Cabinet Office

National Fraud Initiative

Pre-Submission Data Quality Checks

Pre Submission Data Quality Checks

From experience we have found that carrying out a few basic checks on data can save many hours looking at spurious matches caused by poor quality data. These checks can best be made using specialist software, such as IDEA or ACL as there is no limit as to how many records the data contains.

However, many of the checks can be carried out using Excel but only if the number of records does not exceed the size of the worksheet available (1,048,576 Excel 2007 onwards, 65,536 pre 2007).

This document shows examples of these checks using IDEA and Excel.

Please note that whilst the screen shots in this guidance are from older versions of IDEA/Excel, all the functionality demonstrated is available on the more recent 'ribbon' menu versions of these applications.

Summary of checks

1. Sort/Index each field and check first and last entries
    a. Numeric fields (reasonableness)
    b. Character fields (validity)
    c. Date fields (consistent with specification)
    d. Blank fields, is that to be expected (completeness)
2. Control totals for monetary fields (reasonableness)
3. Analyse the main reference number, such as employee number, to establish whether there are any duplicate records (excessive)

Having carried out these checks the data may need to be re-extracted, where the data is shown to have fundamental flaws, or manipulated, where field formats or records can be adjusted to meet the specification.

Whether the data is re-extracted or manipulated we suggest that the pre-submission checks need to be carried out again to ensure that the previous issues have been rectified.

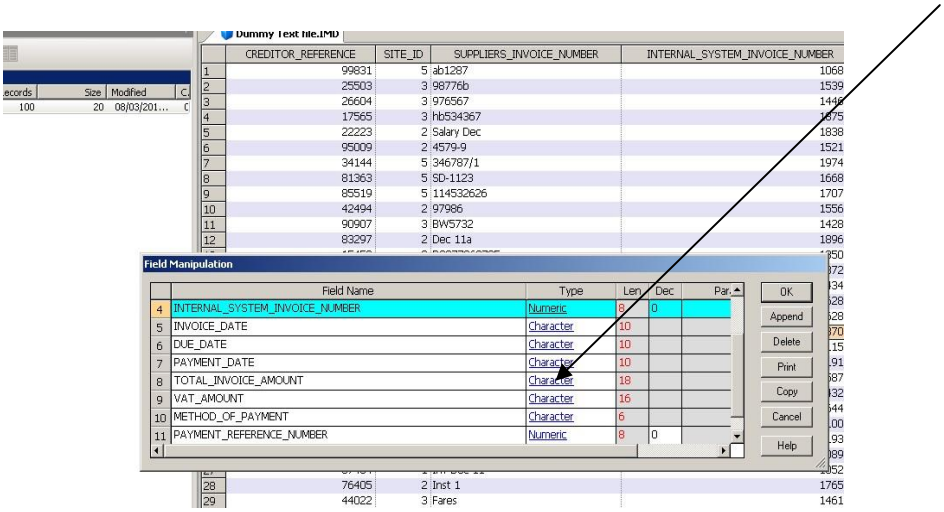
Using IDEA

1. Create a copy of the data to be submitted to ensure that you always have the original intact.

2. Set up client and import data file.

3. If there is no header row with field names then create one from the data specification using ‘Field Manipulation’.

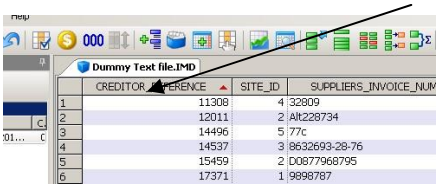
4. In addition, if the monetary or date fields have imported as character fields ensure that they are converted to virtual numeric and date fields.



5. Carry out the following checks:

Sorting

a. Sort the first column by double clicking on the column heading.



b. Inspect the lowest Creditor Reference (11308 in this case) and establish if it is reasonable.

c. Secondly sort descending using same method as a. and establish if the last entry is reasonable.

d. If there are any blanks in the column these will be shown at the top when sorting ascending in a.

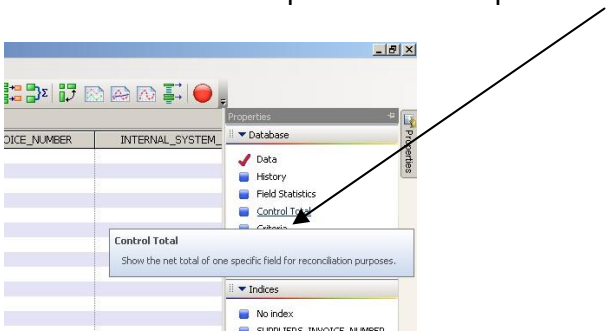
e. Repeat the methodology in b, c and d for at least the other critical fields - ‘Suppliers Invoice Number’, ‘Internal Invoice Number’, ‘Invoice Date’, ‘Payment Date’, ‘Total Invoice Amount’ and ‘VAT Amount’.

f. For date fields the earliest and latest dates should be reasonable and comply with the data specification where appropriate.

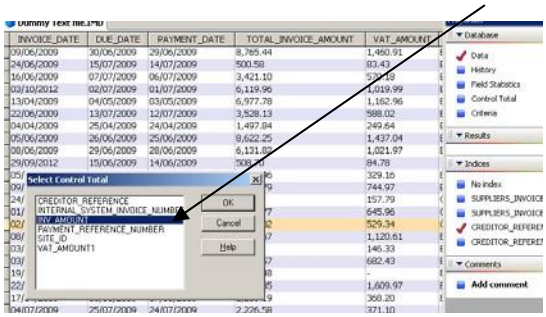
g. For amount fields the smallest (including negatives) and largest amounts should be as expected.

Control Totals

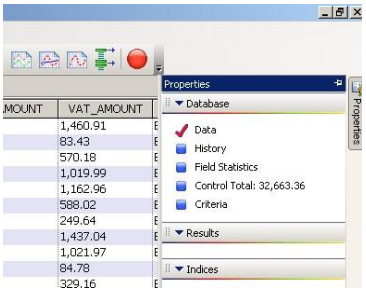
a. Select the Control Total option on the Properties tab.



b. Select each relevant field to obtain control totals.

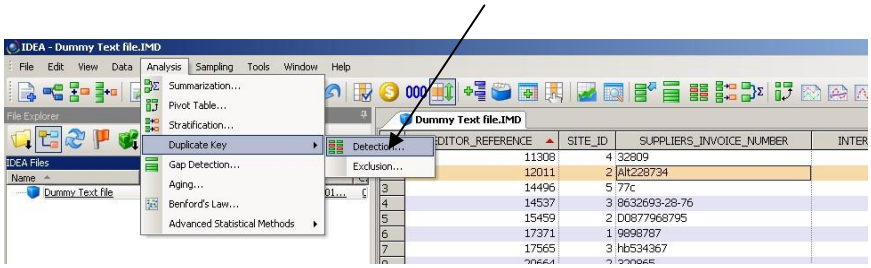


c. Note the control total of the selected field in the Properties tab.

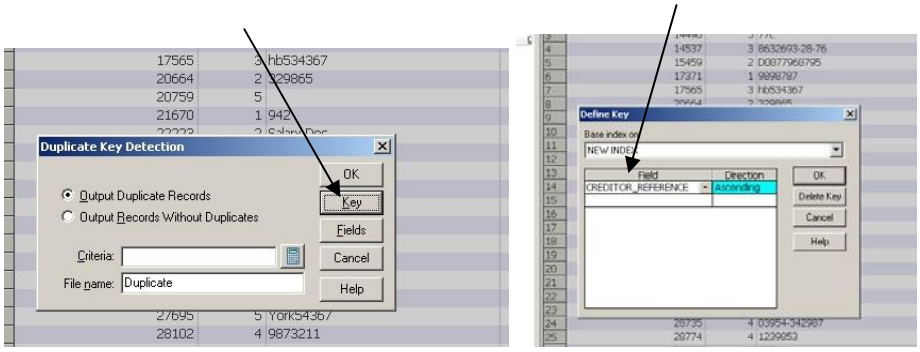


Duplicate records

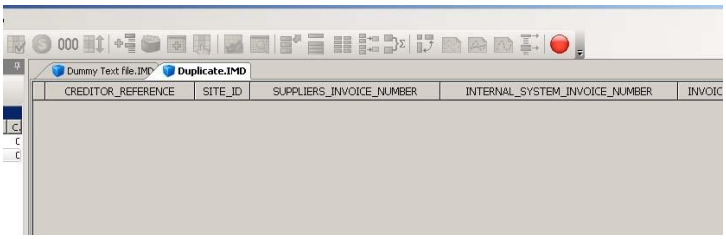
a. From the Analysis menu select Duplicate Key/Detection option.



b. In the Duplicate Key Detection box select the 'Key' button and select the appropriate field to check (in this case the Creditor Reference).



c. If no records are returned (as below) it means that there are no records with duplicate references.



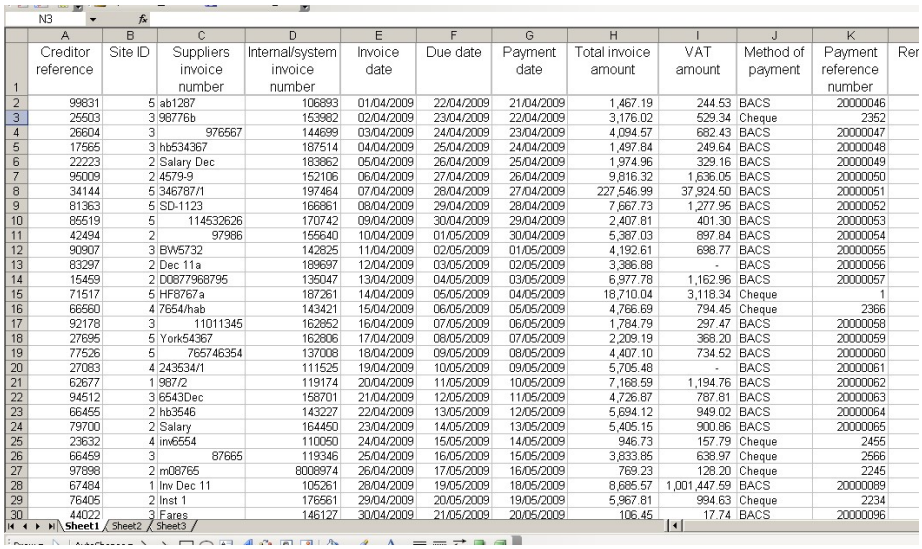
6. Where there are either unusual or unexpected entries at the extremities, there are blank entries, the control totals are unreasonable or there are duplicate references, consider whether there is enough of an issue to merit further investigation as to why the data has these features.

    a. If you are satisfied that the data is as expected and reflects the data specification then submit the original untouched data file.

    b. If the issues appear to be fundamental or extreme enough to doubt the validity or completeness of the data, you need to consider extracting the data again. If this is the case then you need to share these issues with the person responsible for the production of the data as they could have been caused by the extraction process.

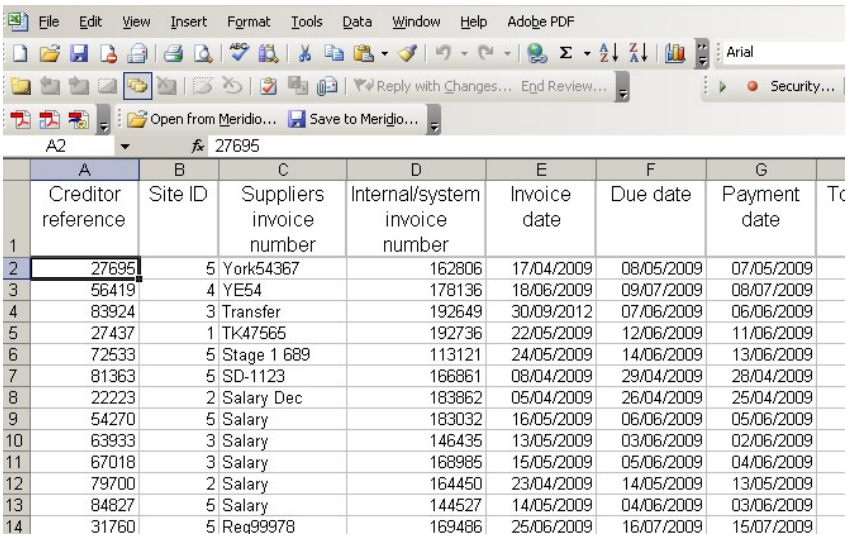
Back to top

1. It has to be assumed that the data does not have more records than the maximum shown by the Excel worksheet.
2. Create a copy of the data to be submitted to ensure that you always have the original intact.
3. If there is no header row (a) insert a row at the top of the worksheet and (b) complete cells in the first row with the field names as per specification (see example below).



	A	B	C	D	E	F	G	H	I	J	K	Rer
1	Creditor reference	Site ID	Suppliers invoice number	Internal/system invoice number	Invoice date	Due date	Payment date	Total invoice amount	VAT amount	Method of payment	Payment reference number	
2	99831	5	ab1287	106893	01/04/2009	22/04/2009	21/04/2009	1,467.19	244.53	BACS	20000046	
3	25503	3	98776b	153962	02/04/2009	23/04/2009	22/04/2009	3,176.02	529.34	Cheque	2352	
4	26604	3	976567	144699	03/04/2009	24/04/2009	23/04/2009	4,094.57	682.43	BACS	20000047	
5	17565	3	hb534367	187514	04/04/2009	25/04/2009	24/04/2009	1,497.84	249.64	BACS	20000048	
6	22223	2	Salary Dec	183862	05/04/2009	26/04/2009	25/04/2009	1,974.96	329.16	BACS	20000049	
7	95009	2	4579-9	152106	06/04/2009	27/04/2009	26/04/2009	9,816.32	1,636.05	BACS	20000050	
8	34144	5	346787/1	197464	07/04/2009	28/04/2009	27/04/2009	227,546.99	37,924.50	BACS	20000051	
9	81363	5	SD-1123	166861	08/04/2009	29/04/2009	28/04/2009	7,667.73	1,277.95	BACS	20000052	
10	85519	5	114532626	170742	09/04/2009	30/04/2009	29/04/2009	2,407.81	401.30	BACS	20000053	
11	42494	2	97986	155640	10/04/2009	01/05/2009	30/04/2009	5,387.03	897.84	BACS	20000054	
12	90907	3	BW5732	142825	11/04/2009	02/05/2009	01/05/2009	4,192.61	698.77	BACS	20000055	
13	83297	2	Dec 11a	189697	12/04/2009	03/05/2009	02/05/2009	3,386.88	-	BACS	20000056	
14	15459	2	D0877968795	135047	13/04/2009	04/05/2009	03/05/2009	6,977.78	1,162.96	BACS	20000057	
15	71517	5	HF8767a	187261	14/04/2009	05/05/2009	04/05/2009	18,710.04	3,118.34	Cheque	1	
16	66960	4	7654/hab	143421	15/04/2009	06/05/2009	05/05/2009	4,766.69	794.45	Cheque	2366	
17	92178	3	11011345	162852	16/04/2009	07/05/2009	06/05/2009	1,784.79	297.47	BACS	20000058	
18	27695	5	York54367	162806	17/04/2009	08/05/2009	07/05/2009	2,209.19	368.20	BACS	20000059	
19	77526	5	765746354	137008	18/04/2009	09/05/2009	08/05/2009	4,407.10	734.52	BACS	20000060	
20	27083	4	243534/1	111525	19/04/2009	10/05/2009	09/05/2009	5,705.48	-	BACS	20000061	
21	62677	1	987/2	119174	20/04/2009	11/05/2009	10/05/2009	7,168.59	1,194.76	BACS	20000062	
22	94512	3	6543Dec	158701	21/04/2009	12/05/2009	11/05/2009	4,726.87	787.81	BACS	20000063	
23	66455	2	hb3546	143227	22/04/2009	13/05/2009	12/05/2009	5,694.12	949.02	BACS	20000064	
24	79700	2	Salary	164450	23/04/2009	14/05/2009	13/05/2009	5,405.15	900.86	BACS	20000065	
25	23632	4	inv6554	110050	24/04/2009	15/05/2009	14/05/2009	946.73	157.79	Cheque	2455	
26	66459	3	87665	119346	25/04/2009	16/05/2009	15/05/2009	3,833.85	638.97	Cheque	2566	
27	97898	2	m08765	8008974	26/04/2009	17/05/2009	16/05/2009	769.23	128.20	Cheque	2245	
28	67484	1	Inv Dec 11	105261	28/04/2009	19/05/2009	18/05/2009	8,685.57	1,001,447.59	BACS	20000089	
29	76405	2	Inst 1	176561	29/04/2009	20/05/2009	19/05/2009	5,967.81	994.63	Cheque	2234	
30	44022	3	Fares	146127	30/04/2009	21/05/2009	20/05/2009	106.45	17.74	BACS	20000096	

4. Highlight all the data in the table – we suggest you place the cursor on the first data cell (A2) and then press shift-ctrl-end simultaneously.

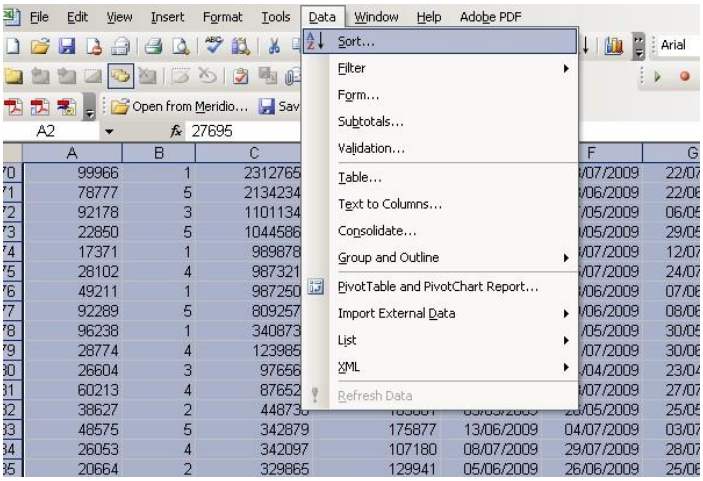


	A	B	C	D	E	F	G	To
1	Creditor reference	Site ID	Suppliers invoice number	Internal/system invoice number	Invoice date	Due date	Payment date	
2	27695	5	York54367	162806	17/04/2009	08/05/2009	07/05/2009	
3	56419	4	YE54	178136	18/06/2009	09/07/2009	08/07/2009	
4	83924	3	Transfer	192649	30/09/2012	07/06/2009	06/06/2009	
5	27437	1	TK47565	192736	22/05/2009	12/06/2009	11/06/2009	
6	72533	5	Stage 1 689	113121	24/05/2009	14/06/2009	13/06/2009	
7	81363	5	SD-1123	166861	08/04/2009	29/04/2009	28/04/2009	
8	22223	2	Salary Dec	183862	05/04/2009	26/04/2009	25/04/2009	
9	54270	5	Salary	183032	16/05/2009	06/06/2009	05/06/2009	
10	63933	3	Salary	146435	13/05/2009	03/06/2009	02/06/2009	
11	67018	3	Salary	168985	15/05/2009	05/06/2009	04/06/2009	
12	79700	2	Salary	164450	23/04/2009	14/05/2009	13/05/2009	
13	84827	5	Salary	144527	14/05/2009	04/06/2009	03/06/2009	
14	31760	5	Req99978	169486	25/06/2009	16/07/2009	15/07/2009	

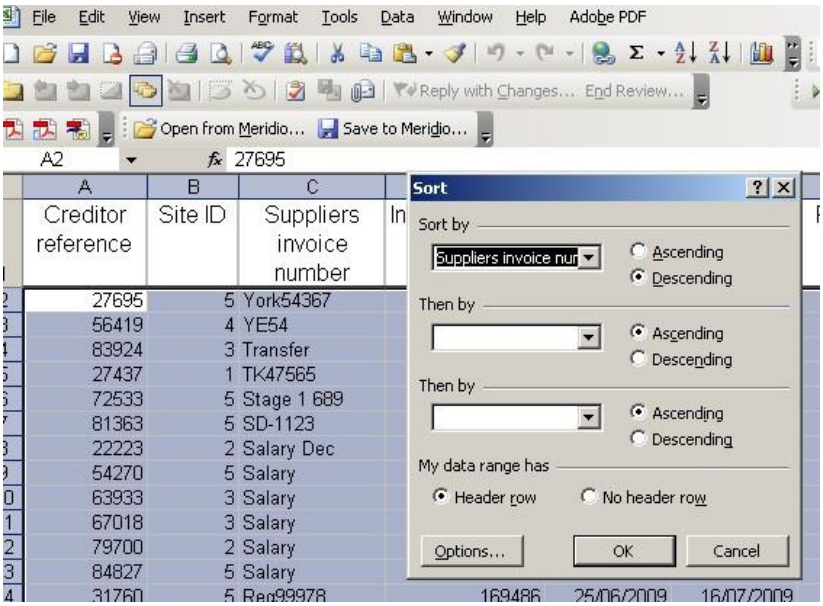
5. Carry out the following checks:

Sorting

a. Select Sort from the Data dropdown menu.



b. In the 'Sort by' dropdown box select the first field to be checked (in this case 'Suppliers invoice number) and select 'OK'.



c. Establish that the first entries are reasonable and as expected.

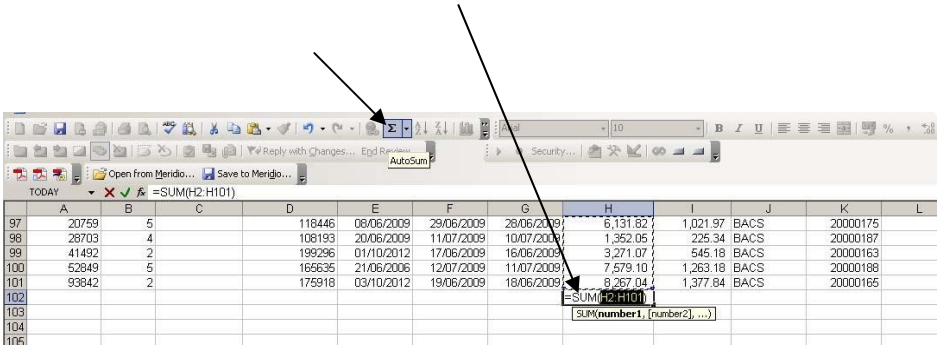
d. Repeat step a. and for the same field select the same field but changing the sort to 'Ascending'. This will enable you to establish whether the data at the other end of this field's range is reasonable. This check should also show whether there are any blank entries.

e. Repeat the methodology in a, b, c and d for at least the other critical fields in this dataset – for example, in a Creditors History file we suggest

'Suppliers Invoice Number', 'Internal Invoice Number', 'Invoice Date', 'Payment Date', 'Total Invoice Amount' and 'VAT Amount'.

Control Totals

a. Select the cell below any monetary column and activate the summarisation function.



b. This will summarise all the entries in that column, which you can then check for reasonableness.

Duplicate records

a. In Excel this is not an easy feature to look for without using fairly advanced techniques. However, it can be quite effective to sort the reference number column and then visually inspect that column as you scroll down page by page.

6. Where there are either unusual or unexpected entries at the extremities, or there are blank entries, consider whether they are reasonable or if there is enough of an issue to merit further investigation as to why the data has these features.

    a. If you are satisfied that the data is as expected and reflects the data specification then submit the original untouched data file.

    b. However, if there are what appears to be a few rogue records or entries then you may find it easier to either delete or refine these records or entries manually. The revised file can then be submitted after carrying out the same checks again.

    c. If the issues appear to be fundamental or extreme enough to doubt the validity or completeness of the data, you need to consider extracting the data again. If this is the case then you need to share these issues with the person responsible for the production of the data as they could have been caused by the extraction process.

Back to top

NFI Team Contact Details

General support

For queries related to the data submission, data matches, investigations, outcome recording or any general NFI query.

Email: nfiqueries@cabinetoffice.gov.uk

Technical queries (NFI Helpdesk)

For queries related to issues specifically with the NFI web application including access issues or user accounts.

Tel: 0845 345 8019

Email: helpdesk@nfi.gov.uk