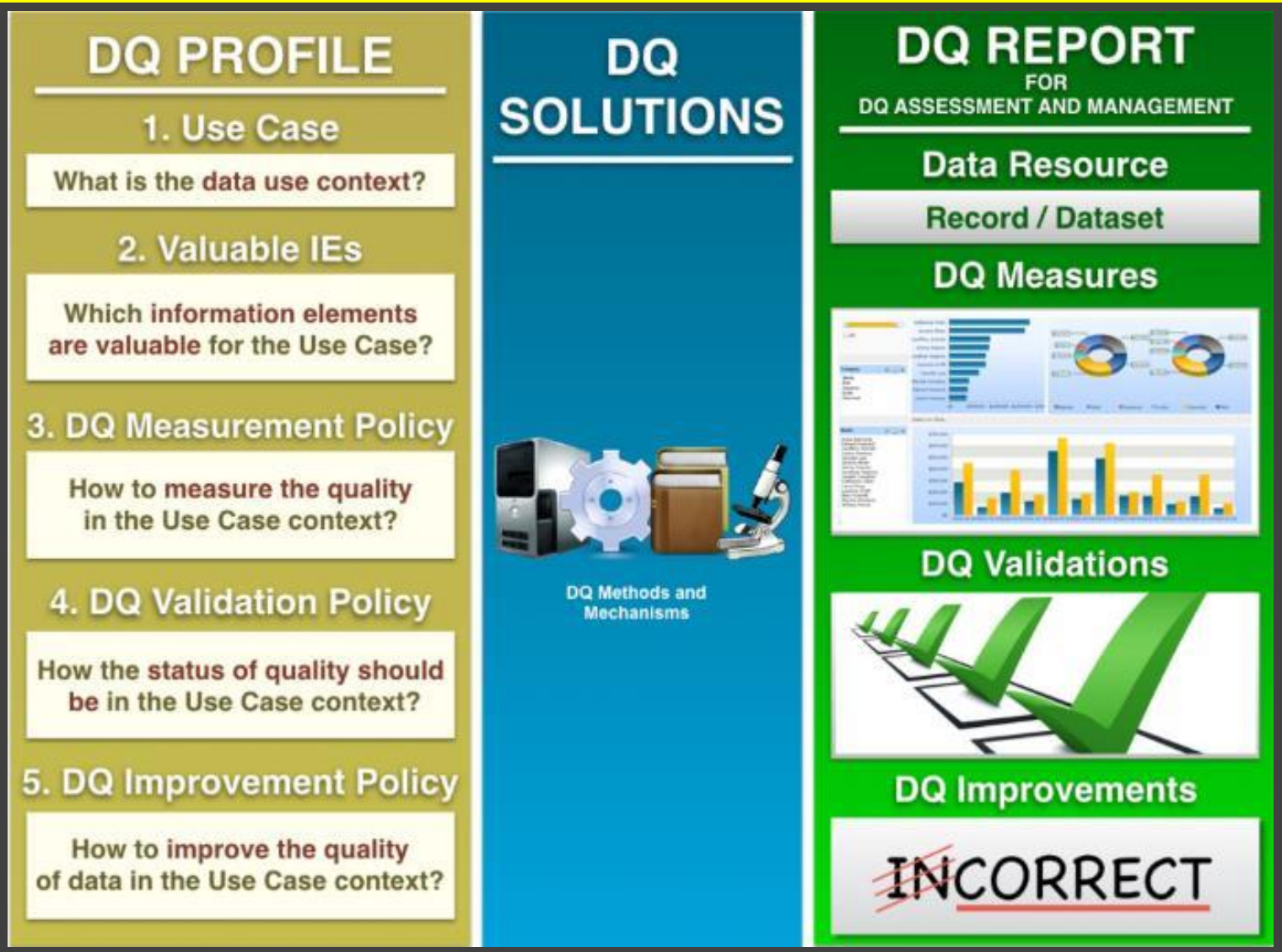
Established in 2014

TDWG DQIG

- Framework for Data Quality
- Consistent – Tests and Assertions
- Use case library for different users/uses
- Data Quality Profiles
- Vocabularies of value
- Documentation of Quality
- Develop an annotations standard for DQ assertions (Annotations IG/DQIG)

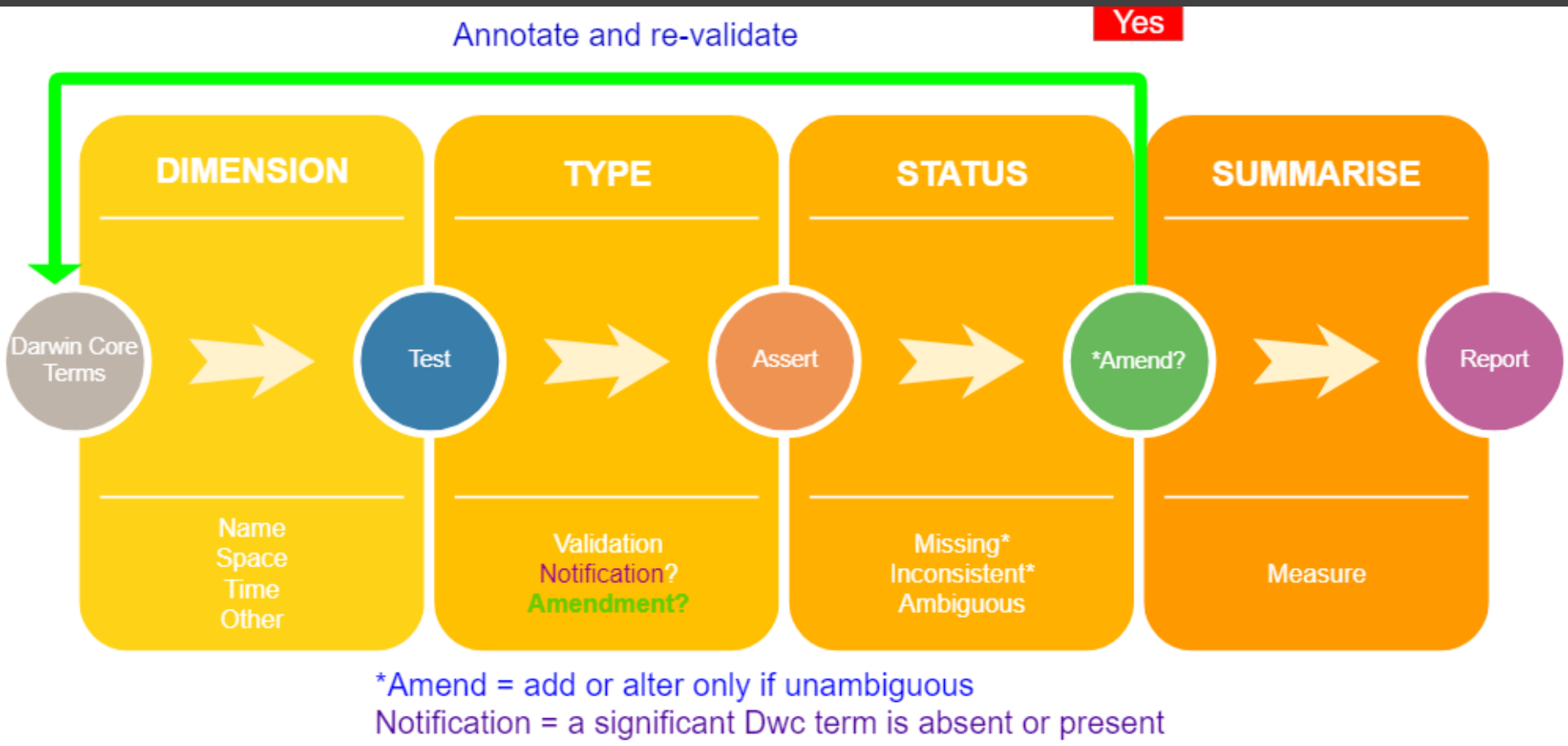
Often not as important to improve the data quality as to assess its quality and to document that quality

TG1 – Framework on Data Quality



From: Veiga, A.K. *et al*. (2017)

TG2 – Core Tests and Assertions



*Amend = add or alter only if unambiguous
Notification = a significant Dwc term is absent or present

Basic tests-assertion concepts

1. **NAME** is missing, ambiguous or inconsistent
2. **SPACE** is missing, ambiguous or inconsistent
3. **TIME** is missing, ambiguous or inconsistent
4. **OTHER** (e.g., basisOfRecord) is missing or inconsistent
5. If we have sufficient unambiguous information, we may be able to AMEND one or more terms

Core Tests

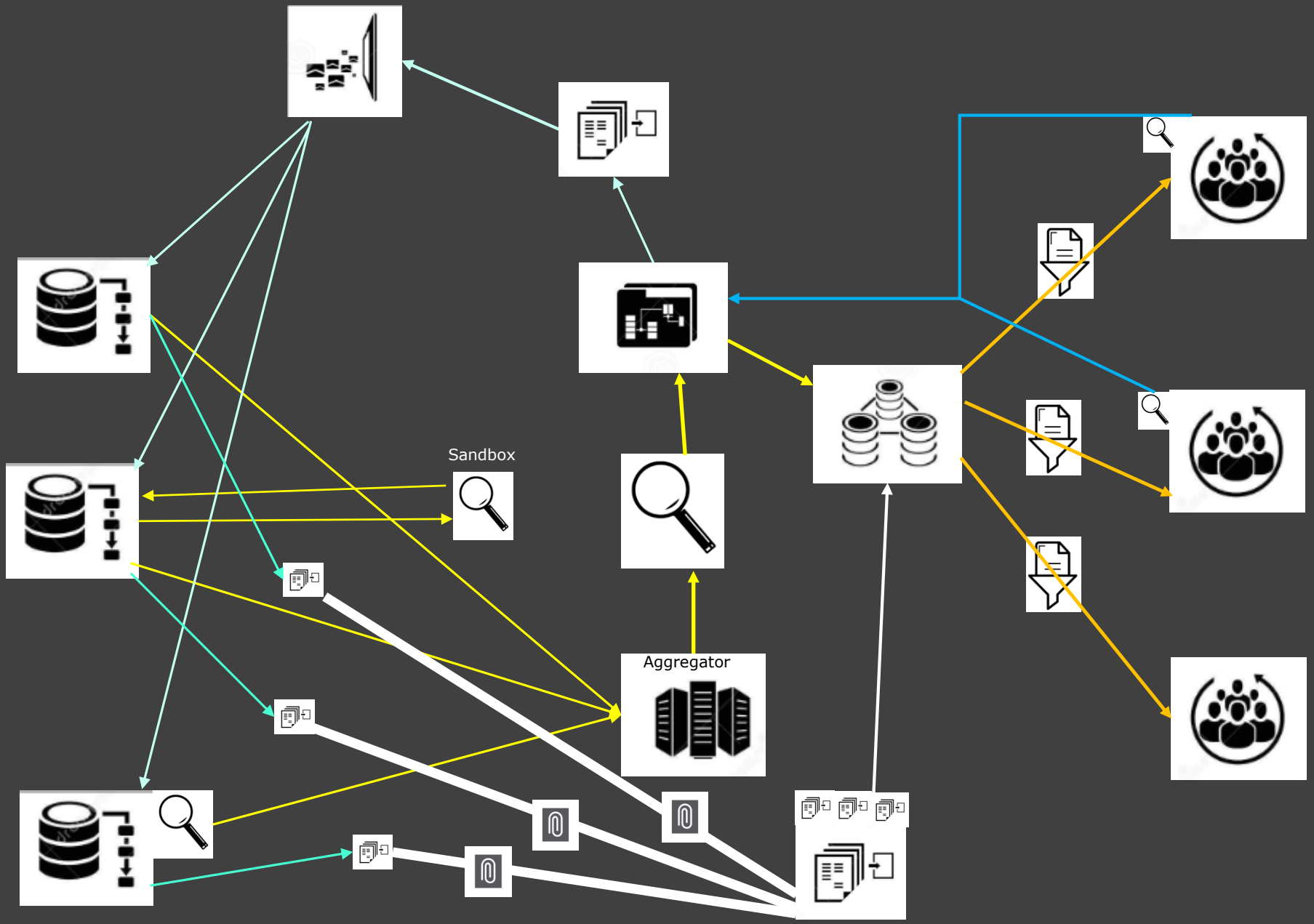
Field	Value
GUID	Globally Unique Identifier
Label	Name of the test
Term-Actions	The Term and Action part of the Label
Output Type	Validation, Notification, Amendment or Measure
Darwin Core Class	The Darwin Core Class that the test references
Information Elements	The Darwin Core Terms referenced by the test
Description	Description of the test of Output Type “Amendment”, “Measure” or “Notification”
Fail Description	Description of the test of Output Type “Validation” if the test fails (NOT_COMPLIANT)
Pass Description	Description of the test of Output Type “Validation” if the test passes (COMPLIANT)
Dimension	Name, Space, Time or Other
Data Quality Dimension	Completeness, Conformance, Consistency, Likeliness, Reliability, Resolution (TG1 Framework)
Warning Type	Nature of the issue (Ambiguous, Amended, Incomplete, Inconsistent, Invalid, Notification, Report, Unlikely)
Example	At least one simple example
Source	The source of the test (agency, individual, etc.)
References	References related to the test
Example Implementations (Mechanisms)	Places/organisations, etc. that have implemented this test as written
Link to Specification Source Code	A link to generic or specific source code for the test
Notes	Notes that pertain to the test – may be issues, clarifications, etc.
Test Prerequisites	Prerequisites that should be considered prior to running the test.

TG2-AMENDMENT_EVENTDATE_FROM_VERBATIM #86

Field	Value
GUID	6d0a0c10-5e4a-4759-b448-88932f399812
Label	AMENDMENT_EVENTDATE_FROM_VERBATIM
Term-Actions	EVENTDATE_FROM_VERBATIM
Output Type	Amendment
Resource Type	SingleRecord
Darwin Core Class	Event
Information Elements	dwc:eventDate
Description	The value of dwc:eventDate was unambiguously interpreted from dwc:verbatimEventDate
Dimension	Time
Data Quality Dimension	Completeness
Warning Type	Amended
Example	dwc:verbatimEventDate="March 2 2013" amends to dwc:eventDate="2013-03-02"
Source	VertNet, FP, Kurator
References	
Example Implementations (Mechanisms)	Kurator:event_date_qc
Link to Specification Source Code	https://github.com/FilteredPush/event_date_qc/blob/5f2e7b30f8a8076977b2a609e0318068db80599a/src/main/java/org/filteredpush/qc/date/DwCEventDQ.java#L169 A minimum set of unit tests is at: https://github.com/FilteredPush/event_date_qc/blob/5f2e7b30f8a8076977b2a609e0318068db80599a/src/test/java/org/filteredpush/qc/date/DwcEventDQTest.java#L310see also unit tests for underlying implementation at https://github.com/FilteredPush/event_date_qc/blob/5f2e7b30f8a8076977b2a609e0318068db80599a/src/test/java/org/filteredpush/qc/date/DateUtilsTest.java#L460and https://github.com/FilteredPush/event_date_qc/blob/5f2e7b30f8a8076977b2a609e0318068db80599a/src/test/java/org/filteredpush/qc/date/DateUtilsTest.java#L616
Notes	
Test Prerequisites	The field dwc:eventDate is EMPTY and the field dwc:verbatimEventDate is not EMPTY and is unambiguously interpretable as an ISO 8601:2004(E) date

Tests Workflow

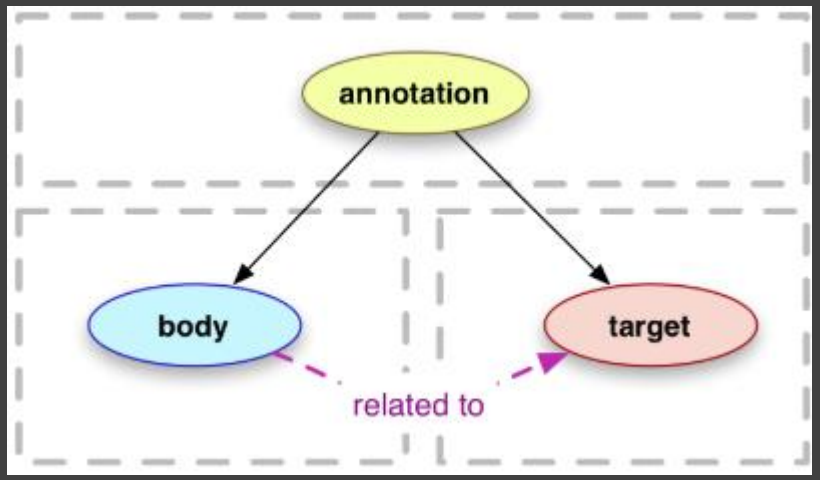
The DQ Process

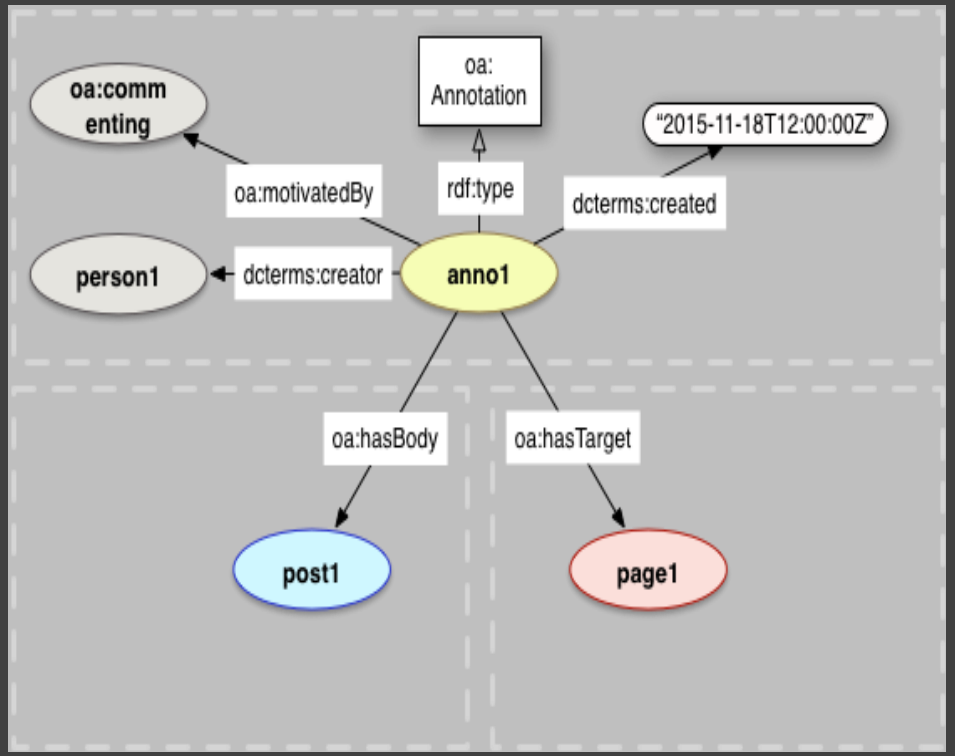


Annotations

Important

1. To have ability to chain annotations
2. That annotations are permanently retained



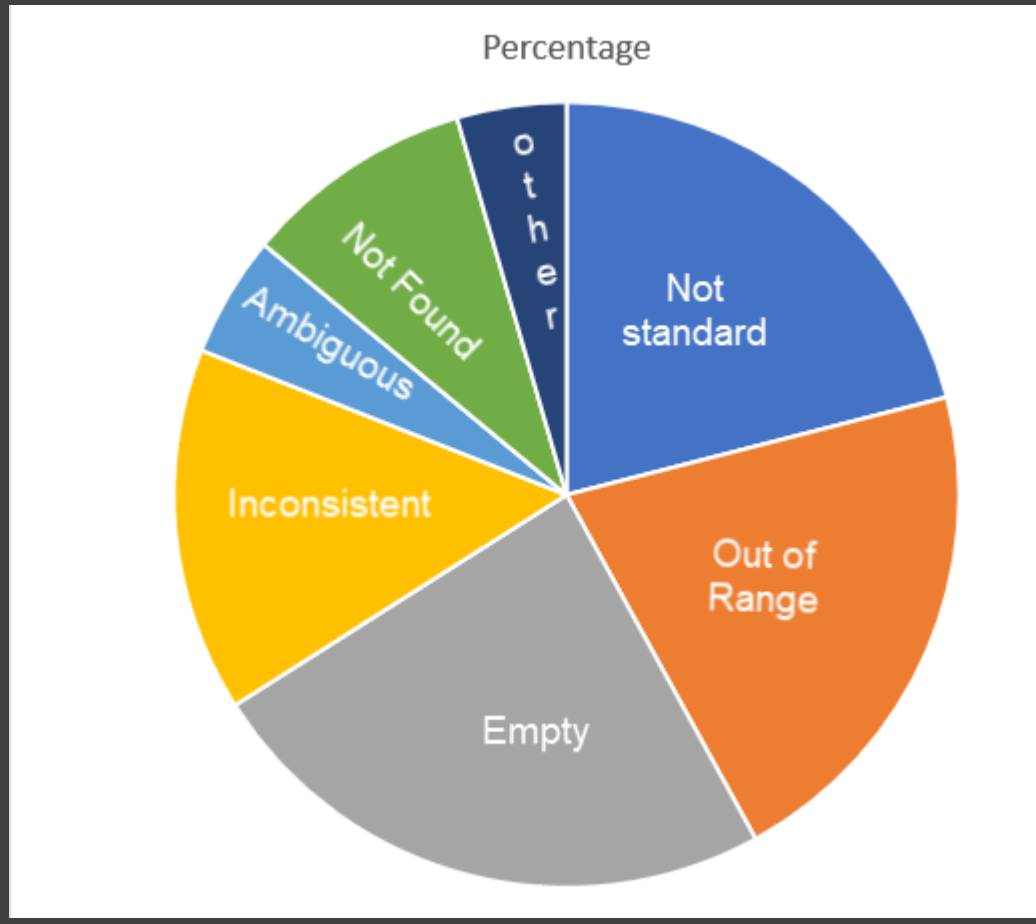


W3C ao:Annotation in conjunction with W3C PROV a key solution

Causes?

Type of Validation

- Not Standard (21%)
  - Vocabularies
  - Pick lists
- Out of Range (21%)
  - Database constraints
- Empty (24%)
  - Where you would expect something
- Ambiguous (5%)
  - Typo?
  - Homonyms
- Not Found (9%)
  - Similar to Not Standard?
- Inconsistent (15%)
  - Between two fields



Where to from here?

Development of generic code

1. Core aggregators (GBIF, ALA, iDigBio)
2. GBIF Nodes based on ALA architecture
3. Other aggregators (OBIS, SiBBr, etc.)
4. Data Custodians (Museums etc.)
   - Standard Annotations
   - Individual code based on generic code
   - Test database to check implementations
5. Database and DBMS developers
6. Sandbox applications

Needs

Collaboration at all levels

- Feedback from aggregators, users
- Data Quality policies within institutions
  - A vision with respect to having good quality data;
  - A policy to implement that vision; and
  - A strategy for implementation
    - Truth in labelling
    - Fitness for purpose labelling
- Resources (funding, staffing)
- Database companies and developers
  - DBMS users groups
- Standards and consistency
- Adherance to standards/vocabularies
- Documentation - and Documentation
  - Metadata of quality
  - Annotations
  - Users need to know the quality

Vocabularies

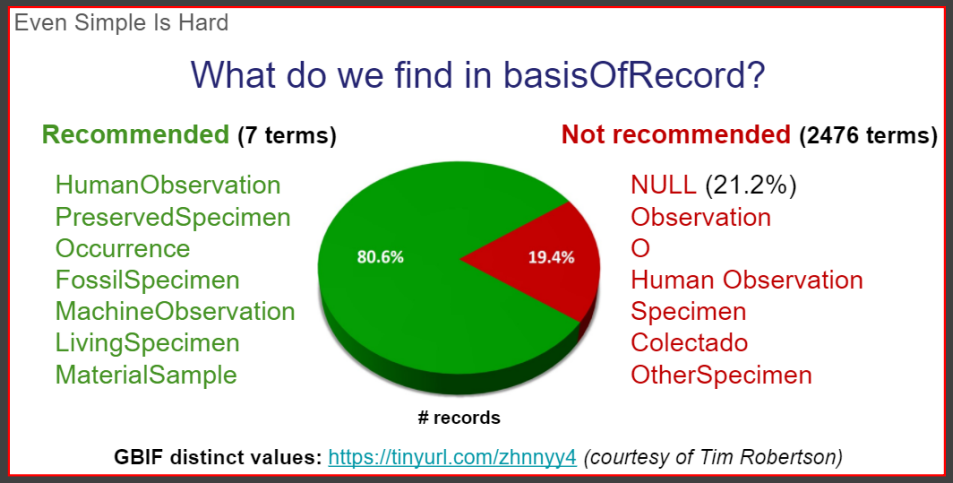
From Darwin Core Webinars

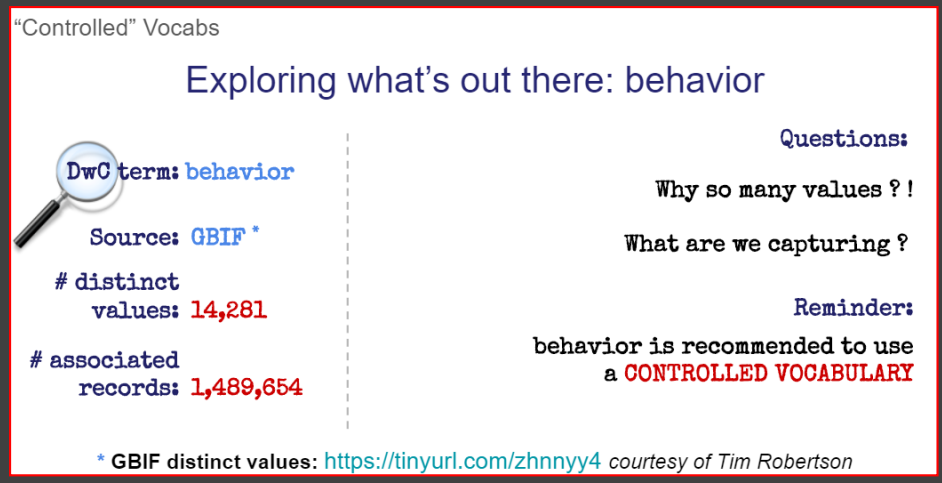
John Wieczorek

"Even Simple is Hard" (Chapter 2)

Paula Zermoglio

"Controlled Vocabularies" (Chapter 3)





Vocabularies

Required for

- Core Tests (29 rely on a Vocabulary)
- Use Cases
- Profiles
- Darwin Core
- Databasing and DBMS developments
- Disciplines (invasive species, etc.)

TG4: Paula Zermoglio

Conclusion

Having poor data is worse than having no data at all.

Maintenance of the data and databases is as important as maintenance of the specimens and the infrastructure, and should be funded accordingly.

Experience has shown that treating data as a long-term asset and managing it within a coordinated framework produces considerable savings and ongoing value

Data can’t always be improved but it can be documented

Documentation is fundamental

Thank You

Ngā mihi nui ki a koe

Ctenophorus caudicinctus macropus (Ring-tailed Dragon)

Lawn Hill Gorge, Queensland, Australia