TDWG/GBIF
Data Quality Interest Group

Arthur D. Chapman, Lee Belbin, Miles Nicholls,
Antonia Saraiva, Allan Koch Veiga, Dmitry Schigel

biodiv_2@achapman.org

TDWG/GBIF Data Quality Interest Group

- Proposed at TDWG2013
- Merged with GBIF DQ group
  - Convenors:
    - Arthur Chapman, Antonio Mauro Saraiva
  - GBIF Liaison:
    - Dmitry Schigel
- Approved by the Exec in October 2014
- Discussions held at TDWG2014

Objectives of DQ Interest Group

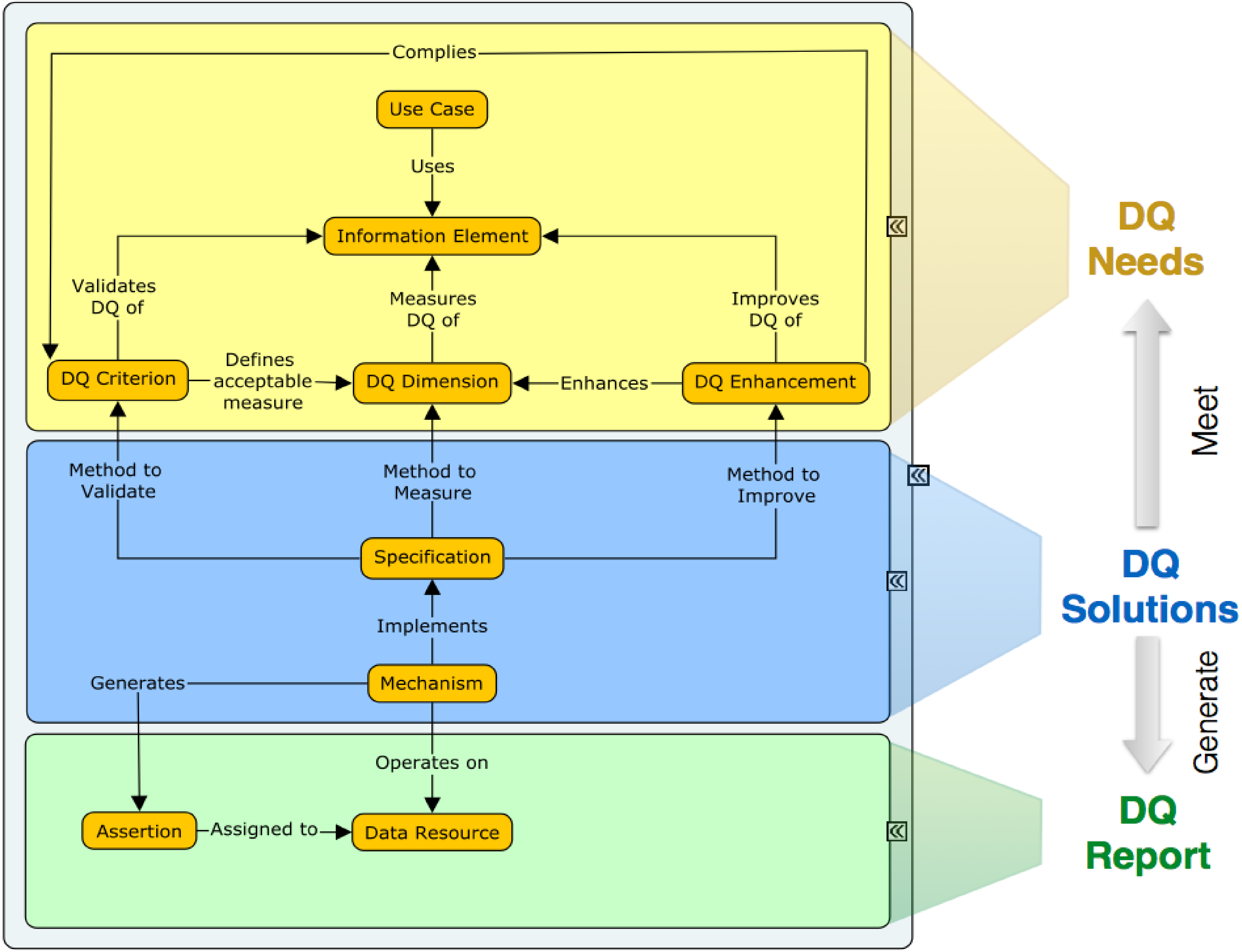
- Discuss, formalise, standardise and develop
  - Concepts, issues, policies, methodologies, metadata, tools and mechanisms, training
- Promote associated best practices throughout the Biodiversity Informatics community
- Collaborate

Activity since TDWG 2014

- Established 3 Task Groups
  - TG1 Framework on Data Quality (Allan Koch Veiga)
  - TG2 Tools, Services and Work Flows (Lee Belbin)
  - TG3 Use Case Library (Miles Nicholls)
- Approx 100 members have expressed interest in participating

  (however very few actually doing so)
- Framework Document has been submitted to Plos1 for publication.

Task Group 1: Framework for Data Quality



Interest Group Workplan for 2016

- Encourage greater participation in Task Groups
- Liaise with GBIF Working Groups
  - Fitness for Use for Agrobiodiversity
  - Fitness for Use for Distribution Modelling
- Held meeting March 2016 in Brazil
- Developed Use Cases for
  - Agrobiodiversity and
  - Species Distribution Modelling
- Consolidate Task Group reports and publications for TDWG2016
- Meet in Costa Rica at TDWG2016

GBIF-TDWG Task Group 2

Tools, Services and Workflows

Lee Belbin

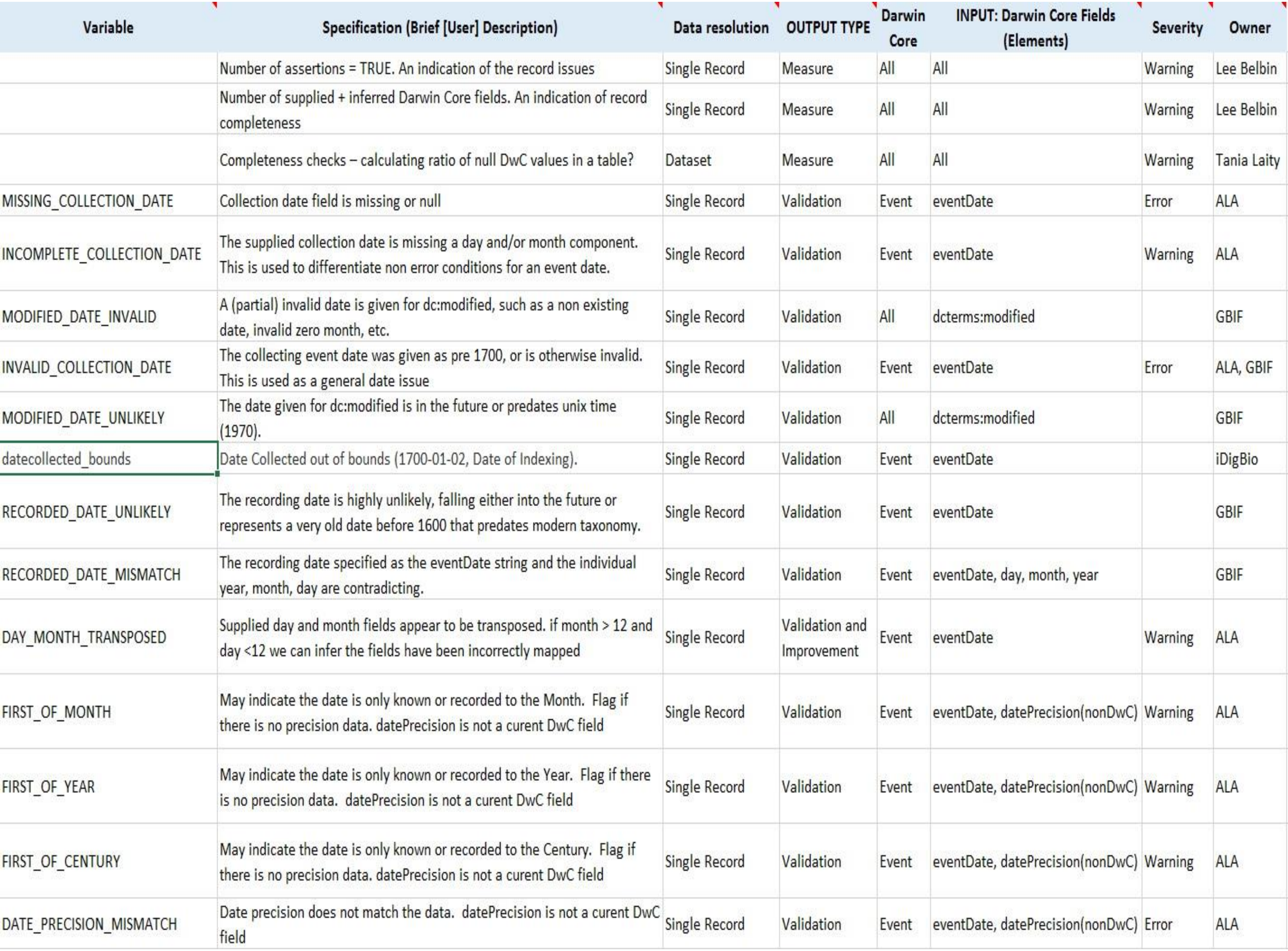
Data Publishers must provide

1. Record and dataset-level evaluations

2. An environment that makes it efficient to determine ‘fitness for use’

3. All the data*

I 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

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

Task Group Purpose

Task Group 3: Use Case Library

(Miles Nicholls)

The third task group of the GBIF TDWG biodiversity data quality interest group

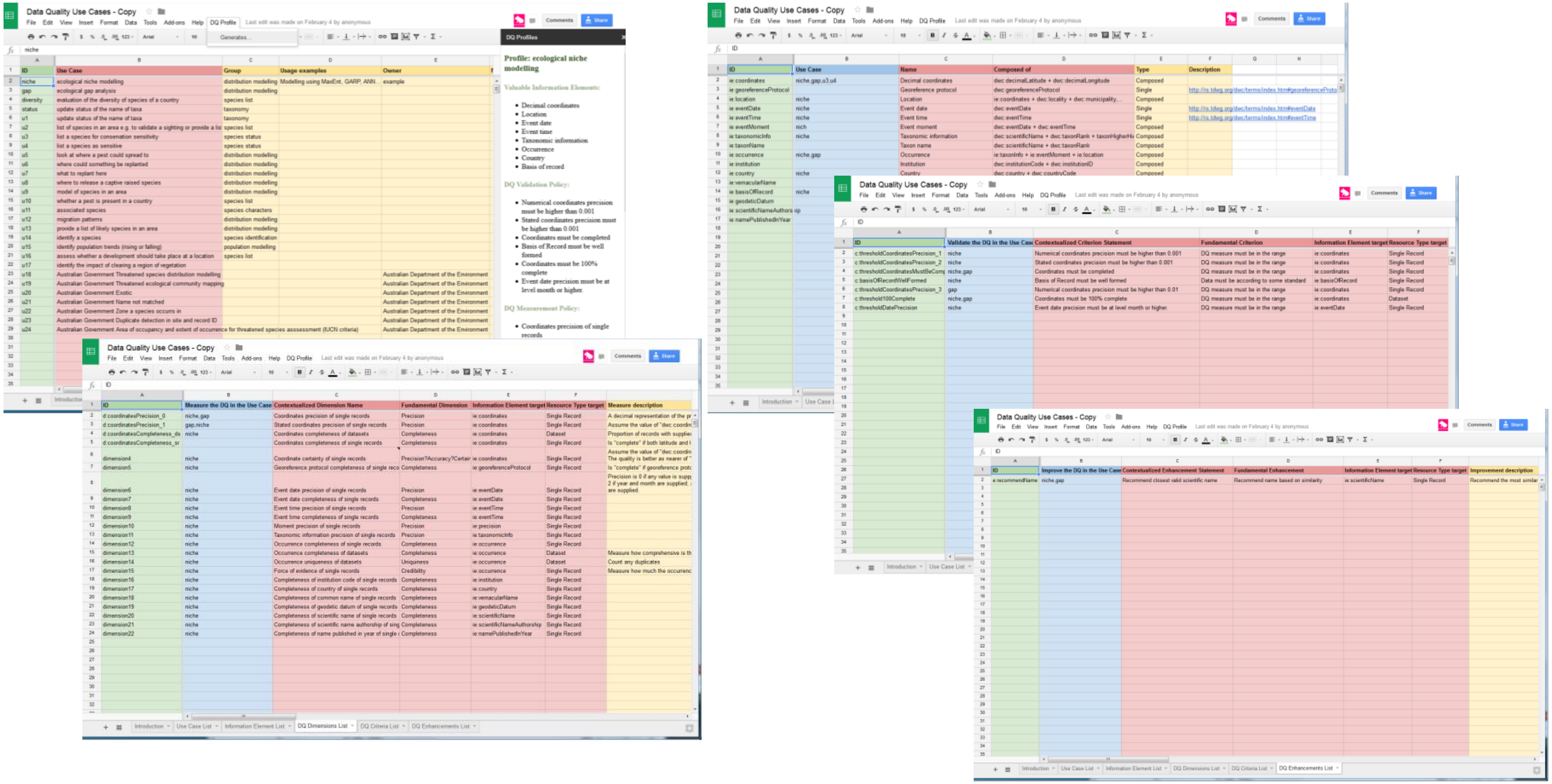
Tasked to assemble a library of use cases describing specific examples of data selection

Activities

- Use the framework from TG1 to develop a use case description store – a series of related worksheets to capture a use case.

- From this point it has been a process of iterating over use case description mechanisms to try and reach a usable tool:

Use case description store



Data Quality Use Cases - Copy

ID	Use Case	Group	Usage examples	Owner
niche	ecological niche modelling	distribution modelling	Modelling using MaxEnt, GARP, ANN...	example
gap	ecological gap analysis	distribution modelling		
diversity	evaluation of the diversity of species of a country	species list		
status	update status of the name of taxa	taxonomy		
u1	update status of the name of taxa	taxonomy		
u2	list of species in an area e.g. to validate a sighting or provide a lis	species list		
u3	list a species for conservation sensitivity	species status		
u4	list a species as sensitive	species status		
u5	look at where a pest could spread to	distribution modelling		
u6	where could something be replanted	distribution modelling		
u7	what to replant here	distribution modelling		
u8	where to release a captive raised species	distribution modelling		
u9	model of species in an area	distribution modelling		
u10	whether a pest is present in a country	species list		
u11	associated species	species characters		
u12	migration patterns	distribution modelling		
u13	provide a list of likely species in an area	distribution modelling		
u14	identify a species	species identification		
u15	identify population trends (rising or falling)	population modelling		
u16	assess whether a development should take place at a location	species list		
u17	identify the impact of clearing a region of vegetation			
u18	Australian Government Threatened species distribution modelling			Australian Department of the Environment
u19	Australian Government Threatened ecological community mapping			Australian Department of the Environment
u20	Australian Government Exotic			Australian Department of the Environment
u21	Australian Government Name not matched			Australian Department of the Environment
u22	Australian Government Zone a species occurs in			Australian Department of the Environment
u23	Australian Government Duplicate detection in site and record ID			Australian Department of the Environment
u24	Australian Government Area of occupancy and extent of occurrence for threatened species asssessment (IUCN criteria)			Australian Department of the Environment

DQ Profiles

Profile: ecological niche modelling

Valuable Information Elements:

- Decimal coordinates
- Location
- Event date
- Event time
- Taxonomic information
- Occurrence
- Country
- Basis of record

DQ Validation Policy:

- Numerical coordinates precision must be higher than 0.001
- Stated coordinates precision must be higher than 0.001
- Coordinates must be completed
- Basis of Record must be well formed
- Coordinates must be 100% complete
- Event date precision must be at level month or higher.

DQ Measurement Policy:

- Coordinates precision of single records

ID	Use Case	Name	Composed of	Type	Description
ie:coordinates	niche,gap,u3,u4	Decimal coordinates	dwc:decimalLatitude + dwc:decimalLongitude	Composed	
ie:georeferenceProtocol		Georeference protocol	dwc:georeferenceProtocol	Single	http://rs.tdwg.org/dwc/terms/index.htm#georeferenceProto
ie:location	niche	Location	ie:coordinates + dwc:locality + dwc:municipality,...	Composed	
ie:eventDate	niche	Event date	dwc:eventDate	Single	http://rs.tdwg.org/dwc/terms/index.htm#eventDate
ie:eventTime	niche	Event time	dwc:eventTime	Single	http://rs.tdwg.org/dwc/terms/index.htm#eventTime
ie:eventMoment	nich	Event moment	dwc:eventDate + dwc:eventTime	Composed	
ie:taxonomicInfo	niche	Taxonomic information	dwc:scientificName + dwc:taxonRank + taxonHigherHi	Composed	
ie:taxonName		Taxon name	dwc:scientificName + dwc:taxonRank	Composed	
ie:occurrence	niche,gap	Occurrence	ie:taxonInfo + ie:eventMoment + ie:location	Composed	
ie:institution		Institution	dwc:institutionCode + dwc:institutionID	Composed	
ie:country	niche	Country	dwc:country + dwc:countryCode	Composed	
ie:vernacularName					
ie:basisOfRecord	niche				
ie:geodeticDatum					
ie:scientificNameAuthorship					
ie:namePublishedInYear					

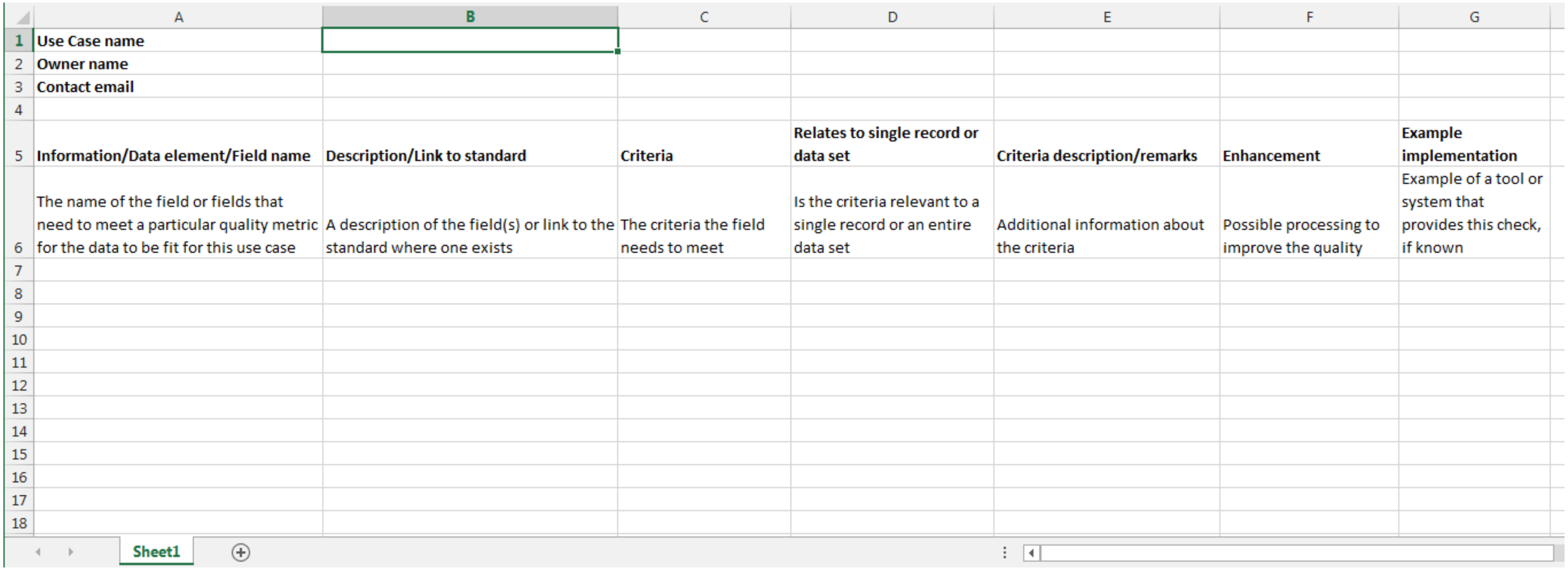
ID	Validate the DQ in the Use Case	Contextualized Criterion Statement	Fundamental Criterion	Information Element target	Resource Type target
c:thresholdCoordinatesPrecision_1	niche	Numerical coordinates precision must be higher than 0.001	DQ measure must be in the range	ie:coordinates	Single Record
c:thresholdCoordinatesPrecision_2	niche	Stated coordinates precision must be higher than 0.001	DQ measure must be in the range	ie:coordinates	Single Record
c:thresholdCoordinatesMustBeComp	niche,gap	Coordinates must be completed	DQ measure must be in the range	ie:coordinates	Single Record
c:basisOfRecordWellFormed	niche	Basis of Record must be well formed	Data must be according to some standard	ie:basisOfRecord	Single Record
c:thresholdCoordinatesPrecision_3	gap	Numerical coordinates precision must be higher than 0.01	DQ measure must be in the range	ie:coordinates	Single Record
c:threshold100Complete	niche,gap	Coordinates must be 100% complete	DQ measure must be in the range	ie:coordinates	Dataset
c:thresholdDatePrecision	niche	Event date precision must be at level month or higher.	DQ measure must be in the range	ie:eventDate	Single Record

ID	Measure the DQ in the Use Case	Contextualized Dimension Name	Fundamental Dimension	Information Element target	Resource Type target	Measure description
d:coordinatesPrecision_0	niche,gap	Coordinates precision of single records	Precision	ie:coordinates	Single Record	A decimal representation of the pr
d:coordinatesPrecision_1	gap,niche	Stated coordinates precision of single records	Precision	ie:coordinates	Single Record	Assume the value of "dwc:coordin
d:coordinatesCompleteness_ds	niche	Coordinates completeness of datasets	Completeness	ie:coordinates	Dataset	Proportion of records with supplie
d:coordinatesCompleteness_sr		Coordinates completeness of single records	Completeness	ie:coordinates	Single Record	Is "complete" if both latitude and l
dimension4	niche	Coordinate certainty of single records	Precision?Accuracy?Certair	ie:coordinates	Single Record	Assume the value of "dwc:coordin The quality is better as nearer of "
dimension5	niche	Georeference protocol completeness of single reco	Completeness	ie:georeferenceProtocol	Single Record	Is "complete" if georeference protc
dimension6	niche	Event date precision of single records	Precision	ie:eventDate	Single Record	Precision is 0 if any value is supp 2 if year and month are supplied, : are supplied
dimension7	niche	Event date completeness of single records	Completeness	ie:eventDate	Single Record	
dimension8	niche	Event time precision of single records	Precision	ie:eventTime	Single Record	
dimension9	niche	Event time completeness of single records	Completeness	ie:eventTime	Single Record	
dimension10	niche	Moment precision of single records	Precision	ie:precision	Single Record	
dimension11	niche	Taxonomic information precision of single records	Precision	ie:taxonomicInfo	Single Record	
dimension12	niche	Occurrence completeness of single records	Completeness	ie:occurrence	Single Record	
dimension13	niche	Occurrence completeness of datasets	Completeness	ie:occurrence	Dataset	Measure how comprehensive is th
dimension14	niche	Occurrence uniqueness of datasets	Uniquiness	ie:occurrence	Dataset	Count any duplicates
dimension15	niche	Force of evidence of single records	Credibility	ie:occurrence	Single Record	Measure how much the occurrenc
dimension16	niche	Completeness of institution code of single records	Completeness	ie:institution	Single Record	
dimension17	niche	Completeness of country of single records	Completeness	ie:country	Single Record	
dimension18	niche	Completeness of common name of single records	Completeness	ie:vernacularName	Single Record	
dimension19	niche	Completeness of geodetic datum of single records	Completeness	ie:geodeticDatum	Single Record	
dimension20	niche	Completeness of scientific name of single records	Completeness	ie:scientificName	Single Record	
dimension21	niche	Completeness of scientific name authorship of sing	Completeness	ie:scientificNameAuthorship	Single Record	
dimension22	niche	Completeness of name published in year of single i	Completeness	ie:namePublishedInYear	Single Record	

ID	Improve the DQ in the Use Case	Contextualized Enhancement Statement	Fundamental Enhancement	Information Element target	Resource Type target	Improvement description
e:recommendName	niche,gap	Recommend closest valid scientific name	Recommend name based on similarity	ie:scientificName	Single Record	Recommend the most similar

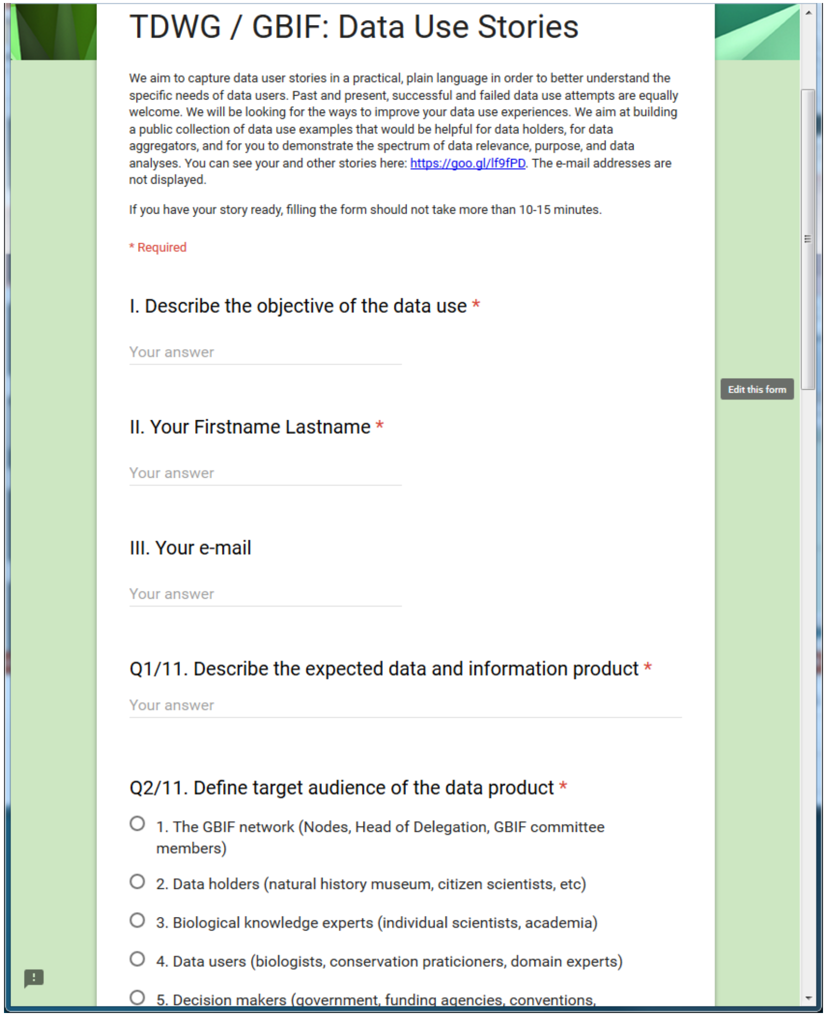
https://docs.google.com/spreadsheets/d/1dr6jSuxlkg2eLFas-VaPL8dAGouAP-ZJxRwVfrVKUGc/edit?ts=561ec9cb#gid=0

Offline worksheet



	A	B	C	D	E	F	G
1	**Use Case name**						
2	**Owner name**						
3	**Contact email**						
4							
5	**Information/Data element/Field name**	**Description/Link to standard**	**Criteria**	**Relates to single record or data set**	**Criteria description/remarks**	**Enhancement**	**Example implementation**
6	The name of the field or fields that need to meet a particular quality metric for the data to be fit for this use case	A description of the field(s) or link to the standard where one exists	The criteria the field needs to meet	Is the criteria relevant to a single record or an entire data set	Additional information about the criteria	Possible processing to improve the quality	Example of a tool or system that provides this check, if known
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							

Google form



TDWG / GBIF: Data Use Stories

We aim to capture data user stories in a practical, plain language in order to better understand the specific needs of data users. Past and present, successful and failed data use attempts are equally welcome. We will be looking for the ways to improve your data use experiences. We aim at building a public collection of data use examples that would be helpful for data holders, for data aggregators, and for you to demonstrate the spectrum of data relevance, purpose, and data analyses. You can see your and other stories here: https://goo.gl/lf9fPD. The e-mail addresses are not displayed.

If you have your story ready, filling the form should not take more than 10-15 minutes.

* Required

I. Describe the objective of the data use *

Your answer

II. Your Firstname Lastname *

Your answer

III. Your e-mail

Your answer

Q1/11. Describe the expected data and information product *

Your answer

Q2/11. Define target audience of the data product *

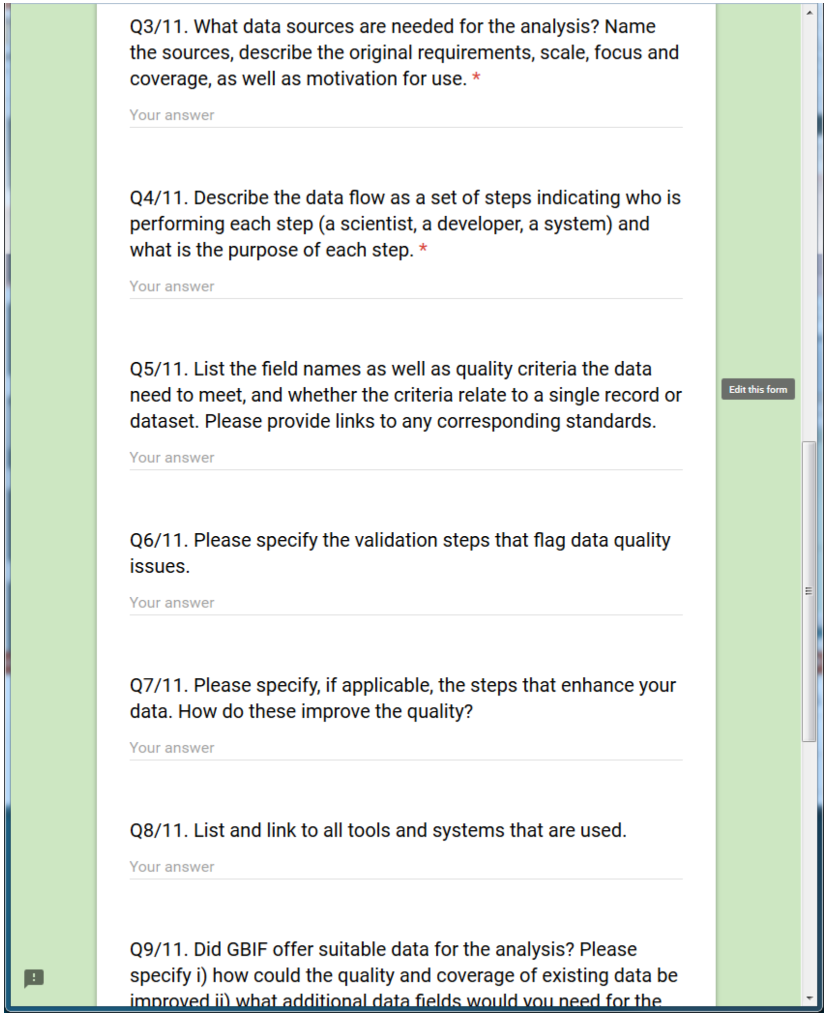
○ 1. The GBIF network (Nodes, Head of Delegation, GBIF committee members)

○ 2. Data holders (natural history museum, citizen scientists, etc)

○ 3. Biological knowledge experts (individual scientists, academia)

○ 4. Data users (biologists, conservation praticioners, domain experts)

○ 5. Decision makers (government, funding agencies, conventions,



Q3/11. What data sources are needed for the analysis? Name the sources, describe the original requirements, scale, focus and coverage, as well as motivation for use. *

Your answer

Q4/11. Describe the data flow as a set of steps indicating who is performing each step (a scientist, a developer, a system) and what is the purpose of each step. *

Your answer

Q5/11. List the field names as well as quality criteria the data need to meet, and whether the criteria relate to a single record or dataset. Please provide links to any corresponding standards.

Your answer

Q6/11. Please specify the validation steps that flag data quality issues.

Your answer

Q7/11. Please specify, if applicable, the steps that enhance your data. How do these improve the quality?

Your answer

Q8/11. List and link to all tools and systems that are used.

Your answer

Q9/11. Did GBIF offer suitable data for the analysis? Please specify i) how could the quality and coverage of existing data be improved ii) what additional data fields would you need for the

https://docs.google.com/forms/d/1udKhE9rdr2txkDE8MiBK-wHHVbr4G5mcdktq6P3rFXk/viewform

Why collect use cases

- To build a use case library as a resource to research how data is selected for particular purposes
- Allow reuse of use cases for consistent quality assessment
- To build profiles of the most used information elements, dimensions and criteria and feed these back into the development of data capture and quality assessment tools
- Begin to develop automated data selection for use cases