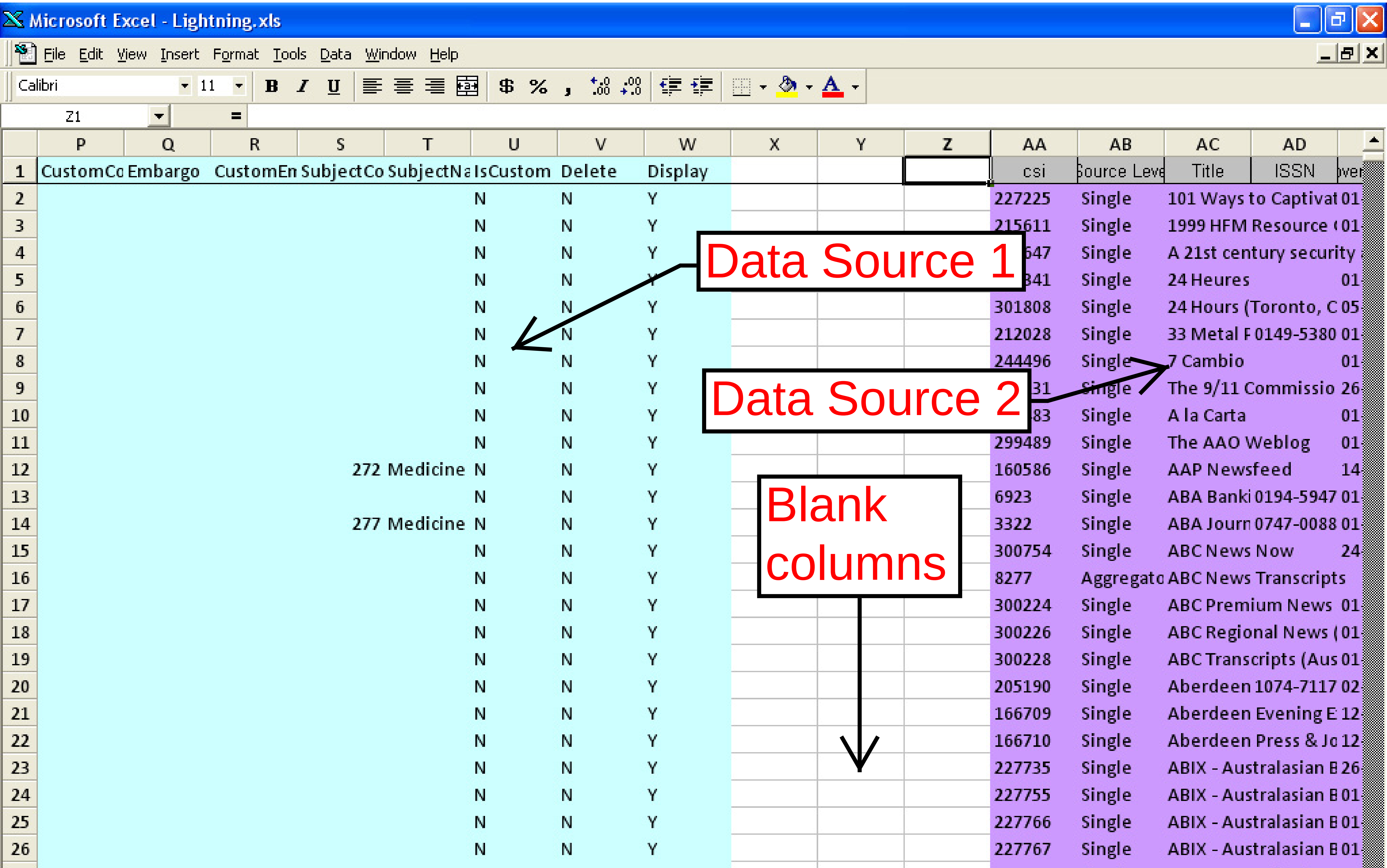
Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner (GwenExner@gmail.com)

When verifying data integrity using two sources of data, it is possible to quickly reduce the amount of data to be manually checked by using Excel or some other spreadsheet.

1) First, place both sets of data in the same worksheet, with at least one blank column between them.

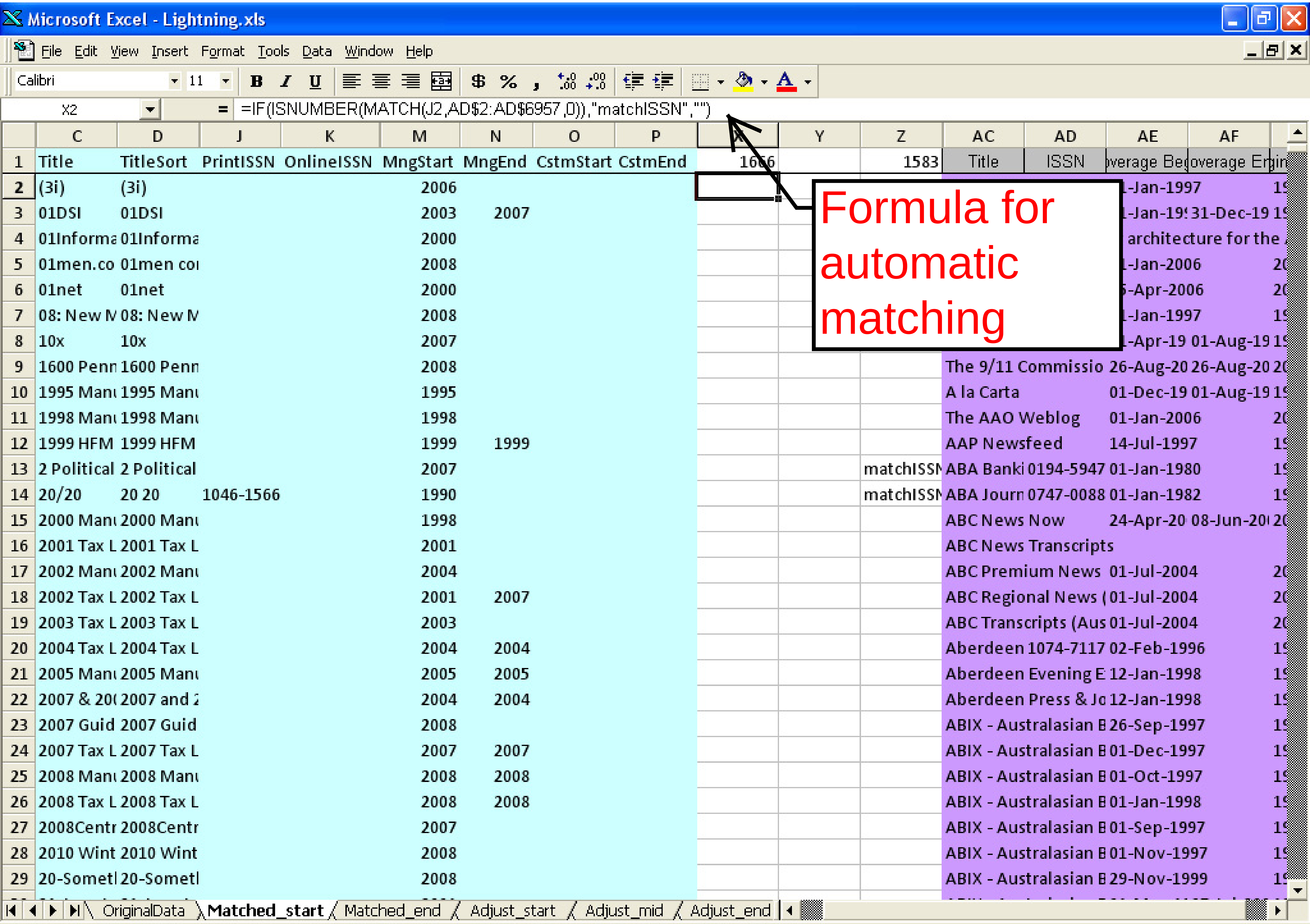

Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri 11
Z1
Data Source 1
Data Source 2
Blank columns

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	
1	CustomCo	Embargo	CustomEn	SubjectCo	SubjectNa	IsCustom	Delete	Display				csi	Source Leve	Title	ISSN	ver
2						N	N	Y				227225	Single	101 Ways to Captivat		01
3						N	N	Y				215611	Single	1999 HFM Resource (		01
4						N	N	Y				647	Single	A 21st century security		
5						N	N	Y				341	Single	24 Heures		01
6						N	N	Y				301808	Single	24 Hours (Toronto, C		05
7						N	N	Y				212028	Single	33 Metal F	0149-5380	01
8						N	N	Y				244496	Single	7 Cambio		01
9						N	N	Y				31	Single	The 9/11 Commissio		26
10						N	N	Y				83	Single	A la Carta		01
11						N	N	Y				299489	Single	The AAO Weblog		01
12				272	Medicine	N	N	Y				160586	Single	AAP Newsfeed		14
13						N	N	Y				6923	Single	ABA Banki	0194-5947	01
14				277	Medicine	N	N	Y				3322	Single	ABA Journ	0747-0088	01
15						N	N	Y				300754	Single	ABC News Now		24
16						N	N	Y				8277	Aggregato	ABC News Transcripts		
17						N	N	Y				300224	Single	ABC Premium News		01
18						N	N	Y				300226	Single	ABC Regional News (		01
19						N	N	Y				300228	Single	ABC Transcripts (Aus		01
20						N	N	Y				205190	Single	Aberdeen	1074-7117	02
21						N	N	Y				166709	Single	Aberdeen Evening E		12
22						N	N	Y				166710	Single	Aberdeen Press & Jo		12
23						N	N	Y				227735	Single	ABIX - Australasian B		26
24						N	N	Y				227755	Single	ABIX - Australasian B		01
25						N	N	Y				227766	Single	ABIX - Australasian B		01
26						N	N	Y				227767	Single	ABIX - Australasian B		01

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

2) Tell the computer to find matches.


Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri
11
OriginalData
Matched_start
Matched_end
Adjust_start
Adjust_mid
Adjust_end

X2 =IF(ISNUMBER(MATCH(J2,AD$2:AD$6957,0)),"matchISSN","")

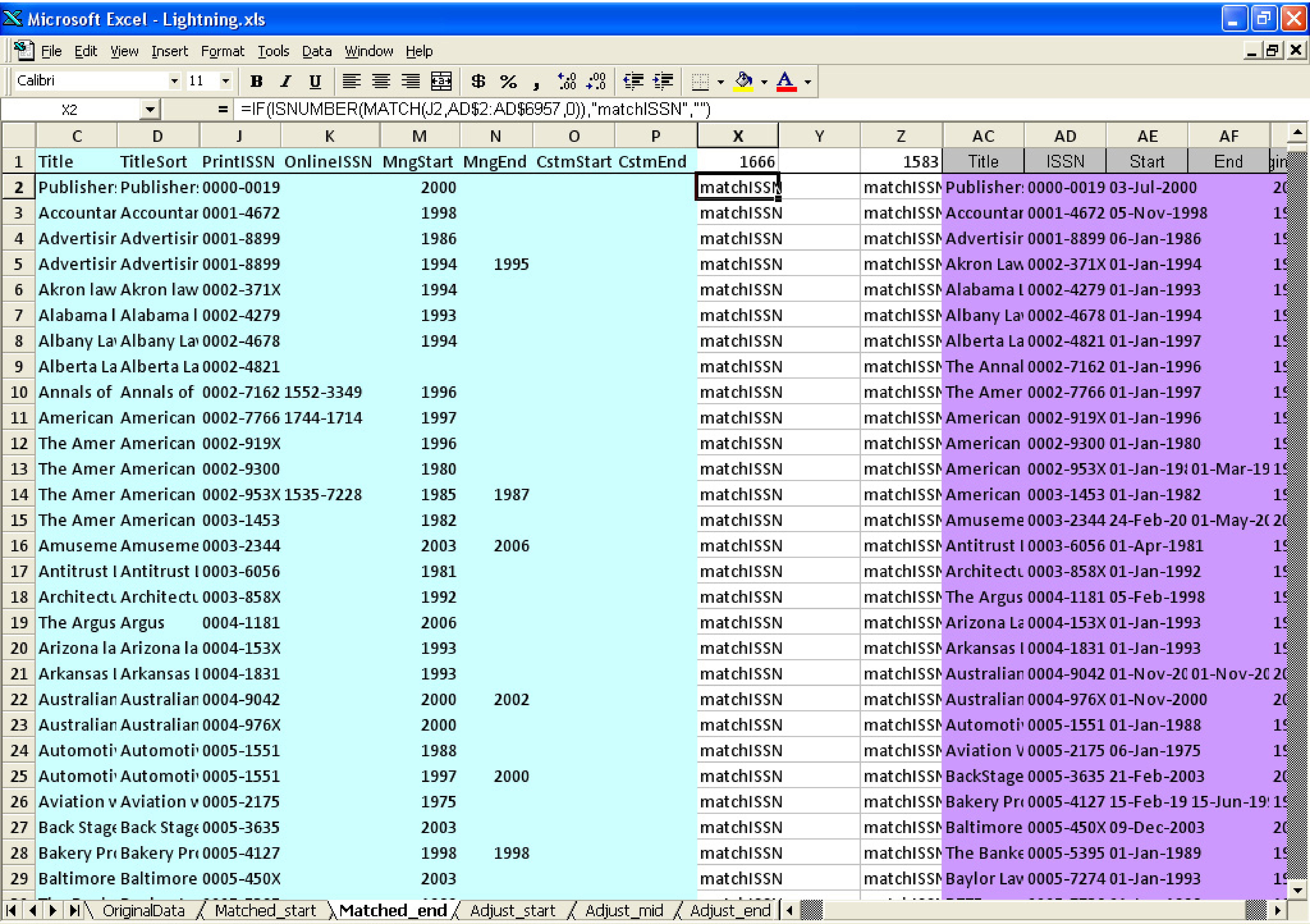
Formula for automatic matching

	C	D	J	K	M	N	O	P	X	Y	Z	AC	AD	AE	AF	
1	Title	TitleSort	PrintISSN	OnlineISSN	MngStart	MngEnd	CstmStart	CstmEnd	1666		1583	Title	ISSN	verage Be	overage Er	gin
2	(3i)	(3i)			2006									1-Jan-1997		19
3	01DSI	01DSI			2003	2007								1-Jan-19!	31-Dec-19	19
4	01Informa	01Informa			2000									architecture for the		
5	01men.co	01men coı			2008									1-Jan-2006		20
6	01net	01net			2000									5-Apr-2006		20
7	08: New M	08: New M			2008									1-Jan-1997		19
8	10x	10x			2007									1-Apr-19	01-Aug-19	19
9	1600 Penn	1600 Penn			2008							The 9/11 Commissio		26-Aug-20	26-Aug-20	20
10	1995 Manı	1995 Manı			1995							A la Carta		01-Dec-19	01-Aug-19	19
11	1998 Manı	1998 Manı			1998							The AAO Weblog		01-Jan-2006		20
12	1999 HFM	1999 HFM			1999	1999						AAP Newsfeed		14-Jul-1997		19
13	2 Political	2 Political			2007						matchISSN	ABA Banki	0194-5947	01-Jan-1980		19
14	20/20	20 20	1046-1566		1990						matchISSN	ABA Journ	0747-0088	01-Jan-1982		19
15	2000 Manı	2000 Manı			1998							ABC News Now		24-Apr-20	08-Jun-20	20
16	2001 Tax L	2001 Tax L			2001							ABC News Transcripts				
17	2002 Manı	2002 Manı			2004							ABC Premium News		01-Jul-2004		20
18	2002 Tax L	2002 Tax L			2001	2007						ABC Regional News (		01-Jul-2004		20
19	2003 Tax L	2003 Tax L			2003							ABC Transcripts (Aus		01-Jul-2004		20
20	2004 Tax L	2004 Tax L			2004	2004						Aberdeen	1074-7117	02-Feb-1996		19
21	2005 Manı	2005 Manı			2005	2005						Aberdeen Evening E		12-Jan-1998		19
22	2007 & 20(	2007 and 2			2004	2004						Aberdeen Press & Jo		12-Jan-1998		19
23	2007 Guid	2007 Guid			2008							ABIX - Australasian B		26-Sep-1997		19
24	2007 Tax L	2007 Tax L			2007	2007						ABIX - Australasian B		01-Dec-1997		19
25	2008 Manı	2008 Manı			2008	2008						ABIX - Australasian B		01-Oct-1997		19
26	2008 Tax L	2008 Tax L			2008	2008						ABIX - Australasian B		01-Jan-1998		19
27	2008Centr	2008Centr			2007							ABIX - Australasian B		01-Sep-1997		19
28	2010 Wint	2010 Wint			2008							ABIX - Australasian B		01-Nov-1997		19
29	20-Somet	20-Somet			2008							ABIX - Australasian B		29-Nov-1999		19

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

3) Sort the matches to the top. (There will almost certainly be unmatched data at the bottom.)


Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri 11
B I U
$ % ,

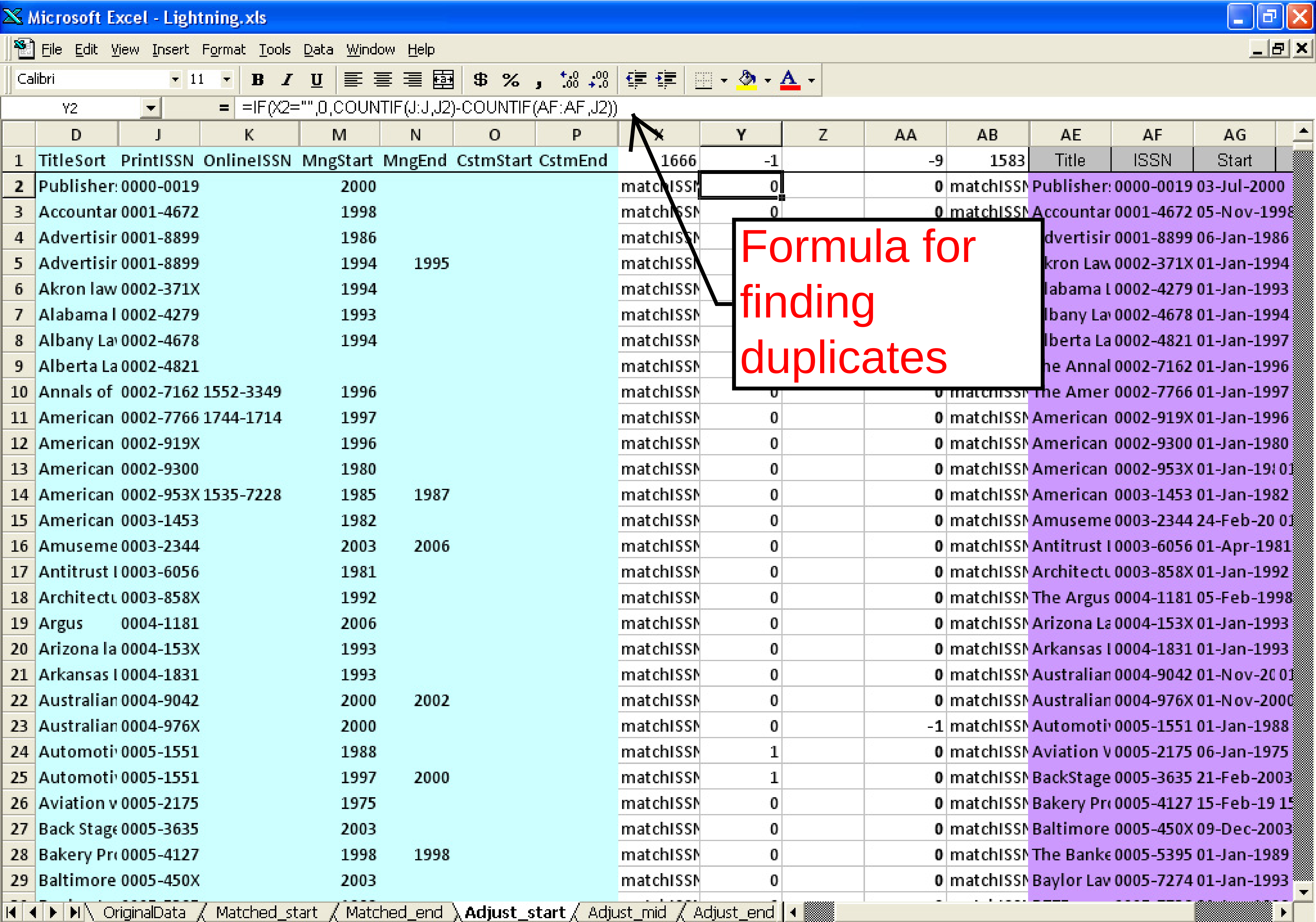
X2 = =IF(ISNUMBER(MATCH(J2,AD$2:AD$6957,0)),"matchISSN","")

	C	D	J	K	M	N	O	P	X	Y	Z	AC	AD	AE	AF	
1	Title	TitleSort	PrintISSN	OnlineISSN	MngStart	MngEnd	CstmStart	CstmEnd	1666		1583	Title	ISSN	Start	End	gin
2	Publisher:	Publisher:	0000-0019		2000				matchISSN		matchISSN	Publisher:	0000-0019	03-Jul-2000		20
3	Accountar	Accountar	0001-4672		1998				matchISSN		matchISSN	Accountar	0001-4672	05-Nov-1998		19
4	Advertisir	Advertisir	0001-8899		1986				matchISSN		matchISSN	Advertisir	0001-8899	06-Jan-1986		19
5	Advertisir	Advertisir	0001-8899		1994	1995			matchISSN		matchISSN	Akron Law	0002-371X	01-Jan-1994		19
6	Akron law	Akron law	0002-371X		1994				matchISSN		matchISSN	Alabama l	0002-4279	01-Jan-1993		19
7	Alabama l	Alabama l	0002-4279		1993				matchISSN		matchISSN	Albany La	0002-4678	01-Jan-1994		19
8	Albany La	Albany La	0002-4678		1994				matchISSN		matchISSN	Alberta La	0002-4821	01-Jan-1997		19
9	Alberta La	Alberta La	0002-4821						matchISSN		matchISSN	The Annal	0002-7162	01-Jan-1996		19
10	Annals of	Annals of	0002-7162	1552-3349	1996				matchISSN		matchISSN	The Amer	0002-7766	01-Jan-1997		19
11	American	American	0002-7766	1744-1714	1997				matchISSN		matchISSN	American	0002-919X	01-Jan-1996		19
12	The Amer	American	0002-919X		1996				matchISSN		matchISSN	American	0002-9300	01-Jan-1980		19
13	The Amer	American	0002-9300		1980				matchISSN		matchISSN	American	0002-953X	01-Jan-19	01-Mar-19	19
14	The Amer	American	0002-953X	1535-7228	1985	1987			matchISSN		matchISSN	American	0003-1453	01-Jan-1982		19
15	The Amer	American	0003-1453		1982				matchISSN		matchISSN	Amuseme	0003-2344	24-Feb-20	01-May-2	20
16	Amuseme	Amuseme	0003-2344		2003	2006			matchISSN		matchISSN	Antitrust l	0003-6056	01-Apr-1981		19
17	Antitrust l	Antitrust l	0003-6056		1981				matchISSN		matchISSN	Architectu	0003-858X	01-Jan-1992		19
18	Architectu	Architectu	0003-858X		1992				matchISSN		matchISSN	The Argus	0004-1181	05-Feb-1998		19
19	The Argus	Argus	0004-1181		2006				matchISSN		matchISSN	Arizona La	0004-153X	01-Jan-1993		19
20	Arizona la	Arizona la	0004-153X		1993				matchISSN		matchISSN	Arkansas l	0004-1831	01-Jan-1993		19
21	Arkansas l	Arkansas l	0004-1831		1993				matchISSN		matchISSN	Australian	0004-9042	01-Nov-20	01-Nov-20	20
22	Australian	Australian	0004-9042		2000	2002			matchISSN		matchISSN	Australian	0004-976X	01-Nov-2000		20
23	Australian	Australian	0004-976X		2000				matchISSN		matchISSN	Automoti	0005-1551	01-Jan-1988		19
24	Automoti	Automoti	0005-1551		1988				matchISSN		matchISSN	Aviation V	0005-2175	06-Jan-1975		19
25	Automoti	Automoti	0005-1551		1997	2000			matchISSN		matchISSN	BackStage	0005-3635	21-Feb-2003		20
26	Aviation v	Aviation v	0005-2175		1975				matchISSN		matchISSN	Bakery Pr	0005-4127	15-Feb-19	15-Jun-19	19
27	Back Stag	Back Stag	0005-3635		2003				matchISSN		matchISSN	Baltimore	0005-450X	09-Dec-2003		20
28	Bakery Pr	Bakery Pr	0005-4127		1998	1998			matchISSN		matchISSN	The Banke	0005-5395	01-Jan-1989		19
29	Baltimore	Baltimore	0005-450X		2003				matchISSN		matchISSN	Baylor Lav	0005-7274	01-Jan-1993		19

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

4) Find out which matches are excessive/unmatched duplicates. If the result is positive then that data in that set is an excessive duplicate. If the result is negative then that data is duplicated in the other set, but not this one.


Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri
11
Y2
OriginalData
Matched_start
Matched_end
Adjust_start
Adjust_mid
Adjust_end

=IF(X2="",0,COUNTIF(J:J,J2)-COUNTIF(AF:AF,J2))

	D	J	K	M	N	O	P	X	Y	Z	AA	AB	AE	AF	AG
1	TitleSort	PrintISSN	OnlineISSN	MngStart	MngEnd	CstmStart	CstmEnd	1666	-1		-9	1583	Title	ISSN	Start
2	Publisher:	0000-0019		2000				matchISSN	0		0	matchISSN	Publisher:	0000-0019	03-Jul-2000
3	Accountar	0001-4672		1998				matchISSN	0		0	matchISSN	Accountar	0001-4672	05-Nov-1998
4	Advertisir	0001-8899		1986				matchISSN					dvertisir	0001-8899	06-Jan-1986
5	Advertisir	0001-8899		1994	1995			matchISSN					kron Law	0002-371X	01-Jan-1994
6	Akron law	0002-371X		1994				matchISSN					labama L	0002-4279	01-Jan-1993
7	Alabama l	0002-4279		1993				matchISSN					lbany La	0002-4678	01-Jan-1994
8	Albany La	0002-4678		1994				matchISSN					lberta La	0002-4821	01-Jan-1997
9	Alberta La	0002-4821						matchISSN					he Annal	0002-7162	01-Jan-1996
10	Annals of	0002-7162	1552-3349	1996				matchISSN	0		0	matchISSN	The Amer	0002-7766	01-Jan-1997
11	American	0002-7766	1744-1714	1997				matchISSN	0		0	matchISSN	American	0002-919X	01-Jan-1996
12	American	0002-919X		1996				matchISSN	0		0	matchISSN	American	0002-9300	01-Jan-1980
13	American	0002-9300		1980				matchISSN	0		0	matchISSN	American	0002-953X	01-Jan-19! 01
14	American	0002-953X	1535-7228	1985	1987			matchISSN	0		0	matchISSN	American	0003-1453	01-Jan-1982
15	American	0003-1453		1982				matchISSN	0		0	matchISSN	Amuseme	0003-2344	24-Feb-20 01
16	Amuseme	0003-2344		2003	2006			matchISSN	0		0	matchISSN	Antitrust l	0003-6056	01-Apr-1981
17	Antitrust l	0003-6056		1981				matchISSN	0		0	matchISSN	Architectu	0003-858X	01-Jan-1992
18	Architectu	0003-858X		1992				matchISSN	0		0	matchISSN	The Argus	0004-1181	05-Feb-1998
19	Argus	0004-1181		2006				matchISSN	0		0	matchISSN	Arizona La	0004-153X	01-Jan-1993
20	Arizona la	0004-153X		1993				matchISSN	0		0	matchISSN	Arkansas l	0004-1831	01-Jan-1993
21	Arkansas l	0004-1831		1993				matchISSN	0		0	matchISSN	Australian	0004-9042	01-Nov-20 01
22	Australian	0004-9042		2000	2002			matchISSN	0		0	matchISSN	Australian	0004-976X	01-Nov-2000
23	Australian	0004-976X		2000				matchISSN	0		-1	matchISSN	Automoti	0005-1551	01-Jan-1988
24	Automoti	0005-1551		1988				matchISSN	1		0	matchISSN	Aviation V	0005-2175	06-Jan-1975
25	Automoti	0005-1551		1997	2000			matchISSN	1		0	matchISSN	BackStage	0005-3635	21-Feb-2003
26	Aviation v	0005-2175		1975				matchISSN	0		0	matchISSN	Bakery Pr	0005-4127	15-Feb-19 15
27	Back Stag	0005-3635		2003				matchISSN	0		0	matchISSN	Baltimore	0005-450X	09-Dec-2003
28	Bakery Pr	0005-4127		1998	1998			matchISSN	0		0	matchISSN	The Banke	0005-5395	01-Jan-1989
29	Baltimore	0005-450X		2003				matchISSN	0		0	matchISSN	Baylor Lav	0005-7274	01-Jan-1993

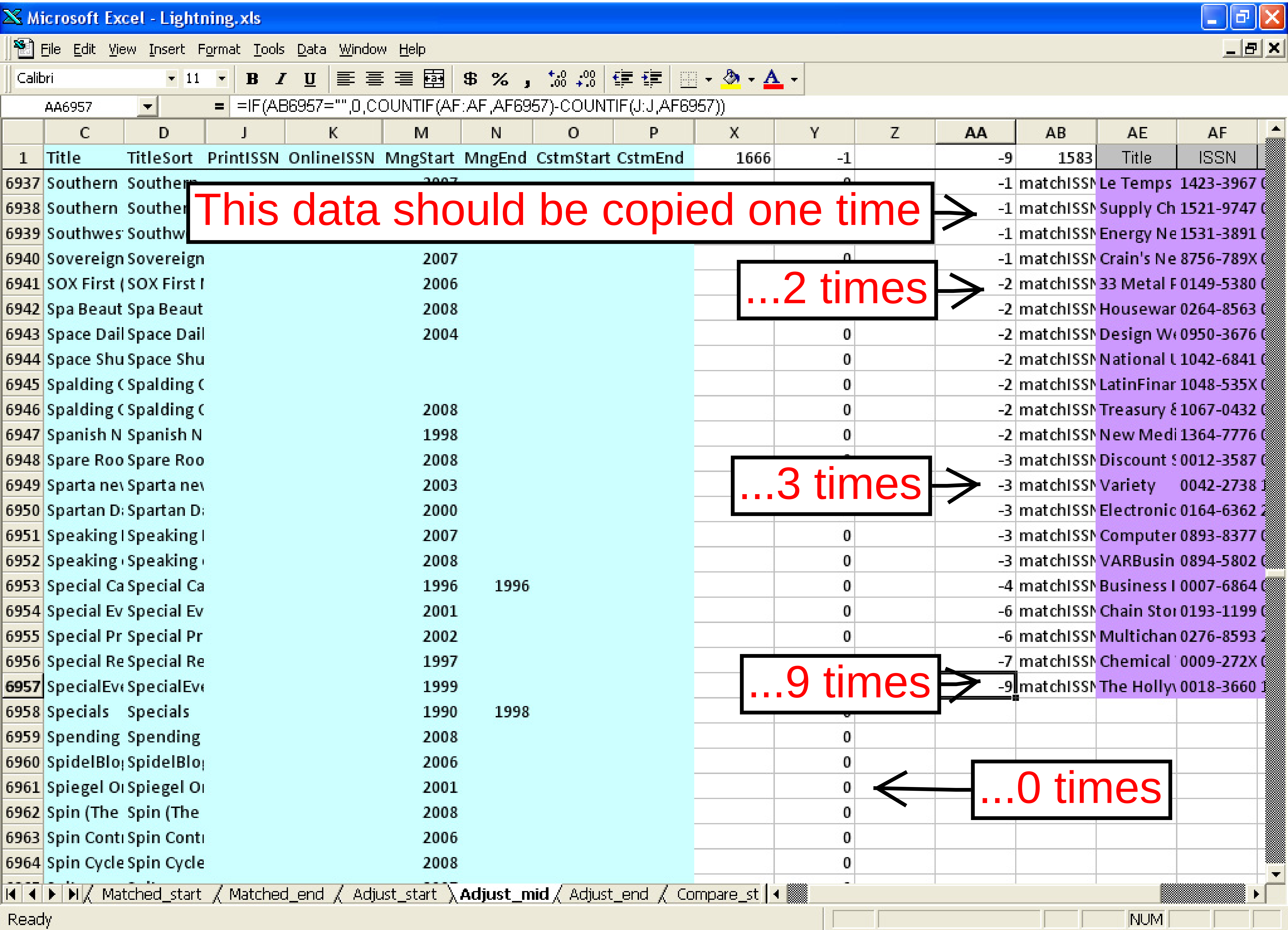
Formula for finding duplicates

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

5) Sort the data so that the ones with the most duplicates in the other set are at the bottom.

6) Duplicate the data in this set until there are no more negative numbers.



Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri
11
Ready
NUM

AA6957 =IF(AB6957="",0,COUNTIF(AF:AF,AF6957)-COUNTIF(J:J,AF6957))

	C	D	J	K	M	N	O	P	X	Y	Z	AA	AB	AE	AF
1	Title	TitleSort	PrintISSN	OnlineISSN	MngStart	MngEnd	CstmStart	CstmEnd	1666	-1		-9	1583	Title	ISSN
6937	Southern	Southern			2007					0		-1	matchISSN	Le Temps	1423-3967
6938	Southern	Souther										-1	matchISSN	Supply Ch	1521-9747
6939	Southwes	Southw										-1	matchISSN	Energy Ne	1531-3891
6940	Sovereign	Sovereign			2007					0		-1	matchISSN	Crain's Ne	8756-789X
6941	SOX First (	SOX First I			2006							-2	matchISSN	33 Metal F	0149-5380
6942	Spa Beaut	Spa Beaut			2008							-2	matchISSN	Housewar	0264-8563
6943	Space Dail	Space Dail			2004					0		-2	matchISSN	Design W	0950-3676
6944	Space Shu	Space Shu								0		-2	matchISSN	National U	1042-6841
6945	Spalding (	Spalding (								0		-2	matchISSN	LatinFinar	1048-535X
6946	Spalding (	Spalding (			2008					0		-2	matchISSN	Treasury &	1067-0432
6947	Spanish N	Spanish N			1998					0		-2	matchISSN	New Medi	1364-7776
6948	Spare Roo	Spare Roo			2008					0		-3	matchISSN	Discount S	0012-3587
6949	Sparta nev	Sparta nev			2003							-3	matchISSN	Variety	0042-2738
6950	Spartan D:	Spartan D:			2000							-3	matchISSN	Electronic	0164-6362
6951	Speaking I	Speaking I			2007					0		-3	matchISSN	Computer	0893-8377
6952	Speaking	Speaking			2008					0		-3	matchISSN	VARBusin	0894-5802
6953	Special Ca	Special Ca			1996	1996				0		-4	matchISSN	Business I	0007-6864
6954	Special Ev	Special Ev			2001					0		-6	matchISSN	Chain Stor	0193-1199
6955	Special Pr	Special Pr			2002					0		-6	matchISSN	Multichan	0276-8593
6956	Special Re	Special Re			1997							-7	matchISSN	Chemical	0009-272X
6957	SpecialEve	SpecialEve			1999							-9	matchISSN	The Holly	0018-3660
6958	Specials	Specials			1990	1998				0					
6959	Spending	Spending			2008					0					
6960	SpidelBlo	SpidelBlo			2006					0					
6961	Spiegel Oı	Spiegel Oı			2001					0					
6962	Spin (The	Spin (The			2008					0					
6963	Spin Conti	Spin Conti			2006					0					
6964	Spin Cycle	Spin Cycle			2008					0					

Matched_start / Matched_end / Adjust_start / **Adjust_mid** / Adjust_end / Compare_st

This data should be copied one time

...2 times

...3 times

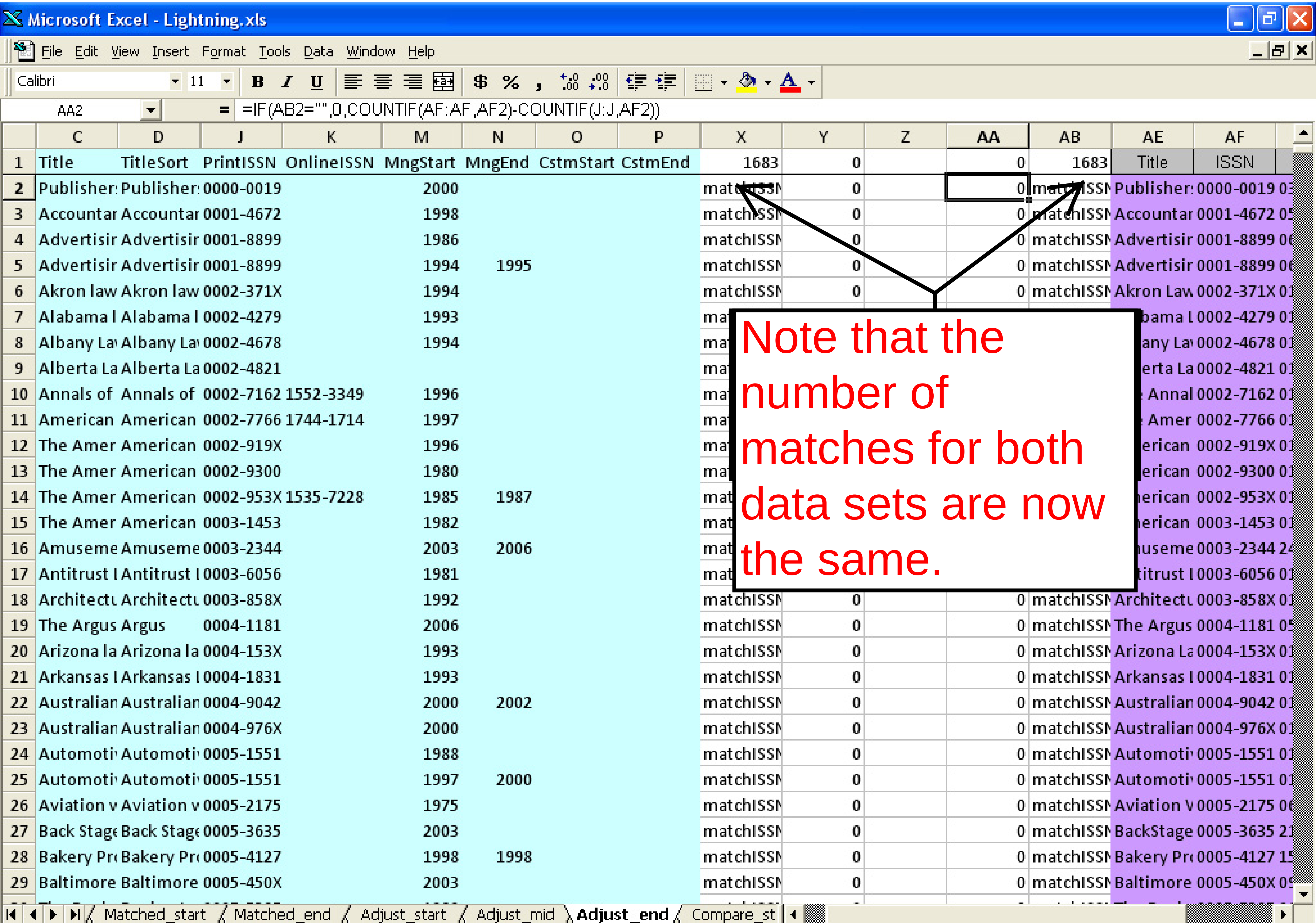
...9 times

...0 times

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

7) Sort the data by the chosen match element (in this case ISSN)



Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri 11
AA2
Matched_start / Matched_end / Adjust_start / Adjust_mid / Adjust_end / Compare_st

AA2 =IF(AB2="",0,COUNTIF(AF:AF,AF2)-COUNTIF(J:J,AF2))

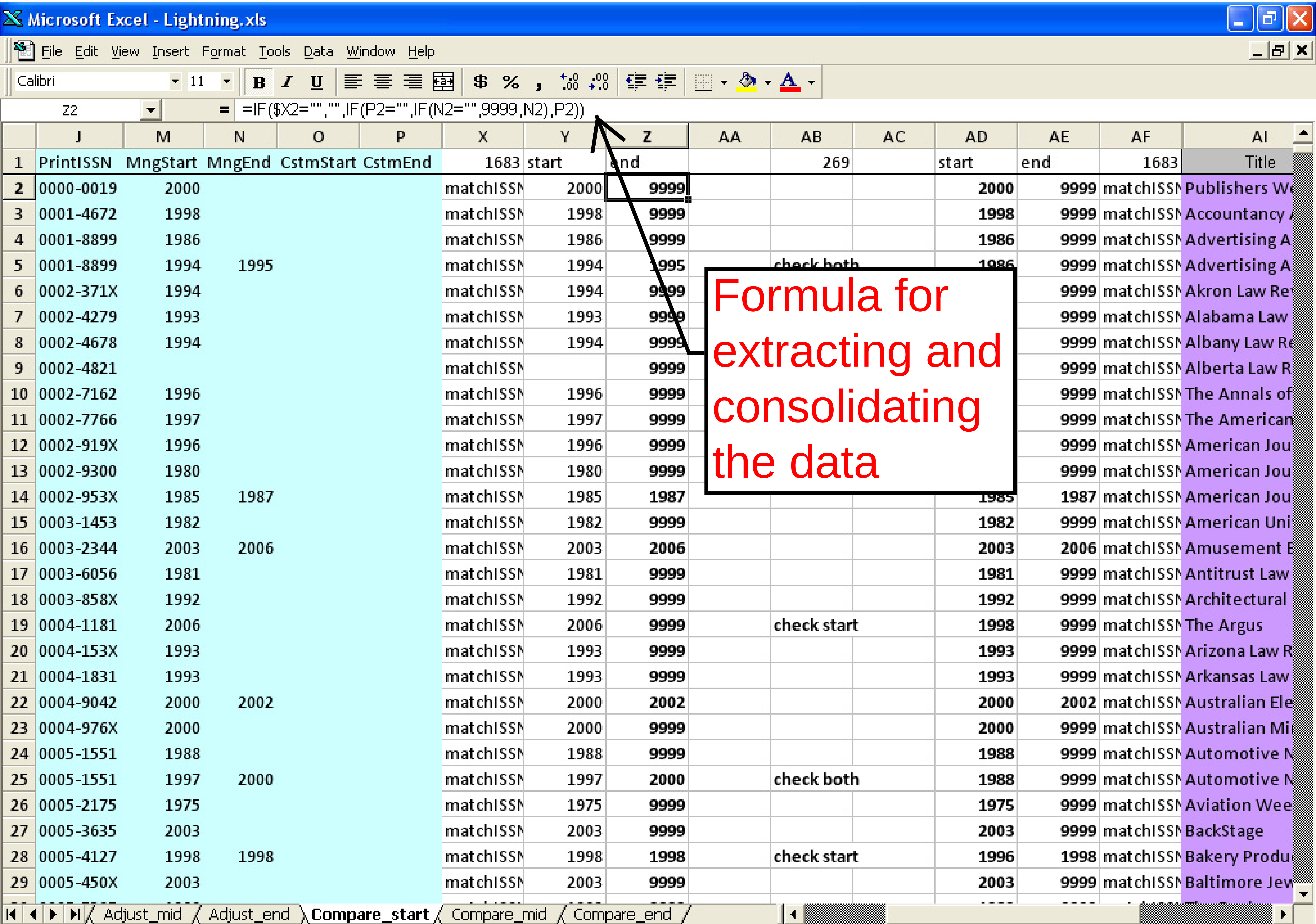
	C	D	J	K	M	N	O	P	X	Y	Z	AA	AB	AE	AF
1	Title	TitleSort	PrintISSN	OnlineISSN	MngStart	MngEnd	CstmStart	CstmEnd	1683	0		0	1683	Title	ISSN
2	Publisher:	Publisher:	0000-0019		2000				matchISSN	0		0	matchISSN	Publisher:	0000-0019 03
3	Accountar	Accountar	0001-4672		1998				matchISSN	0		0	matchISSN	Accountar	0001-4672 05
4	Advertisir	Advertisir	0001-8899		1986				matchISSN	0		0	matchISSN	Advertisir	0001-8899 06
5	Advertisir	Advertisir	0001-8899		1994	1995			matchISSN	0		0	matchISSN	Advertisir	0001-8899 06
6	Akron law	Akron law	0002-371X		1994				matchISSN	0		0	matchISSN	Akron Law	0002-371X 01
7	Alabama l	Alabama l	0002-4279		1993				mat					bama L	0002-4279 01
8	Albany La	Albany La	0002-4678		1994				mat					any La	0002-4678 01
9	Alberta La	Alberta La	0002-4821						mat					erta La	0002-4821 01
10	Annals of	Annals of	0002-7162	1552-3349	1996				mat					Annal	0002-7162 01
11	American	American	0002-7766	1744-1714	1997				mat					Amer	0002-7766 01
12	The Amer	American	0002-919X		1996				mat					erican	0002-919X 01
13	The Amer	American	0002-9300		1980				mat					erican	0002-9300 01
14	The Amer	American	0002-953X	1535-7228	1985	1987			mat					erican	0002-953X 01
15	The Amer	American	0003-1453		1982				mat					erican	0003-1453 01
16	Amuseme	Amuseme	0003-2344		2003	2006			mat					useme	0003-2344 24
17	Antitrust l	Antitrust l	0003-6056		1981				mat					titrust l	0003-6056 01
18	Architectu	Architectu	0003-858X		1992				matchISSN	0		0	matchISSN	Architectu	0003-858X 01
19	The Argus	Argus	0004-1181		2006				matchISSN	0		0	matchISSN	The Argus	0004-1181 05
20	Arizona la	Arizona la	0004-153X		1993				matchISSN	0		0	matchISSN	Arizona La	0004-153X 01
21	Arkansas l	Arkansas l	0004-1831		1993				matchISSN	0		0	matchISSN	Arkansas l	0004-1831 01
22	Australian	Australian	0004-9042		2000	2002			matchISSN	0		0	matchISSN	Australian	0004-9042 01
23	Australian	Australian	0004-976X		2000				matchISSN	0		0	matchISSN	Australian	0004-976X 01
24	Automoti	Automoti	0005-1551		1988				matchISSN	0		0	matchISSN	Automoti	0005-1551 01
25	Automoti	Automoti	0005-1551		1997	2000			matchISSN	0		0	matchISSN	Automoti	0005-1551 01
26	Aviation v	Aviation v	0005-2175		1975				matchISSN	0		0	matchISSN	Aviation V	0005-2175 06
27	Back Stage	Back Stage	0005-3635		2003				matchISSN	0		0	matchISSN	BackStage	0005-3635 21
28	Bakery Pr	Bakery Pr	0005-4127		1998	1998			matchISSN	0		0	matchISSN	Bakery Pr	0005-4127 15
29	Baltimore	Baltimore	0005-450X		2003				matchISSN	0		0	matchISSN	Baltimore	0005-450X 05

Note that the number of matches for both data sets are now the same.

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

8) Extract and consolidate the data to be checked, to make the formats match.



Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri 11

Z2 = =IF($X2="","",IF(P2="",IF(N2="",9999,N2),P2))

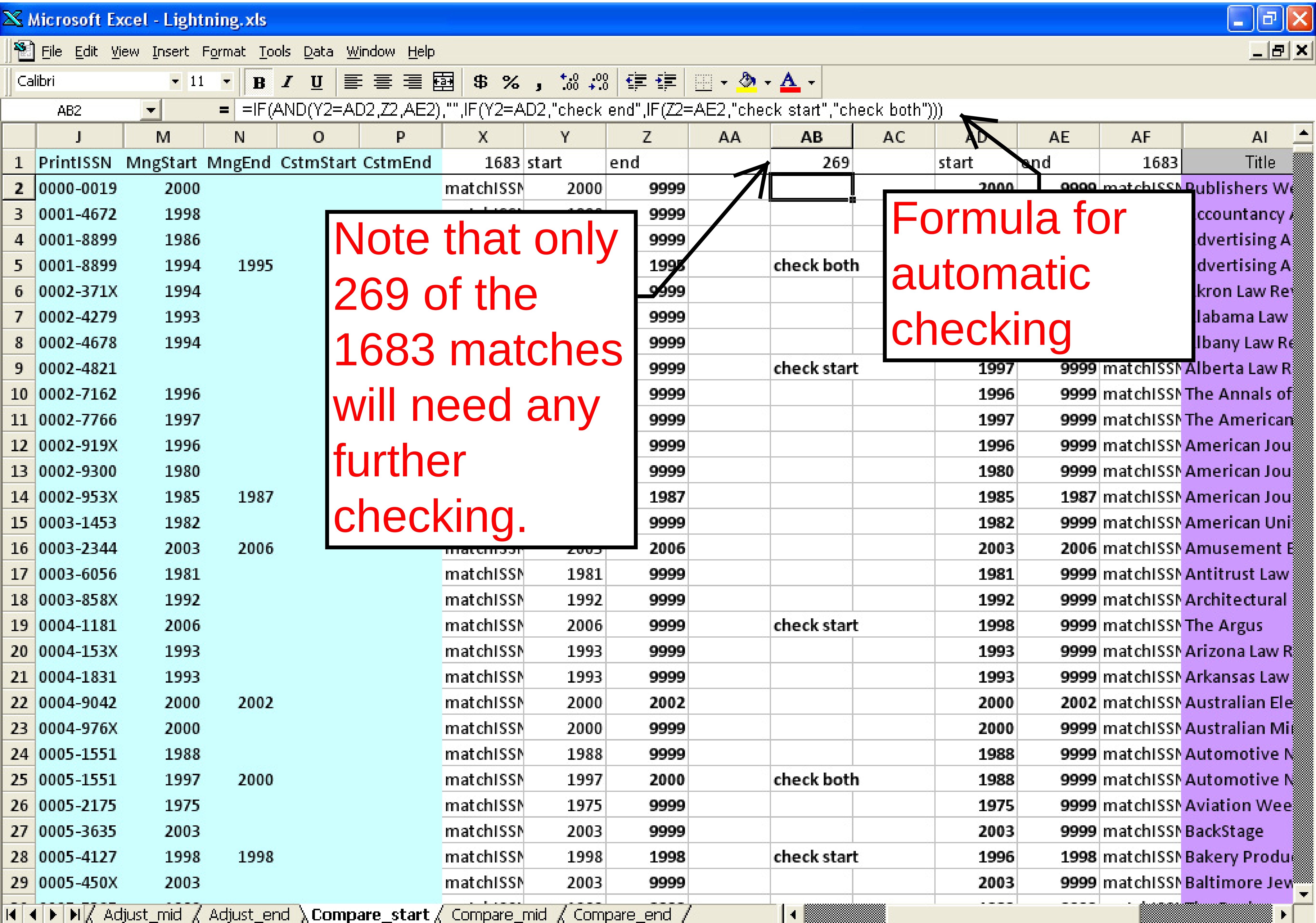
	J	M	N	O	P	X	Y	Z	AA	AB	AC	AD	AE	AF	AI
1	PrintISSN	MngStart	MngEnd	CstmStart	CstmEnd	1683	start	end		269		start	end	1683	Title
2	0000-0019	2000				matchISSN	2000	9999				2000	9999	matchISSN	Publishers We
3	0001-4672	1998				matchISSN	1998	9999				1998	9999	matchISSN	Accountancy A
4	0001-8899	1986				matchISSN	1986	9999				1986	9999	matchISSN	Advertising A
5	0001-8899	1994	1995			matchISSN	1994	1995		check both		1986	9999	matchISSN	Advertising A
6	0002-371X	1994				matchISSN	1994	9999					9999	matchISSN	Akron Law Re
7	0002-4279	1993				matchISSN	1993	9999					9999	matchISSN	Alabama Law
8	0002-4678	1994				matchISSN	1994	9999					9999	matchISSN	Albany Law Re
9	0002-4821					matchISSN		9999					9999	matchISSN	Alberta Law R
10	0002-7162	1996				matchISSN	1996	9999					9999	matchISSN	The Annals of
11	0002-7766	1997				matchISSN	1997	9999					9999	matchISSN	The American
12	0002-919X	1996				matchISSN	1996	9999					9999	matchISSN	American Jou
13	0002-9300	1980				matchISSN	1980	9999					9999	matchISSN	American Jou
14	0002-953X	1985	1987			matchISSN	1985	1987				1985	1987	matchISSN	American Jou
15	0003-1453	1982				matchISSN	1982	9999				1982	9999	matchISSN	American Uni
16	0003-2344	2003	2006			matchISSN	2003	2006				2003	2006	matchISSN	Amusement E
17	0003-6056	1981				matchISSN	1981	9999				1981	9999	matchISSN	Antitrust Law
18	0003-858X	1992				matchISSN	1992	9999				1992	9999	matchISSN	Architectural
19	0004-1181	2006				matchISSN	2006	9999		check start		1998	9999	matchISSN	The Argus
20	0004-153X	1993				matchISSN	1993	9999				1993	9999	matchISSN	Arizona Law R
21	0004-1831	1993				matchISSN	1993	9999				1993	9999	matchISSN	Arkansas Law
22	0004-9042	2000	2002			matchISSN	2000	2002				2000	2002	matchISSN	Australian Ele
23	0004-976X	2000				matchISSN	2000	9999				2000	9999	matchISSN	Australian Mi
24	0005-1551	1988				matchISSN	1988	9999				1988	9999	matchISSN	Automotive N
25	0005-1551	1997	2000			matchISSN	1997	2000		check both		1988	9999	matchISSN	Automotive N
26	0005-2175	1975				matchISSN	1975	9999				1975	9999	matchISSN	Aviation Wee
27	0005-3635	2003				matchISSN	2003	9999				2003	9999	matchISSN	BackStage
28	0005-4127	1998	1998			matchISSN	1998	1998		check start		1996	1998	matchISSN	Bakery Produ
29	0005-450X	2003				matchISSN	2003	9999				2003	9999	matchISSN	Baltimore Jew

Formula for extracting and consolidating the data

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

9) Check for data that doesn't match



Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri 11
Adjust_mid
Adjust_end
Compare_start
Compare_mid
Compare_end

AB2 =IF(AND(Y2=AD2,Z2,AE2),"",IF(Y2=AD2,"check end",IF(Z2=AE2,"check start","check both")))

	J	M	N	O	P	X	Y	Z	AA	AB	AC	AD	AE	AF	AI
1	PrintISSN	MngStart	MngEnd	CstmStart	CstmEnd	1683	start	end		269		start	end	1683	Title
2	0000-0019	2000				matchISSN	2000	9999				2000	9999	matchISSN	Publishers We
3	0001-4672	1998						9999							ccountancy
4	0001-8899	1986						9999							dvertising A
5	0001-8899	1994	1995					1995		check both					dvertising A
6	0002-371X	1994						9999							kron Law Rev
7	0002-4279	1993						9999							labama Law
8	0002-4678	1994						9999							lbany Law Re
9	0002-4821							9999		check start		1997	9999	matchISSN	Alberta Law R
10	0002-7162	1996						9999				1996	9999	matchISSN	The Annals of
11	0002-7766	1997						9999				1997	9999	matchISSN	The American
12	0002-919X	1996						9999				1996	9999	matchISSN	American Jou
13	0002-9300	1980						9999				1980	9999	matchISSN	American Jou
14	0002-953X	1985	1987					1987				1985	1987	matchISSN	American Jou
15	0003-1453	1982						9999				1982	9999	matchISSN	American Uni
16	0003-2344	2003	2006			matchISSN	2003	2006				2003	2006	matchISSN	Amusement E
17	0003-6056	1981				matchISSN	1981	9999				1981	9999	matchISSN	Antitrust Law
18	0003-858X	1992				matchISSN	1992	9999				1992	9999	matchISSN	Architectural
19	0004-1181	2006				matchISSN	2006	9999		check start		1998	9999	matchISSN	The Argus
20	0004-153X	1993				matchISSN	1993	9999				1993	9999	matchISSN	Arizona Law R
21	0004-1831	1993				matchISSN	1993	9999				1993	9999	matchISSN	Arkansas Law
22	0004-9042	2000	2002			matchISSN	2000	2002				2000	2002	matchISSN	Australian Ele
23	0004-976X	2000				matchISSN	2000	9999				2000	9999	matchISSN	Australian Mi
24	0005-1551	1988				matchISSN	1988	9999				1988	9999	matchISSN	Automotive N
25	0005-1551	1997	2000			matchISSN	1997	2000		check both		1988	9999	matchISSN	Automotive N
26	0005-2175	1975				matchISSN	1975	9999				1975	9999	matchISSN	Aviation Wee
27	0005-3635	2003				matchISSN	2003	9999				2003	9999	matchISSN	BackStage
28	0005-4127	1998	1998			matchISSN	1998	1998		check start		1996	1998	matchISSN	Bakery Produ
29	0005-450X	2003				matchISSN	2003	9999				2003	9999	matchISSN	Baltimore Jew

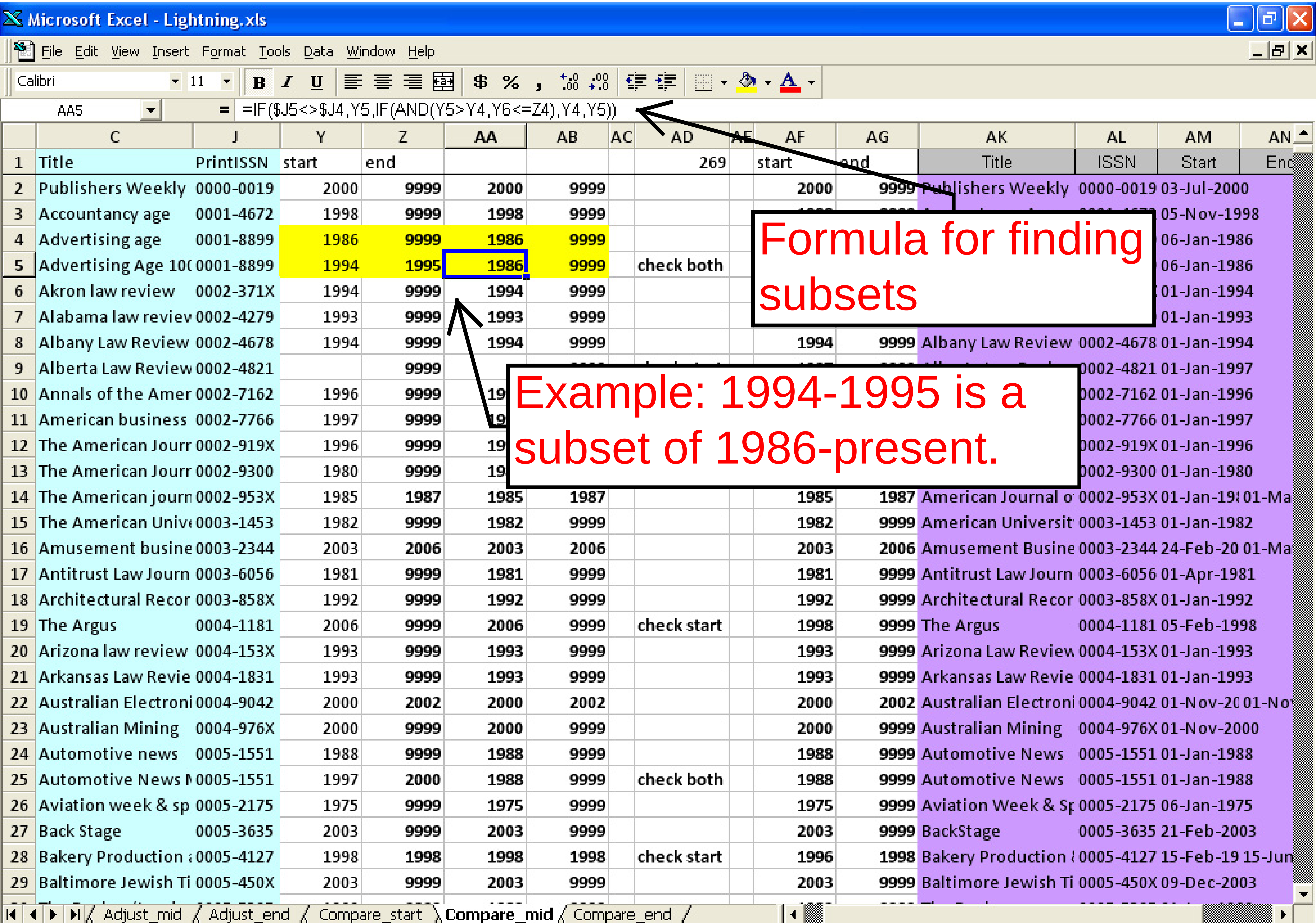
Note that only 269 of the 1683 matches will need any further checking.

Formula for automatic checking

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

10) Check for "duplicate" items that are subsets of other items, and adjust dates accordingly.


Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri
11
Adjust_mid
Adjust_end
Compare_start
Compare_mid
Compare_end

AA5 = =IF($J5<>$J4,Y5,IF(AND(Y5>Y4,Y6<=Z4),Y4,Y5))

	C	J	Y	Z	AA	AB	AC	AD	AE	AF	AG	AK	AL	AM	AN
1	Title	PrintISSN	start	end				269		start	end	Title	ISSN	Start	End
2	Publishers Weekly	0000-0019	2000	9999	2000	9999				2000	9999	Publishers Weekly	0000-0019	03-Jul-2000	
3	Accountancy age	0001-4672	1998	9999	1998	9999								05-Nov-1998	
4	Advertising age	0001-8899	1986	9999	1986	9999								06-Jan-1986	
5	Advertising Age 10(	0001-8899	1994	1995	1986	9999		check both						06-Jan-1986	
6	Akron law review	0002-371X	1994	9999	1994	9999								01-Jan-1994	
7	Alabama law reviev	0002-4279	1993	9999	1993	9999								01-Jan-1993	
8	Albany Law Review	0002-4678	1994	9999	1994	9999				1994	9999	Albany Law Review	0002-4678	01-Jan-1994	
9	Alberta Law Review	0002-4821		9999									0002-4821	01-Jan-1997	
10	Annals of the Amer	0002-7162	1996	9999	19								0002-7162	01-Jan-1996	
11	American business	0002-7766	1997	9999	19								0002-7766	01-Jan-1997	
12	The American Jourr	0002-919X	1996	9999	19								0002-919X	01-Jan-1996	
13	The American Jourr	0002-9300	1980	9999	19								0002-9300	01-Jan-1980	
14	The American journ	0002-953X	1985	1987	1985	1987				1985	1987	American Journal o	0002-953X	01-Jan-19!	01-Ma
15	The American Univ	0003-1453	1982	9999	1982	9999				1982	9999	American Universit	0003-1453	01-Jan-1982	
16	Amusement busine	0003-2344	2003	2006	2003	2006				2003	2006	Amusement Busine	0003-2344	24-Feb-20	01-Ma
17	Antitrust Law Journ	0003-6056	1981	9999	1981	9999				1981	9999	Antitrust Law Journ	0003-6056	01-Apr-1981	
18	Architectural Recor	0003-858X	1992	9999	1992	9999				1992	9999	Architectural Recor	0003-858X	01-Jan-1992	
19	The Argus	0004-1181	2006	9999	2006	9999		check start		1998	9999	The Argus	0004-1181	05-Feb-1998	
20	Arizona law review	0004-153X	1993	9999	1993	9999				1993	9999	Arizona Law Review	0004-153X	01-Jan-1993	
21	Arkansas Law Revie	0004-1831	1993	9999	1993	9999				1993	9999	Arkansas Law Revie	0004-1831	01-Jan-1993	
22	Australian Electroni	0004-9042	2000	2002	2000	2002				2000	2002	Australian Electroni	0004-9042	01-Nov-20	01-No
23	Australian Mining	0004-976X	2000	9999	2000	9999				2000	9999	Australian Mining	0004-976X	01-Nov-2000	
24	Automotive news	0005-1551	1988	9999	1988	9999				1988	9999	Automotive News	0005-1551	01-Jan-1988	
25	Automotive News I	0005-1551	1997	2000	1988	9999		check both		1988	9999	Automotive News	0005-1551	01-Jan-1988	
26	Aviation week & sp	0005-2175	1975	9999	1975	9999				1975	9999	Aviation Week & Sp	0005-2175	06-Jan-1975	
27	Back Stage	0005-3635	2003	9999	2003	9999				2003	9999	BackStage	0005-3635	21-Feb-2003	
28	Bakery Production	0005-4127	1998	1998	1998	1998		check start		1996	1998	Bakery Production	0005-4127	15-Feb-19	15-Jun
29	Baltimore Jewish Ti	0005-450X	2003	9999	2003	9999				2003	9999	Baltimore Jewish Ti	0005-450X	09-Dec-2003	

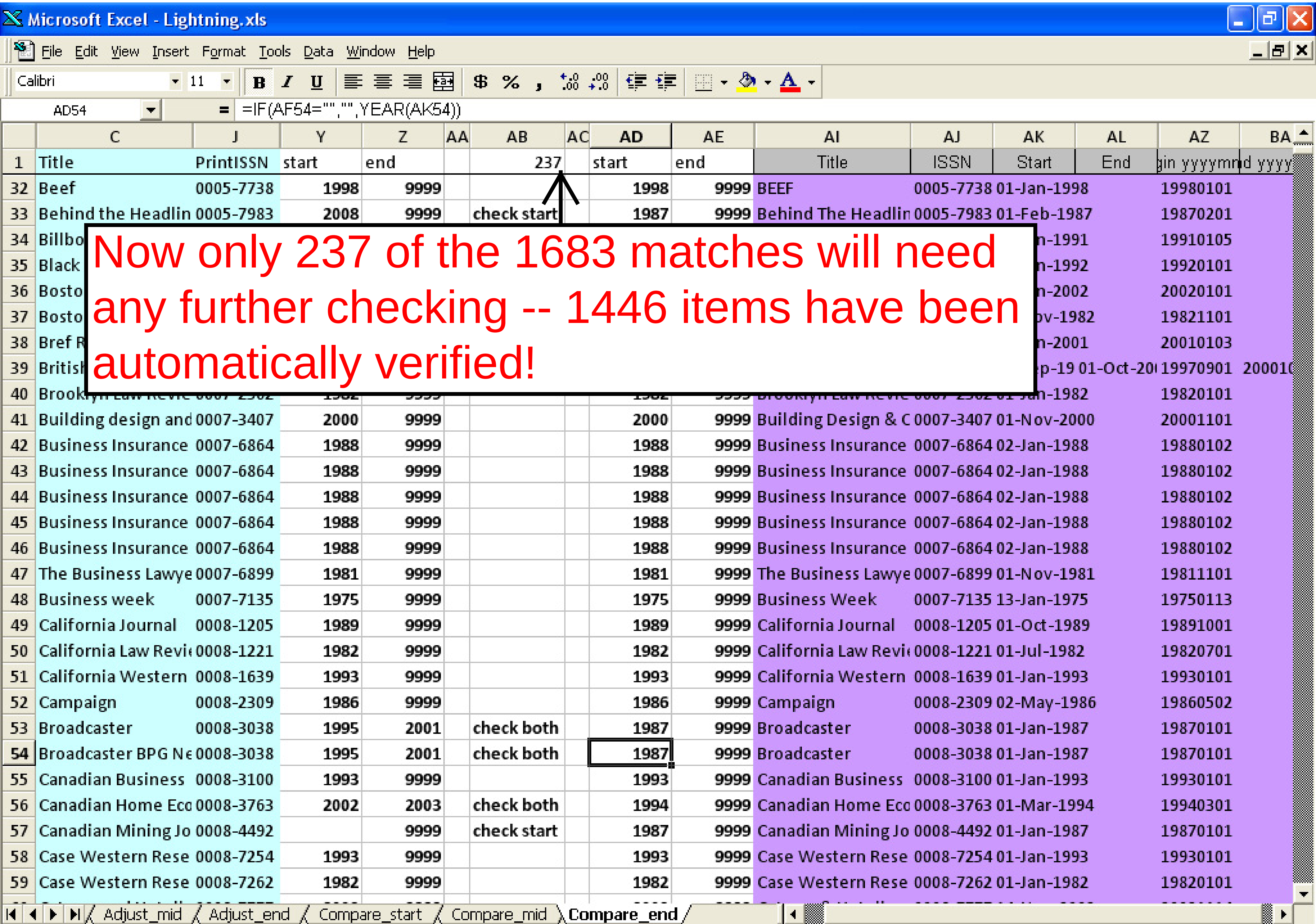
Formula for finding subsets

Example: 1994-1995 is a subset of 1986-present.

Data Integrity (Cheap, Fast, and Easy)

by Gwen Exner

11) Overwrite the extracted data with the adjusted data.



Microsoft Excel - Lightning.xls
File Edit View Insert Format Tools Data Window Help
Calibri 11

AD54 = =IF(AF54="","",YEAR(AK54))

	C	J	Y	Z	AA	AB	AC	AD	AE	AI	AJ	AK	AL	AZ	BA
1	Title	PrintISSN	start	end		237		start	end	Title	ISSN	Start	End	gin yyyymm	d yyyy
32	Beef	0005-7738	1998	9999				1998	9999	BEEF	0005-7738	01-Jan-1998		19980101	
33	Behind the Headlin	0005-7983	2008	9999		check start		1987	9999	Behind The Headlin	0005-7983	01-Feb-1987		19870201	
34	Billbo											n-1991		19910105	
35	Black											n-1992		19920101	
36	Bosto											n-2002		20020101	
37	Bosto											ov-1982		19821101	
38	Bref R											n-2001		20010103	
39	British											p-19	01-Oct-20	19970901	20001C
40	Brooklyn Law Revie	0007-2302	1982	9999				1982	9999	Brooklyn Law Revie	0007-2302	01-Jan-1982		19820101	
41	Building design and	0007-3407	2000	9999				2000	9999	Building Design & C	0007-3407	01-Nov-2000		20001101	
42	Business Insurance	0007-6864	1988	9999				1988	9999	Business Insurance	0007-6864	02-Jan-1988		19880102	
43	Business Insurance	0007-6864	1988	9999				1988	9999	Business Insurance	0007-6864	02-Jan-1988		19880102	
44	Business Insurance	0007-6864	1988	9999				1988	9999	Business Insurance	0007-6864	02-Jan-1988		19880102	
45	Business Insurance	0007-6864	1988	9999				1988	9999	Business Insurance	0007-6864	02-Jan-1988		19880102	
46	Business Insurance	0007-6864	1988	9999				1988	9999	Business Insurance	0007-6864	02-Jan-1988		19880102	
47	The Business Lawye	0007-6899	1981	9999				1981	9999	The Business Lawye	0007-6899	01-Nov-1981		19811101	
48	Business week	0007-7135	1975	9999				1975	9999	Business Week	0007-7135	13-Jan-1975		19750113	
49	California Journal	0008-1205	1989	9999				1989	9999	California Journal	0008-1205	01-Oct-1989		19891001	
50	California Law Revi	0008-1221	1982	9999				1982	9999	California Law Revi	0008-1221	01-Jul-1982		19820701	
51	California Western	0008-1639	1993	9999				1993	9999	California Western	0008-1639	01-Jan-1993		19930101	
52	Campaign	0008-2309	1986	9999				1986	9999	Campaign	0008-2309	02-May-1986		19860502	
53	Broadcaster	0008-3038	1995	2001		check both		1987	9999	Broadcaster	0008-3038	01-Jan-1987		19870101	
54	Broadcaster BPG Ne	0008-3038	1995	2001		check both		1987	9999	Broadcaster	0008-3038	01-Jan-1987		19870101	
55	Canadian Business	0008-3100	1993	9999				1993	9999	Canadian Business	0008-3100	01-Jan-1993		19930101	
56	Canadian Home Eco	0008-3763	2002	2003		check both		1994	9999	Canadian Home Eco	0008-3763	01-Mar-1994		19940301	
57	Canadian Mining Jo	0008-4492		9999		check start		1987	9999	Canadian Mining Jo	0008-4492	01-Jan-1987		19870101	
58	Case Western Rese	0008-7254	1993	9999				1993	9999	Case Western Rese	0008-7254	01-Jan-1993		19930101	
59	Case Western Rese	0008-7262	1982	9999				1982	9999	Case Western Rese	0008-7262	01-Jan-1982		19820101	

Now only 237 of the 1683 matches will need any further checking -- 1446 items have been automatically verified!