OPA Excel Tips: Removing duplicates

When downloading data from IMPACII and QVR, the number of records will depend on the search you carry out but also the fields requested, for example you could end up with a row for each investigator on a grant, or a row for each re-submission. Depending on the question you may, or may not, want these multiple rows. For example an analysis of applications by IC might mean counting all re-submissions, whereas counting projects by IC would require duplicate project numbers to be removed. Within Excel there is a handy function to do this.

Example 1: Removing duplicate project numbers.

In this example the Transplantation dataset is used.

A quick pivot table (see Excel tip 4) shows the number of rows in the dataset by IC (Table 1) but creating a pivot table of project number and counting ApplIDs shows significant duplication (Table 2).

Table 1: Count of projects by IC

Row Labels	Count of Project
AA	9
AG	40
AI	345
AR	113
AT	1
CA	211
DC	7
DE	55
DK	307
EB	35
ES	5
EY	51
GM	31
HD	29
HG	1
HL	414
MD	6
MH	7
NR	11
NS	107
OD	8
RR	6
TR	3
Grand Total	**1802**

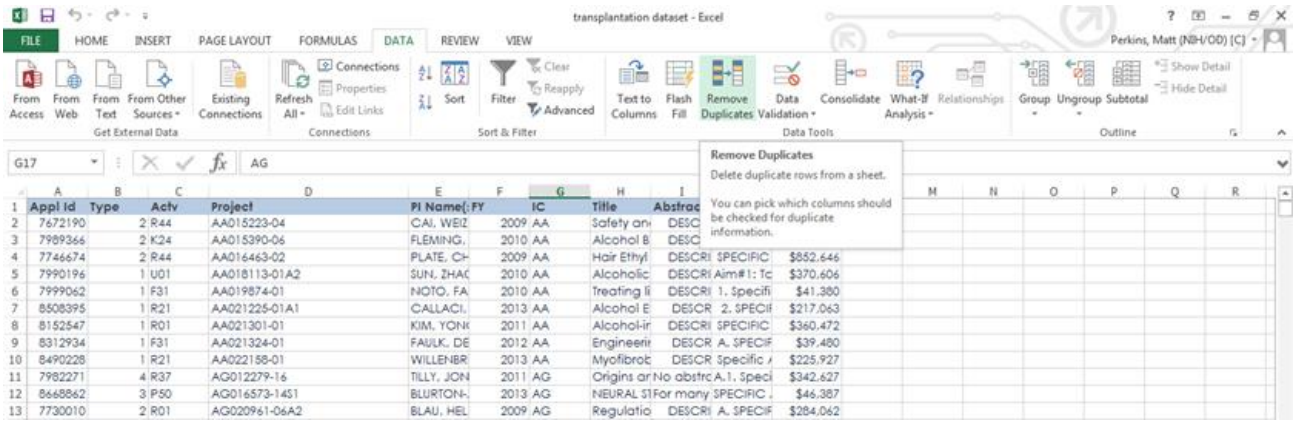
Table 2: Count of ApplIDs by Project

Row Labels	Count of Appl Id
CA023766-34	9
NS055976-06A1	8
DK013083-40A1	8
CA049605-24A1	8
HL075462-06A1	7
CA065493-16	7
AI063603-08	7
CA142106-06A1	6
CA015396-37	6
CA039542-23A1	6
CA018029-37A1	6
AI084853-01	6
AI102405-01	6
AI045897-11A1	6
AI087586-01	6
AI046629-10A1	6
AI051731-11	6
HL094374-01A1	5
CA078902-11	5
AI089556-01A1	5
HL018646-31A1	4
CA111412-06	4
AI097113-01	4

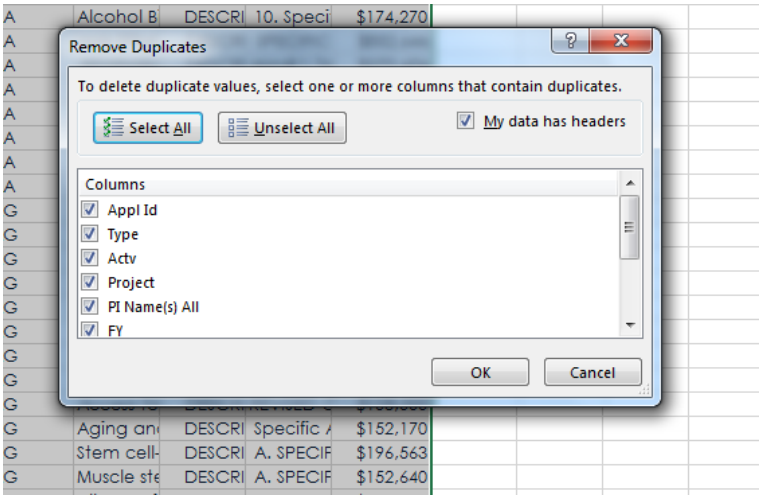
Looking at the first project number in the list (by double clicking on the count in the cell) shows:

	Appl Id	Ty	Act	Project	PI	FY	IC	Title	Abstract Text (only)	SA Text	Awd
2	8435576	2	P01	CA023766-34	HELL	2013	CA	Biostatistics Core	The Biostatistics Core (Core B) will provide bi	SPECIFICAIMS; The	2E+05
3	8435567	2	P01	CA023766-34	SCHE	2013	CA	Potentiating Anti-WT1 r	This project brings together three important,	Specific aims: This pr	5E+05
4	8435564	2	P01	CA023766-34	VAN	2013	CA	IL-22 and Mucosal Immu	There is currently little understanding of how	SPECIFIC AIMS: Donc	4E+05
5	8435574	2	P01	CA023766-34	PERA	2013	CA	Graft Characterization an	PROJECT SUMIVIARY (See instructions): Allo	Specific Aims. The p	3E+05
6	8435563	2	P01	CA023766-34	YOUI	2013	CA	Distinguishing the Affere	PrOJSCt 3 This project addresses two major	Specific Aim 1: Deter	5E+05
7	8435571	2	P01	CA023766-34	O'RE	2013	CA	Clinical Trials Evaluating T	PROJECT SUMMARY (See instructions): Proje	Objective: To condu	4E+05
8	8435577	2	P01	CA023766-34	O'RE	2013	CA	Administrative Core	C o r e C The Administrative Core (Core C) p	OBJECTIVE: The Ad	2E+05
9	8435562	2	P01	CA023766-34	PAME	2013	CA	Monocytes and the Inte	P r o j e c t 2 Monocytes provide defense ag	Monocytes provide d	3E+05
10	8435557	2	P01	CA023766-34	HSU,	2013	CA	NK Receptor Function in	P r o j e c t 1 Allogeneic hematopoietic stem	Allogeneic hematopo	3E+05

So we only want to count this project once as our original table includes 9 awards under a P01 grant in the same year. Excel has a function to do this for us called 'Remove Duplicates'. It is under the 'Data' tab.

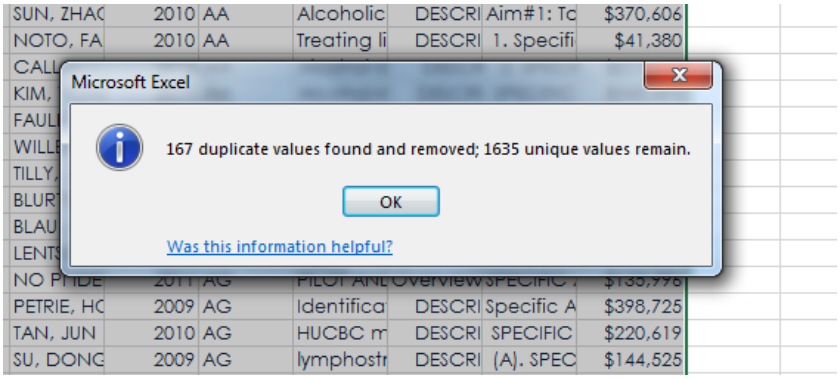


Clicking on the icon brings up the following box:



The list of columns in the spreadsheet appear in the list. Leaving all boxes selected will only remove rows that are matches in every column in the dataset. For this example we select 'Unselect All' then select the 'Project' tick box and the IC tick box (to ensure that we count any projects funded by multiple ICs in the dataset), then select 'OK'.

The following message appears:



Reproducing the IC pivot table above now gives different results. Instead of 1,802 records we only have 1,635. Looking at the first few rows, AA and AG and unchanged but the AI count has fallen to 275.

Row Labels	Count of Project
AA	9
AG	40
AI	275
AR	110
AT	1
CA	158
DC	7
DE	55
DK	299
EB	35
ES	4
EY	51
GM	30
HD	29
HG	1
HL	394
MD	4
MH	7
NR	11
NS	98
OD	8
RR	6
TR	3
Grand Total	**1635**

CAUTION: Care must be taken when removing duplicates, especially when working with large datasets where it is not possible to check all the records that are likely to be removed. Think carefully about which fields you need to select in order to ensure only duplicate records are removed. For example if the download included multiple rows per project due to funding coming from multiple ICs, you may want to tick the IC box as well, to ensure you keep a record for each IC that funded the project.

More help

Go to the ‘Remove Duplicates’ icon, select it, then in the options box select the ‘?’ in the top right corner of the box: