TG3: Use Case Library

Data Quality

Miles Nicholls
Data Manager, Atlas of Living Australia
miles.nicholls@csiro.au

Context and purpose

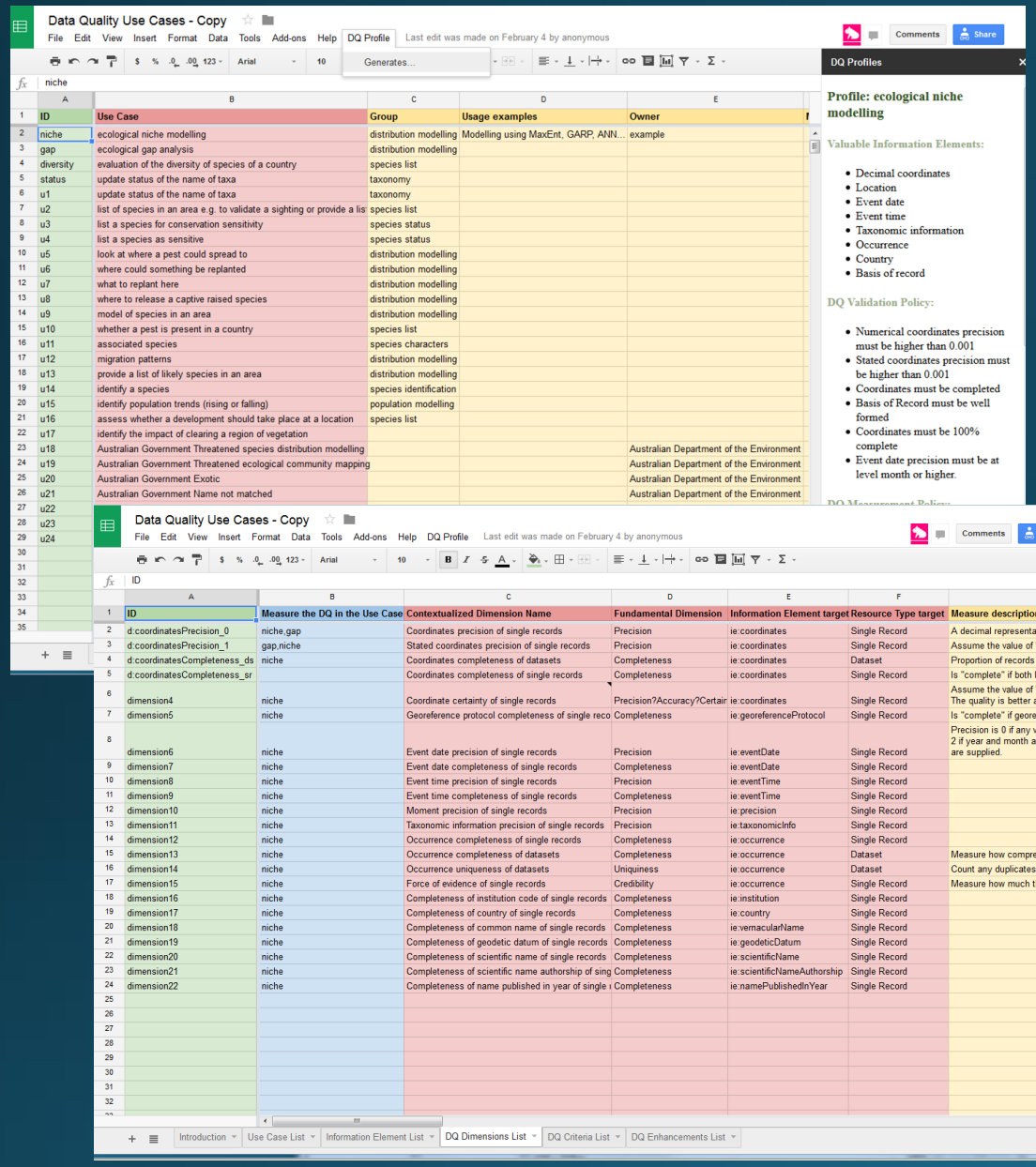
- The third task group of the GBIF TDWG biodiversity data quality interest group
- Tasked to assemble a library of use cases describing specific examples of data selection

Anticipated Process

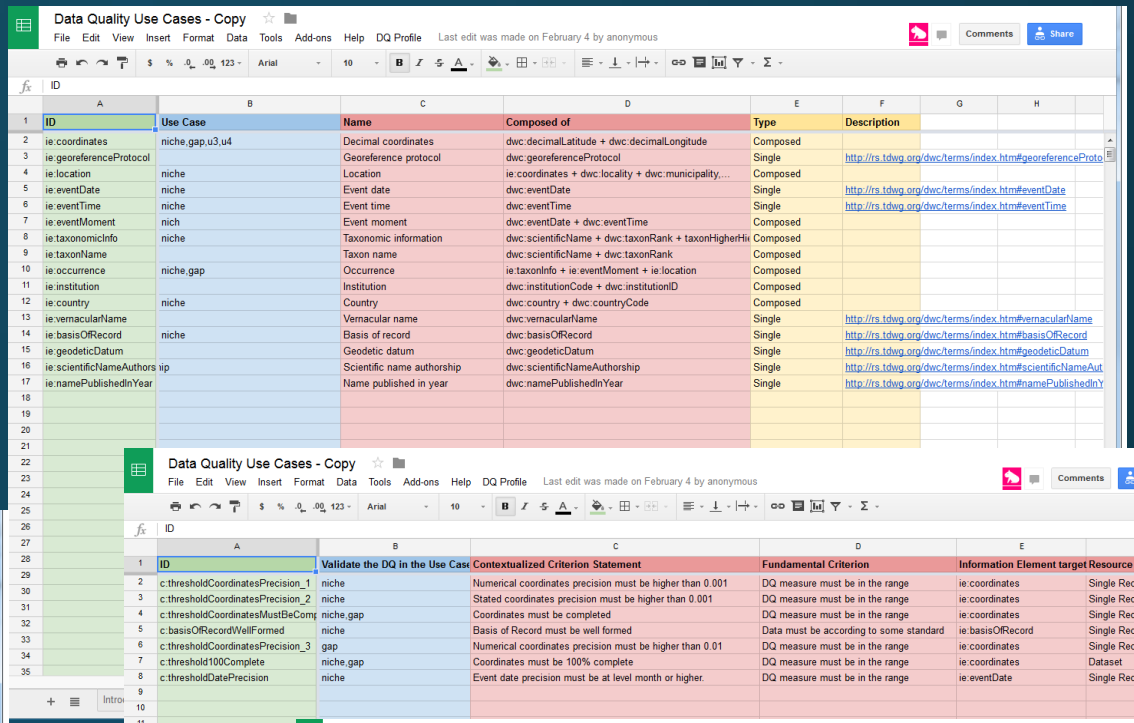
1. Use the framework from TG1 to understand the information that needs to be captured in a data quality use case
2. Create a model to store use case descriptions
3. Capture use cases
4. Analyse the use case elements
5. Cross reference the use cases against the tools from TG2
6. Identify areas for additional quality tools

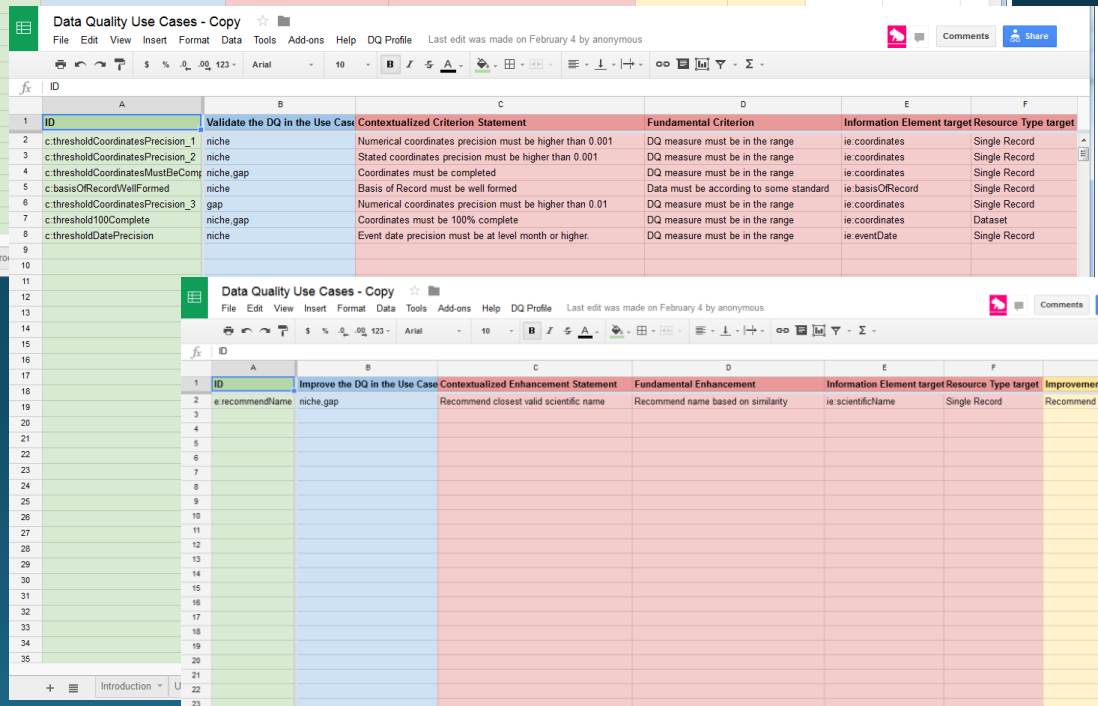
Actual process

- The first few steps were relatively straight forward:
- Running through a few examples helped in understanding the practical application of the framework and led to the development of a use case description store – a series of related worksheets to capture a use case:
  - Sheets that collect the elements from the framework: Use Case Title, Information Elements, Data Quality Dimensions, Data Quality Criteria and Data Quality Enhancements
- Allan Koch Veiga added a DQ profile tool which produces a report based on the information in the worksheets


Data Quality Use Cases - Copy
File Edit View Insert Format Data Tools Add-ons Help DQ Profile
Last edit was made on February 4 by anonymous
Comments
Share
Generates...
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.
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
u23
u24

Data Quality Use Cases - Copy
File Edit View Insert Format Data Tools Add-ons Help DQ Profile
Last edit was made on February 4 by anonymous
Comments
Share
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 proto
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
Uniqueness
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
Completeness
ie:namePublishedInYear
Single Record
Introduction
Use Case List
Information Element List
DQ Dimensions List
DQ Criteria List
DQ Enhancements List


Data Quality Use Cases - Copy
File Edit View Insert Format Data Tools Add-ons Help DQ Profile
Last edit was made on February 4 by anonymous
Comments
Share
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
Vernacular name
dwc:vernacularName
Single
http://rs.tdwg.org/dwc/terms/index.htm#vernacularName
ie:basisOfRecord
niche
Basis of record
dwc:basisOfRecord
Single
http://rs.tdwg.org/dwc/terms/index.htm#basisOfRecord
ie:geodeticDatum
Geodetic datum
dwc:geodeticDatum
Single
http://rs.tdwg.org/dwc/terms/index.htm#geodeticDatum
ie:scientificNameAuthors
hip
Scientific name authorship
dwc:scientificNameAuthorship
Single
http://rs.tdwg.org/dwc/terms/index.htm#scientificNameAut
ie:namePublishedInYear
Name published in year
dwc:namePublishedInYear
Single
http://rs.tdwg.org/dwc/terms/index.htm#namePublishedInY


Data Quality Use Cases - Copy
File Edit View Insert Format Data Tools Add-ons Help DQ Profile
Last edit was made on February 4 by anonymous
Comments
Share
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:thresholdCoordinatesMustBeCom
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
Introduction

Data Quality Use Cases - Copy
File Edit View Insert Format Data Tools Add-ons Help DQ Profile
Last edit was made on February 4 by anonymous
Comments
Share
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
Introduction
Use Case List
Information Element List
DQ Dimensions List
DQ Criteria List
DQ Enhancements List

Actual process

- Review of the initial worksheets indicated that although it collected the required information the level of technical understanding required to fill in the series of related sheets was likely too high for it to be widely used.
- That the worksheets were online and didn't export well was also a barrier to data entry
- So an additional offline sheet was developed focussed on capturing a single use case as simply as possible

	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							

Sheet1

- Once the template was filled in it would be converted to the more complex data structure manually

Actual Process

- The template was then posted up on the GBIF community site with some help material, a worked example and a request for contributions
- …
- A further post was added to indicate that an email with a text description would be enough to start the process
- …
- Unfortunately there has been very little participation

Key challenge so far

- How to address the limited number of contributions:
  - Is the community aware of the working groups and what we're doing?
    - More and varied communication channels
    - More frequent updates on progress
    - More communication on the purpose and outcomes of the work to encourage participation
  - Are the proposed tools too difficult to use?
    - Needs feedback on the usability of the framework and worksheets from the community using them

What's next

- Continue to collect use cases
  - Improve communication
  - Publicise contributions
- Transfer use cases to the use case store
- Improve use case collection tools based on feedback

The future

- Encourage use of the use case library as a resource to research how data is selected for particular purposes
- Begin to build profiles of the most used information elements, dimensions and criteria
  - Feed these back to tools and initiatives that collect data
- Feedback use cases into the development of data quality assessment tools
- Begin to develop automated data selection for use cases

Thank You

Do you have a data quality use case?

Task Group 3 - Use Case Library for Data Quality