GIS GIGO (Garbage In Garbage Out): 30 checks for data errors

- Published on February 25, 2015

Picture: Photobucket

By: Nathan Heazlewood

Business Consulting Practice Lead at Eagle Technology

GIS GIGO (Garbage In Garbage Out):

A checklist of data errors

GARBAGE IN GARBAGE OUT: This is one of the most famous acronyms in the history of computing. What this means is that it doesn't matter if you have the most brilliant software and hardware: if your data is rubbish then your outputs or analysis will be rubbish as well. When I am working with data, particularly migrating or loading data between systems I check the accuracy of that data as part of the process (rather than loading it and then trying to work out why I am getting strange results later).

Some famous examples of where this has gone wrong include the Mars Orbiter where the basic mistake of assuming that an attribute was measured in imperial (English) units when in fact the data was metric. This simple error caused the destruction of an expensive spacecraft.

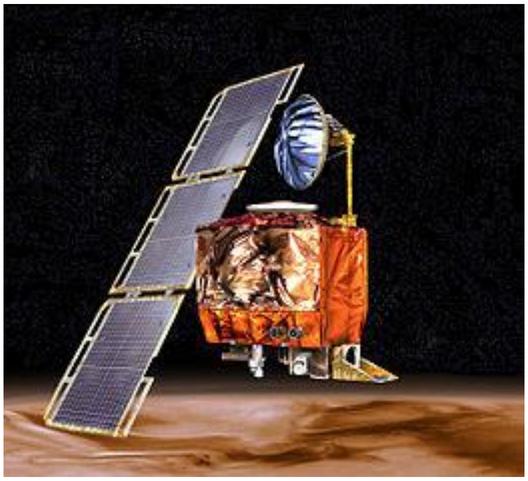


Image: Wikipedia

GIS data requires many more checks compared to standard 'text and numbers' data so I recommend working through a checklist of tests as I have provided in the list below. It should be noted that different GIS/CAD etc. software has different rules for things like how topology works. The following list is slightly targeted towards ArcGIS but should also provide some hints for users of other GIS/CAD software:

Positional accuracy checks

1 The data is of the required positional accuracy Data may have been acquired by a variety of means and at different times. Check whether the spatial accuracy of the data is fit for the use that you want to put it to. It should be noted that for some purposes even data that is hundreds of meters out of place can still be accurate enough for certain purposes.

2 The data is sufficiently spatially accurate when combined with the other data that will be used One recurring issue with GIS data is that because it is easy to 'overlay' different datasets on top of each other sometimes data that has different spatial accuracy characteristics can infer spatial relationships that are incorrect. For example, consider a situation of overlaying vector underground power cable data over an aerial photograph: where the power cable data is actually inaccurate with a common shift where the data is 2 meters in the wrong direction to the East, and the aerial photograph is 10 meters to the West: the aggregate error is now 12 meters if the two datasets are used together. Imagine now that someone uses this data to decide which side of a fence (visible in the aerial photograph) the vector power cable is on. It would be very easy for that person to infer from the data that the underground power cable was on the other side of a fence when this was an error- with potentially catastrophic results if this data is used to decide whether it is safe to excavate.

Check for topological logic

3 Line features must not overlap or self-overlap. If this logical for this feature type: e.g. a stream cannot intersect itself, however a road such as a highway off-ramp can cross over itself (via a flyover etc.)

4 Many polygon features generally should not overlap. E.g. regional government boundaries should not overlap. Often land parcels and other polygons shouldn’t overlap but there may be differences in this approach from country to country.

5 Some point features generally should be coincident with a line feature. E.g. generally railway station points need to sit on a railway line (exceptions to this could be newly built but not operational train stations or disused train stations where the line has been removed).

6 Certain polygons must not overlap with other types of polygons. E.g. a lake polygon should not overlap or be coincident with an ocean polygon.

There are other logical topological relationships to represent real world situations similar to those above but you get the idea- extend this list for the particular data that you are working with in your situation.

Geometric data considerations

7 One and only one geometry type is stored in each feature class. E.g. only polygon geometries used for a building footprint feature class- not points. It is important to note that some GIS systems make this distinction but others don’t: this can make migrating data between those systems difficult.

8 Only one geometry column per table

Projections and coordinate systems

9 The projected co-ordinate system along with datum matches any desired spatial projection/ referencing within the target system. Unless the system that you are loading data into has the capability to re-project data then you may need to ensure that it is in the correct projection when loading or unless re-projection 'on the fly' is appropriate for some reason- however note that this generally reduces system speed.

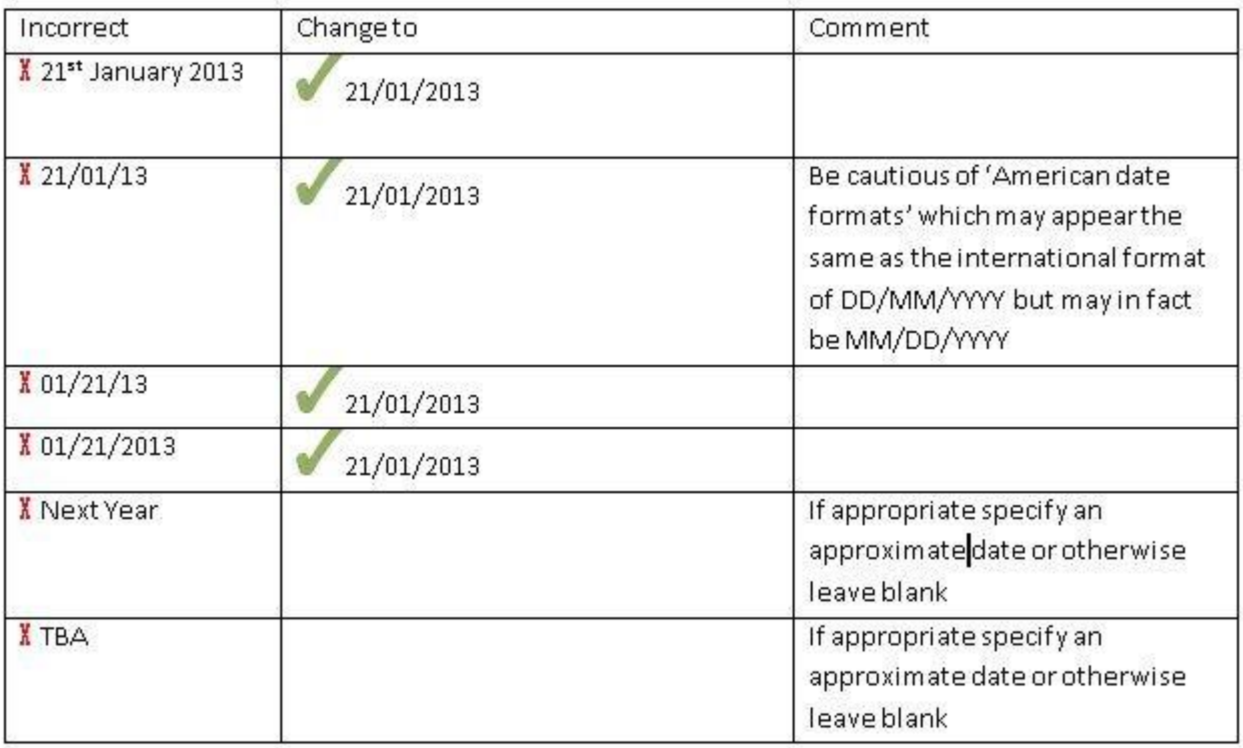
10 Negative values remain negative. One common error when migrating data between coordinate/projections is to accidentally lose negative values. This can also occur with other attributes that should permit negative values but where some part of a migration process causes an error.

Attribute and data structure checks

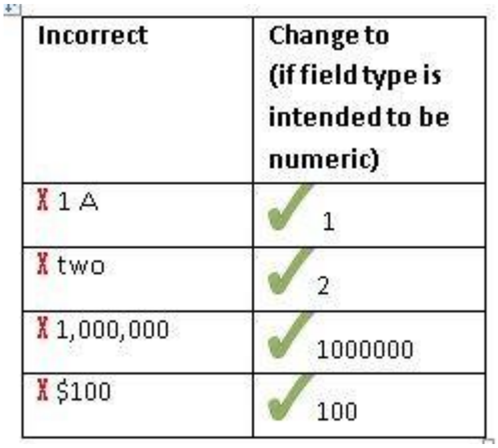
11 Attribute field names are correct and no 'reserved' names are used. Many systems have 'reserved' names that are already used internally within the system and therefore cannot be used for configurable parts of the system such as table or attribute names.

12 Attribute field data types are correct. E.g. an attribute called 'purchase date' is actually a date data type etc

13 All Attribute values are as per data types. E.g. a field that has a date data type only has dates recorded in that field.

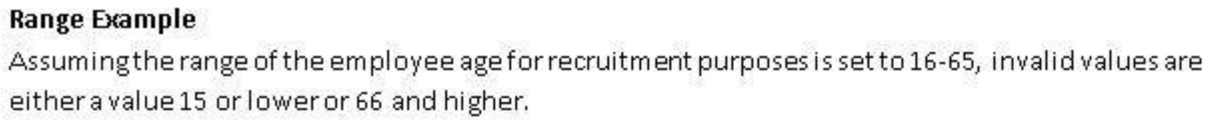


Incorrect	Change to	Comment
✗ 21st January 2013	✓ 21/01/2013	
✗ 21/01/13	✓ 21/01/2013	Be cautious of 'American date formats' which may appear the same as the international format of DD/MM/YYYY but may in fact be MM/DD/YYYY
✗ 01/21/13	✓ 21/01/2013	
✗ 01/21/2013	✓ 21/01/2013	
✗ Next Year		If appropriate specify an approximate date or otherwise leave blank
✗ TBA		If appropriate specify an approximate date or otherwise leave blank



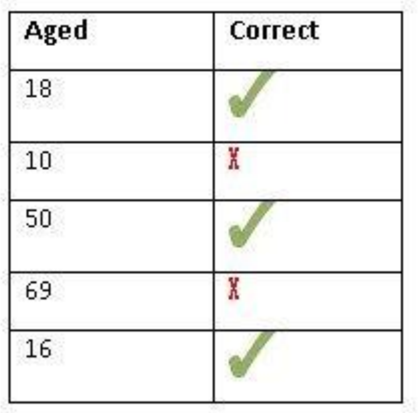
Incorrect	Change to (if field type is intended to be numeric)
✗ 1 A	✓ 1
✗ two	✓ 2
✗ 1,000,000	✓ 1000000
✗ $100	✓ 100

14 There are no values that are outside of the defined range of permitted values. E.g. if dates must be >=1/1/2015 then there should be no entries with dates earlier than this.



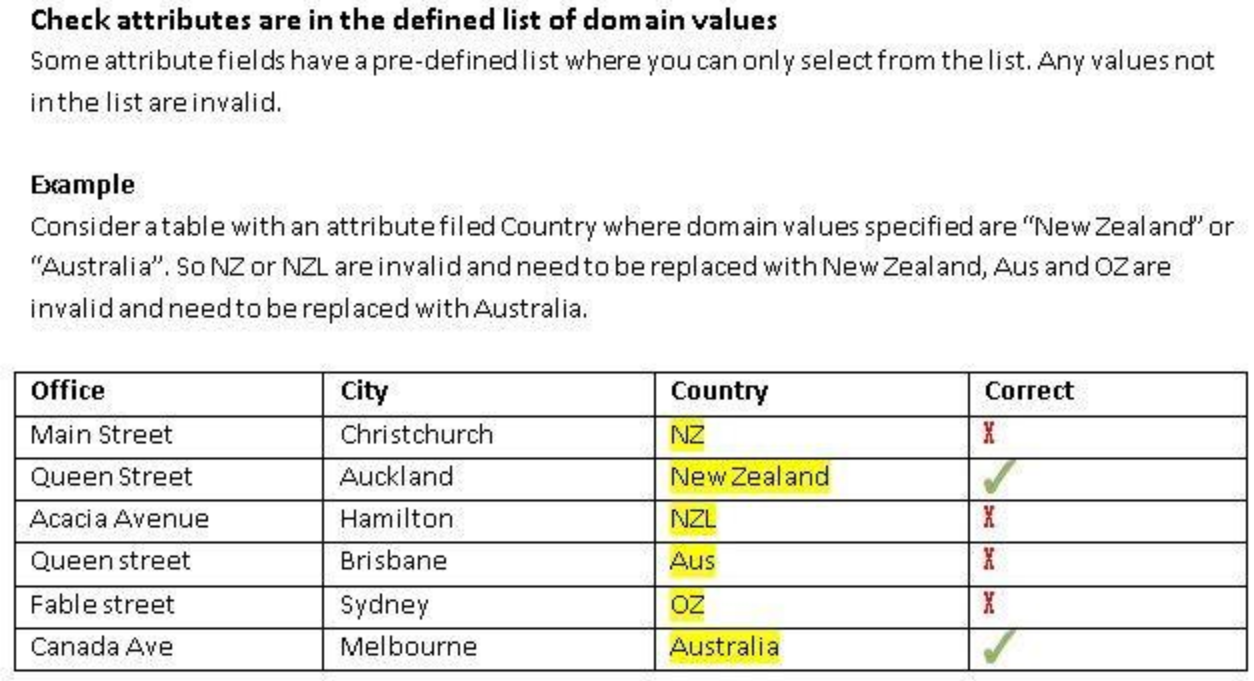
Range Example

Assuming the range of the employee age for recruitment purposes is set to 16-65, invalid values are either a value 15 or lower or 66 and higher.



Aged	Correct
18	✓
10	✗
50	✓
69	✗
16	✓

There are no values other than the domain values. E.g. if a domain of permitted values is a list of the states within the USA then an entry of “Ontario” would be an error.



Check attributes are in the defined list of domain values

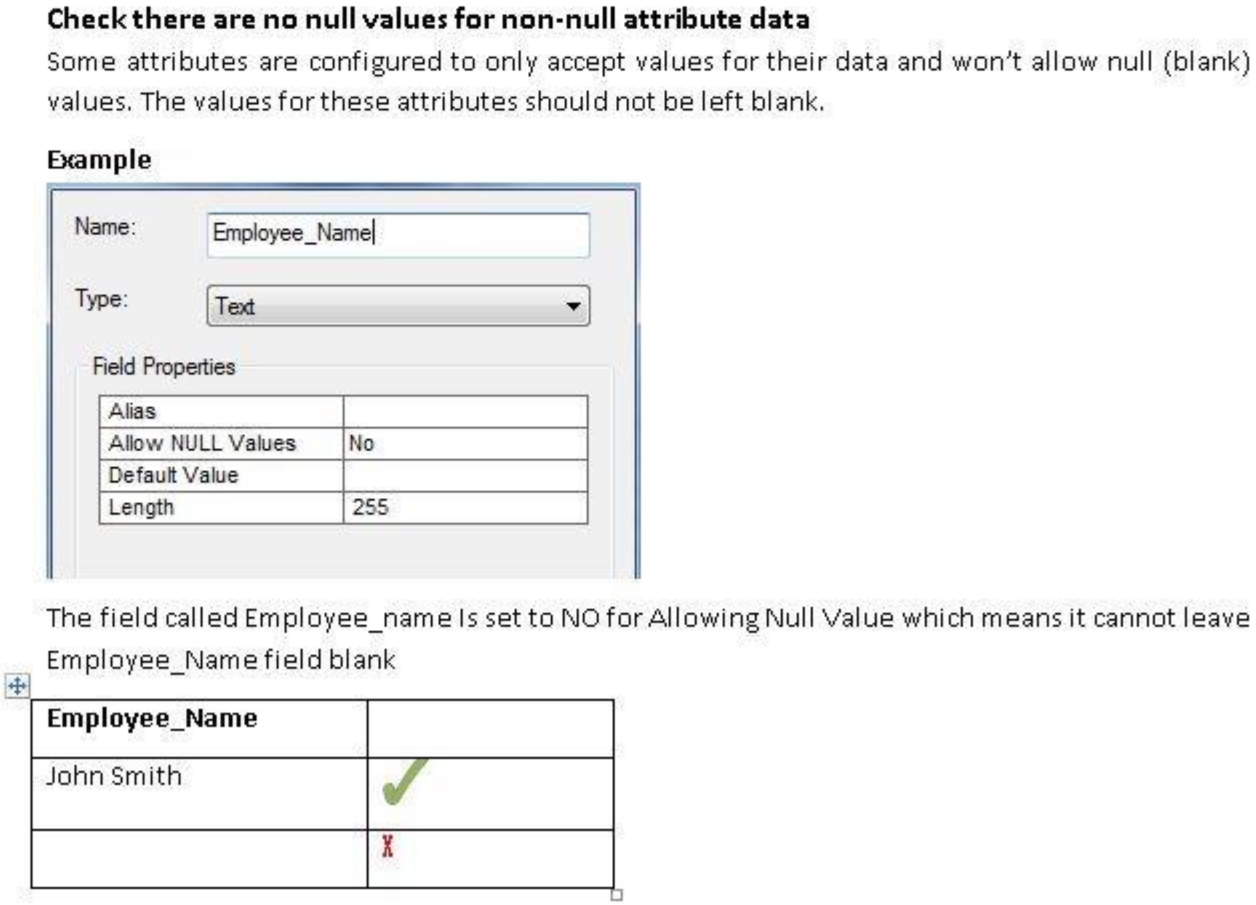
Some attribute fields have a pre-defined list where you can only select from the list. Any values not in the list are invalid.

Example

Consider a table with an attribute filed Country where domain values specified are “New Zealand” or “Australia”. So NZ or NZL are invalid and need to be replaced with New Zealand, Aus and OZ are invalid and need to be replaced with Australia.

Office	**City**	**Country**	**Correct**
Main Street	Christchurch	NZ	✗
Queen Street	Auckland	New Zealand	✓
Acacia Avenue	Hamilton	NZL	✗
Queen street	Brisbane	Aus	✗
Fable street	Sydney	OZ	✗
Canada Ave	Melbourne	Australia	✓

15 There are no entries with null values for non-null attribute data e.g. for a primary key.



Check there are no null values for non-null attribute data

Some attributes are configured to only accept values for their data and won't allow null (blank) values. The values for these attributes should not be left blank.

Example

Name: Employee_Name

Type: Text

Field Properties

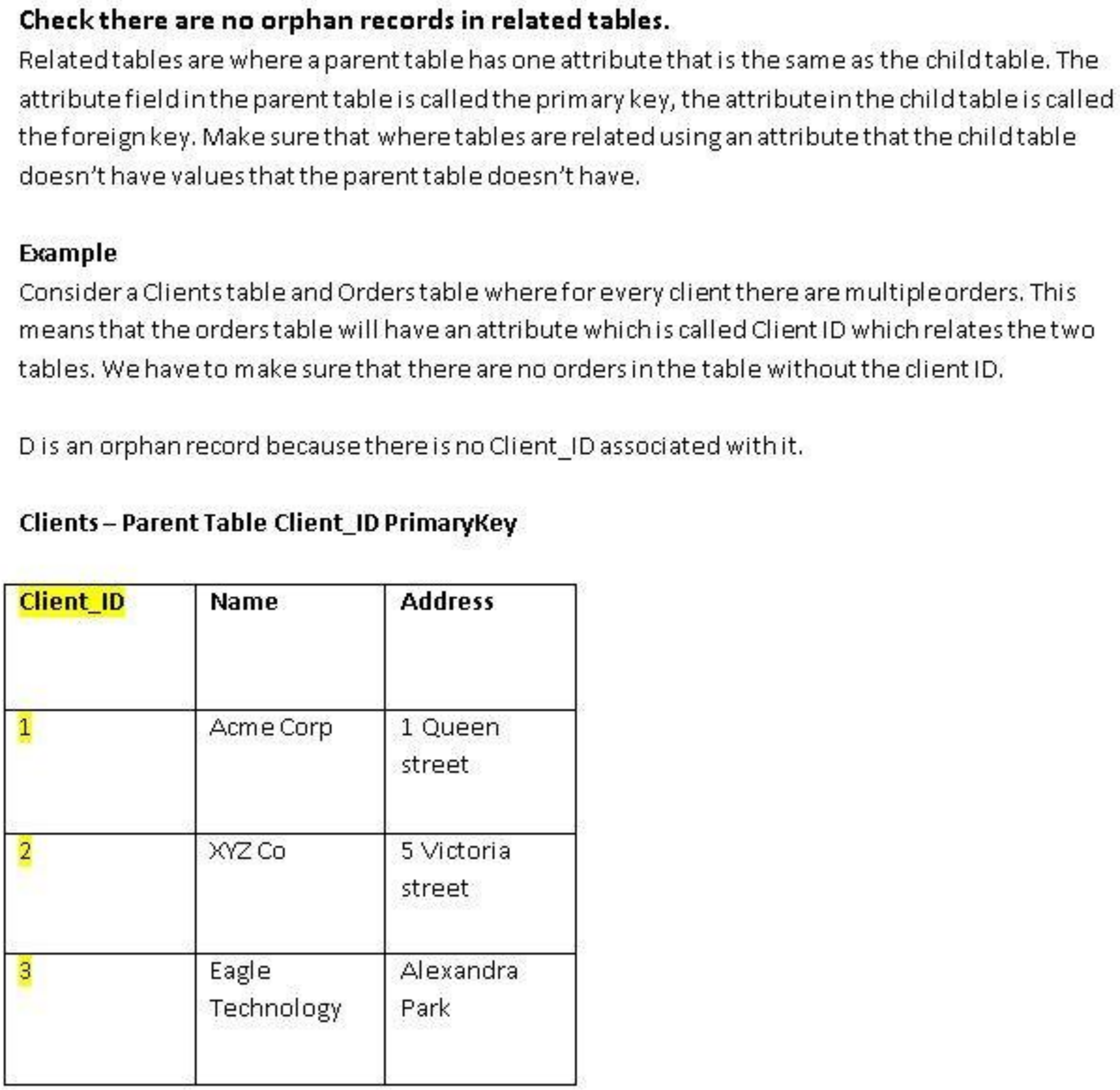
Alias	
Allow NULL Values	No
Default Value	
Length	255

The field called Employee_name Is set to NO for Allowing Null Value which means it cannot leave Employee_Name field blank

Employee_Name	
John Smith	✓
	✗

16 There are no duplicate entries for attributes where there should not be duplicates e.g. for a primary key.

17 There are no ‘orphan’ records in the related tables.



Check there are no orphan records in related tables.

Related tables are where a parent table has one attribute that is the same as the child table. The attribute field in the parent table is called the primary key, the attribute in the child table is called the foreign key. Make sure that where tables are related using an attribute that the child table doesn’t have values that the parent table doesn’t have.

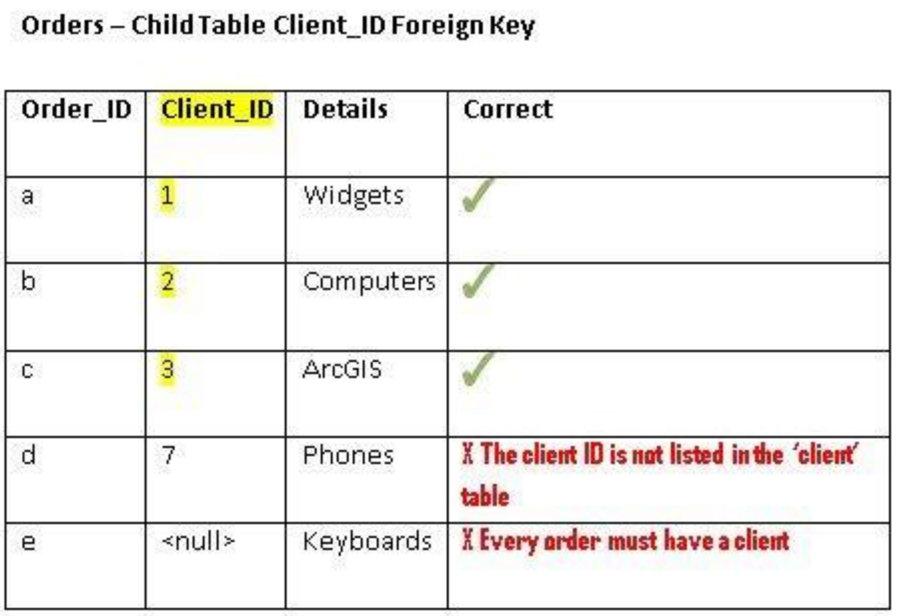
Example

Consider a Clients table and Orders table where for every client there are multiple orders. This means that the orders table will have an attribute which is called Client ID which relates the two tables. We have to make sure that there are no orders in the table without the client ID.

D is an orphan record because there is no Client_ID associated with it.

Clients – Parent Table Client_ID PrimaryKey

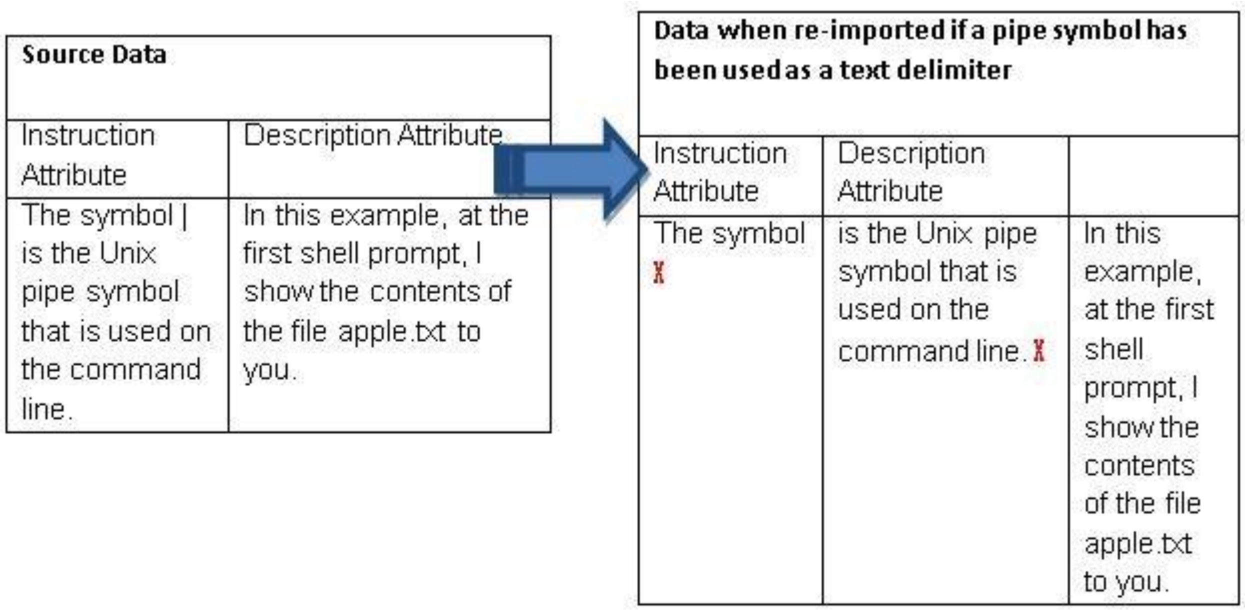
Client_ID	**Name**	**Address**
1	Acme Corp	1 Queen street
2	XYZ Co	5 Victoria street
3	Eagle Technology	Alexandra Park



Orders – Child Table Client_ID Foreign Key

Order_ID	Client_ID	Details	Correct
a	1	Widgets	✓
b	2	Computers	✓
c	3	ArcGIS	✓
d	7	Phones	X The client ID is not listed in the 'client' table
e	<null>	Keyboards	X Every order must have a client

18 Check whether special characters (punctuation symbols) are present in the data. If so check whether any systems that will present, the data will be affected by those characters. Certain characters such as < > / \ | are sometimes used for special purposes within systems and if these are present within the data it can cause issues (this issue is becoming less prevalent- but it still can affect some legacy systems)



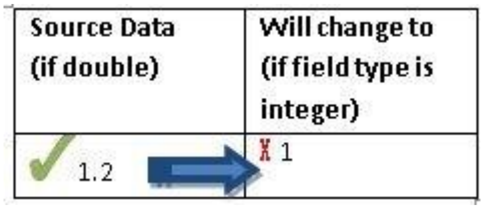
Source Data	
Instruction Attribute	Description Attribute
The symbol | is the Unix pipe symbol that is used on the command line.	In this example, at the first shell prompt, I show the contents of the file apple.txt to you.

Data when re-imported if a pipe symbol has been used as a text delimiter		
Instruction Attribute	Description Attribute	
The symbol X	is the Unix pipe symbol that is used on the command line. X	In this example, at the first shell prompt, I show the contents of the file apple.txt to you.

19 Check whether international characters are present If they are not able to be handled by the target software. Certain characters such as the umlaut: **ö** are common in Germanic languages or the macron common in M**ā**ori can cause difficulties for some systems. Other languages have their own alphabetic symbols. If these characters are present then check how they are to be dealt with in the target system. It is also relatively common to find that systems can store these characters but that search/query functions struggle to deal with them e.g. searching Google gives different results for "Wait**ā**kere" vs. "Wait**a**kere" even though the person searching is probably seeking the same results from either entry.

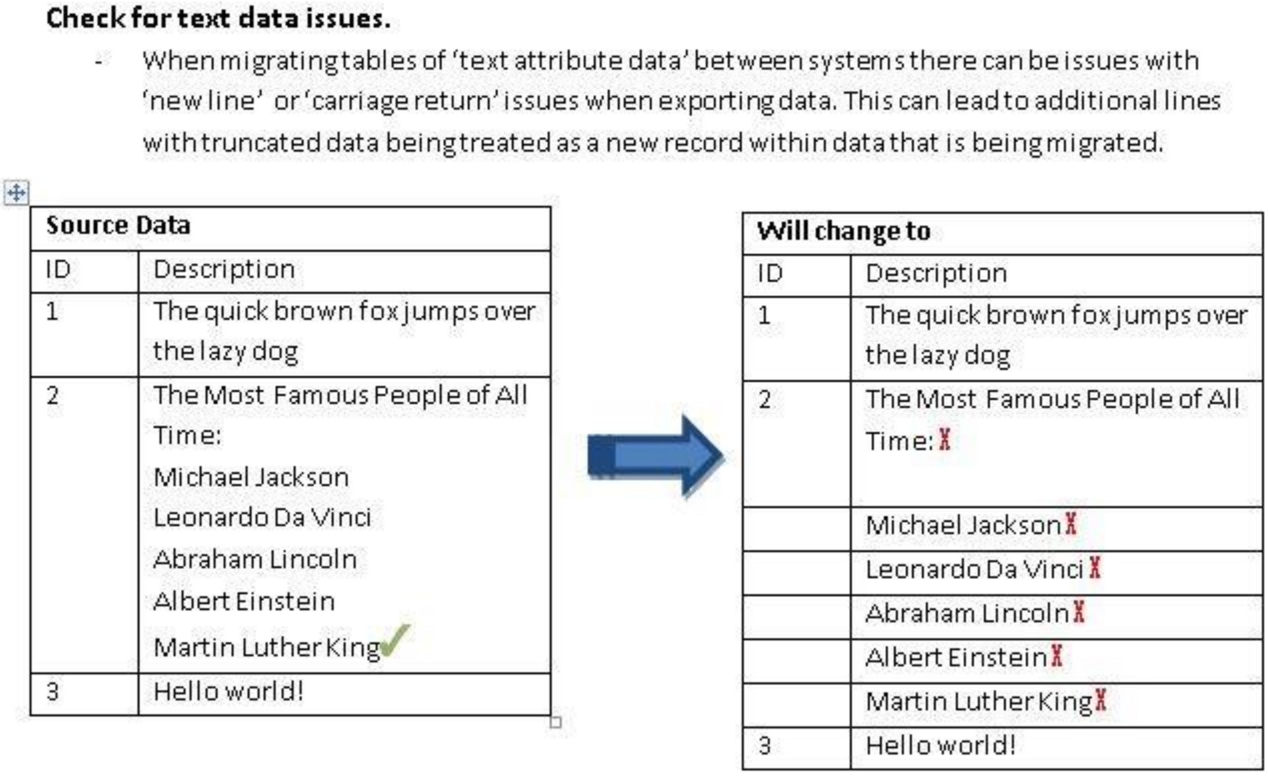
20 Check what aspects of data have been exported from a source system For example, it is common that line-work and attributes associated with vector data can be exported from CAD systems to GIS: however often this type of export does not bring annotation or symbologies through the export process, and so to get a comparable view of the same data significant additional work needs to be done in the GIS system. This is also common when moving 3D data between various systems: standards such as CityGML tend to mean that geometries/shapes can be transferred, however different systems have different ways of treating other elements of 3D models such as textures.

21 Check that the precision of source data has been retained It is a common error when moving data from system to system to 'lose the data after the decimal point' because someone forgets to set the number of decimal points correctly. Also in source data the number of decimal places may have been 5 decimal figures, but once through a transformation sometimes the data has been changed to just 2 decimal figures etc.



Source Data (if double)	Will change to (if field type is integer)
✓ 1.2	✗ 1

22 Check that Carriage Return (New Line) symbols won't cause issues



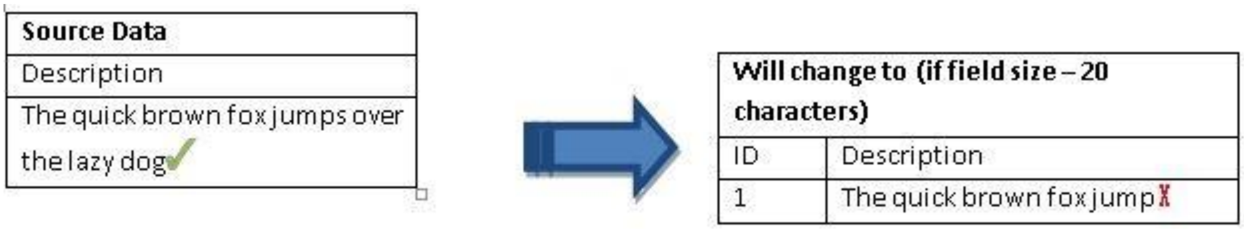
Check for text data issues.

- When migrating tables of 'text attribute data' between systems there can be issues with 'new line' or 'carriage return' issues when exporting data. This can lead to additional lines with truncated data being treated as a new record within data that is being migrated.

Source Data	
ID	Description
1	The quick brown fox jumps over the lazy dog
2	The Most Famous People of All Time: Michael Jackson Leonardo Da Vinci Abraham Lincoln Albert Einstein Martin Luther King✓
3	Hello world!

Will change to	
ID	Description
1	The quick brown fox jumps over the lazy dog
2	The Most Famous People of All Time: ✗
	Michael Jackson ✗
	Leonardo Da Vinci ✗
	Abraham Lincoln ✗
	Albert Einstein ✗
	Martin Luther King ✗
3	Hello world!

23 Check that data hasn't been truncated. There is a common error and problem with exporting with some systems that data is truncated. For example, a source system may have an attribute that allows up to say 1000 characters of text to be stored, but then someone will export the data and inadvertently only 255 characters will be exported.

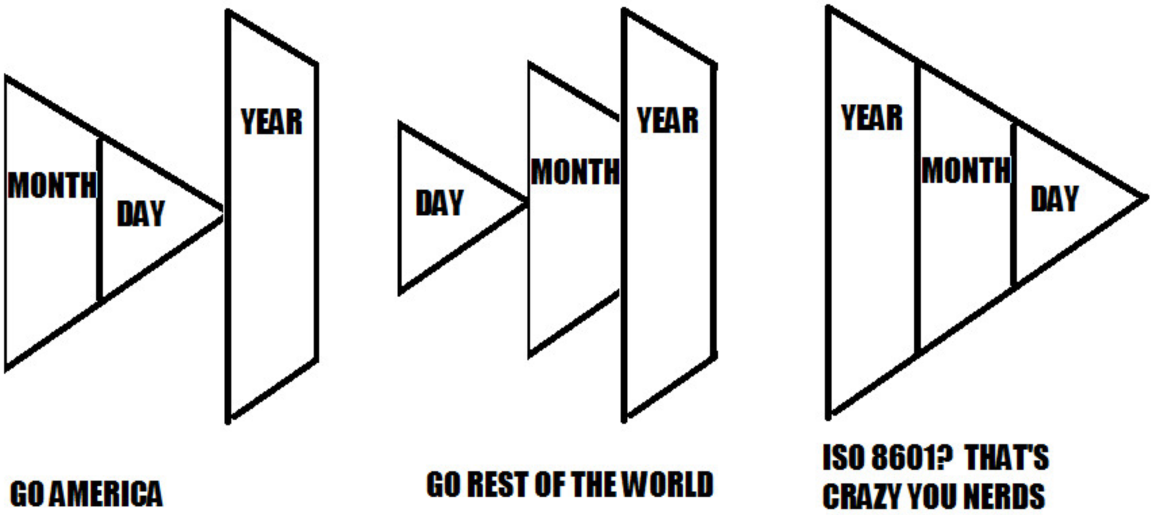


Source Data
Description
The quick brown fox jumps over the lazy dog✓

Will change to (if field size – 20 characters)	
ID	Description
1	The quick brown fox jump ✗

24 Check that attribute names won't be affected by truncation. Exporting from some systems to Shapefiles (or other formats) sometimes means that the attribute names are shortened- check whether or not this will be an issue.

25 Check that the super dumb American date format is being used or not used if you are moving data from another country. Or better yet everyone should just go YYYY/MM/DD mm:ss (big, medium, small - I like logic). In this day and age how is it possible that "1/2/2015" might mean either the First of February or the Second of January and there is no way of telling just by looking at it?



26 Check that units of measurement are consistent from source to target. For example, check that if weight is an attribute that you don't need to convert pounds to kilograms etc. during the transfer process.

27 Check that currency values have been converted if necessary A common error is to see an attribute named 'cost' and to assume that the values contained are in your local currency: if you are dealing with data from another country then you should check whether this is actually the case.

28 Check whether currency values include or exclude any relevant taxes A common error working with data is to assume that an attribute has already been recalculated for any applicable sales tax when in fact it hasn't.

29 Check that commonly confused characters do not cause issues. Postcodes (zip codes) in the UK are comprised of number and letter characters. Often the characters zero "0" and upper case O get confused within these codes which can affect processes like geocoding. Other easy to confuse characters include: lower case l and the number 1 and upper case I.

30 Check the timeliness of the data How up to date is the data? It should be noted that for some purposes even data that is several years old can still be accurate enough for certain purposes. It is common for some users to look at things like aerial photographs and because it looks 'real' they mistake it for being recent.

Check the metadata

The list above is a guideline describing the typical scenarios that organizations will come across when dealing with data quality, this is not an exhaustive list of every aspect of data quality. As each organization will have unique situations and may be migrating data from a variety of legacy software (both alternative GIS systems and/or non GIS systems) it is recommended that data owners/managers should take whatever additional steps required to ensure data quality of their own data.