Att hämta organisationers publikationsposter ur DiVA

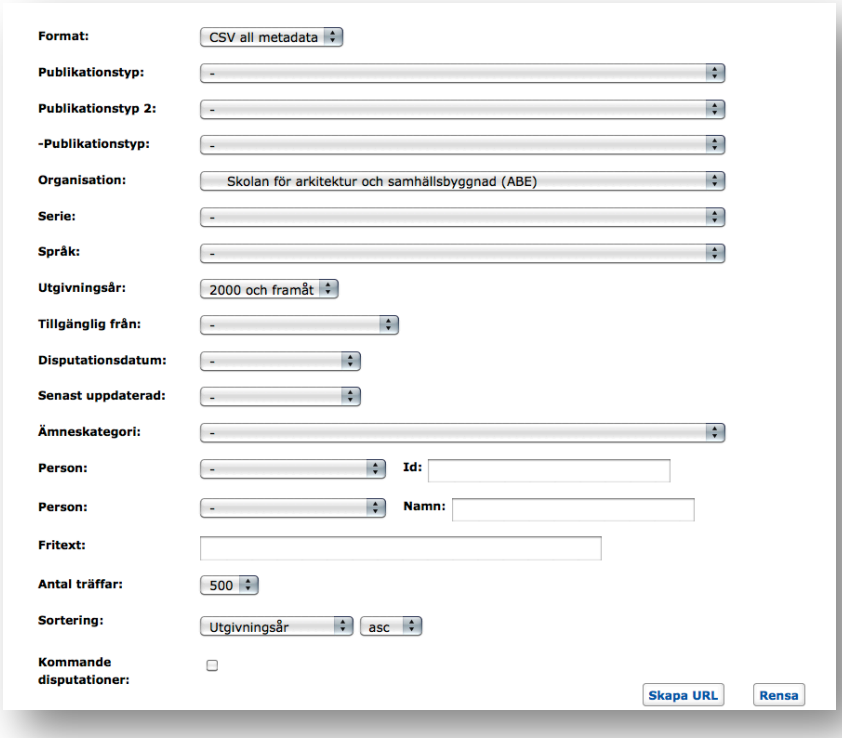
Ulf Kronman, 2011-08-08. Version 1.0

Följande guide beskriver hur man kan ladda ned kompletta publikationsposter i så kallat CSV-format[1] för organisationer från DiVA. Guiden är ursprungligen skriven för uttag av poster för KTH:s skolor, varför adresserna till DiVA-server är KTH:s. Andra organisationer byter lämpligen ut *kth*.diva-portal.org mot *<egen organisation>*.diva-portal.org.

Sida för utsökning ur DiVA

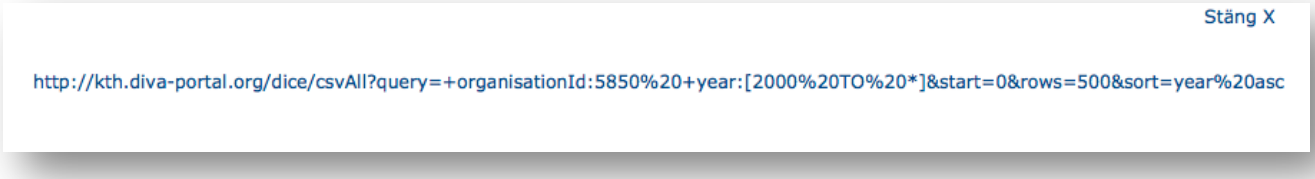
Om man vill få ut all data för alla publikationer för en organisatorisk enhet går man till sidan *Utsökning* > *Skapa länk utsökning* på adress: http://*kth*.diva-portal.org/smash/builder.jsf?type=createLink

På sidan gör man sedan ett urval och väljer överst i rullgardinsmenyn att man vill ha *CSV all metadata*. Möjligheterna att göra detaljerade urval på publiceringsår är begränsade, så dessa parametrar får man sedan justera manuellt i den länk som skapas av formuläret när man trycker på *Skapa URL* längst ned. Antalet träffar kan också behöva justeras manuellt i länken efteråt, eftersom det maximala antalet man kan erhålla via formuläret är 500.



[1] Comma Separated Values. Publikationsposterna listas en per rad och fält avskiljs med kommatecken.

Efter att man tryckt på *Skapa URL* får man en klickbar länk på skärmen.



Man bör inte klicka på länken, utan istället klippa ut den och spara den i ett redigeringsprogram för att ändra parametrarna i länken, så man får det urval man önskar[2].

Parametrar i utsökningslänken

Efter att man har klistrat in sin länk (URL) i redigeringsprogrammet kan man justera parametrarna så att man får exakt det urval man önskar. Följande parametrar är bra att känna till:

organisationId

+organisationId:5850

Varje organisatorisk enhet som finns registrerad i DiVA har ett unikt organisationId. Det finns vad jag vet ingen samlad förteckning över dessa ID:s, så det bästa sättet att få fram dem är att välja en organisation från rullgardinsmenyn i formuläret och låta formuläret generera en URL. I URL:en finns organisationens ID angiven i strängen +organisationId:5850.

year

+year:[2005%20TO%202010]

Intervallet för de år som ska levereras anges i klammer på följande vis: [2005 TO 2010]. Eftersom blanktecknet mellan åren och TO inte är ett tillåtet tecken i en URL kommer det att ersättas med sin kod %20 (se nedan), varför det kan vara lite knepigt att urskilja siffrorna i själva årtalen när man ska ändra dem.

rows

&rows=10000

I den URL som skapas av webbsidan får man ut ett värde på hur många poster som maximalt ska levereras av utsökningen. Det högsta värdet man kan få ut via formuläret är &rows=500 så det värdet kan man behöva ändra för att få ut alla poster från en skola samtidigt. Jag har valt att sätta &rows=10000 för att vara säker på att få med alla poster och det verkar fungera bra.

%20

En URL får inte innehålla några tomrum, för då kan webbservern uppfatta det som om URL:en tar slut där mellanslaget dyker upp. För att undvika detta kodar man mellanslag med sin ASCII-kod, som råkar vara nummer 20. Tecknet som används för att tala om att man använder en ASCII-kod är procenttecknet, och således blir tecknet mellanslag %20 när den finns i en URL.

---

[2] För denna typ av redigeringar rekommenderar jag för Windows det fria programmet PSPad http://www.pspad.com/en/ och för Mac OS X gratisprogrammet TextWrangler http://www.barebones.com/products/textwrangler/

sort

&sort=year%20asc

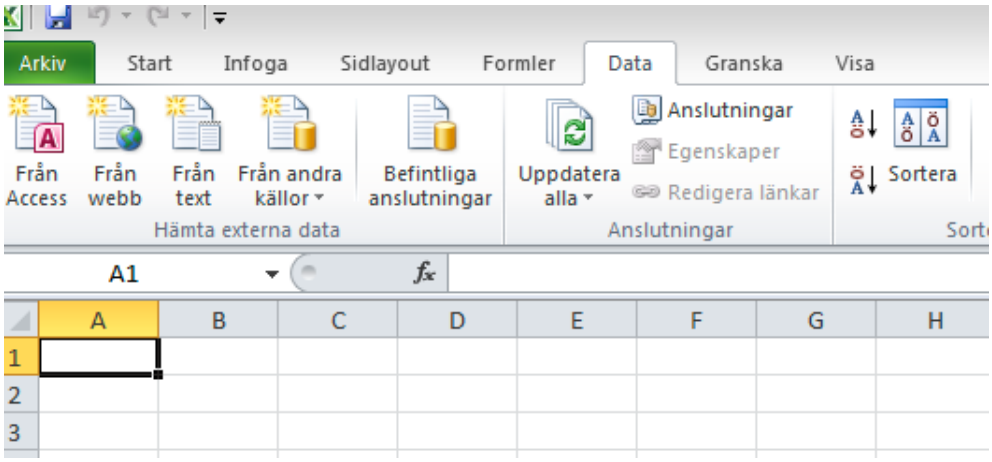
Med parametern sort väljer man efter vilket fält posterna ska sorteras vid leveransen. Om man avser att använda posterna i Excel senare är sorteringen inte viktig, för man kan alltid sortera posterna i Excel. I exemplet ovan har jag valt att sortera efter publiceringsår i stigande ordning (asc = ascending).

Nedladdning av CSV-fil

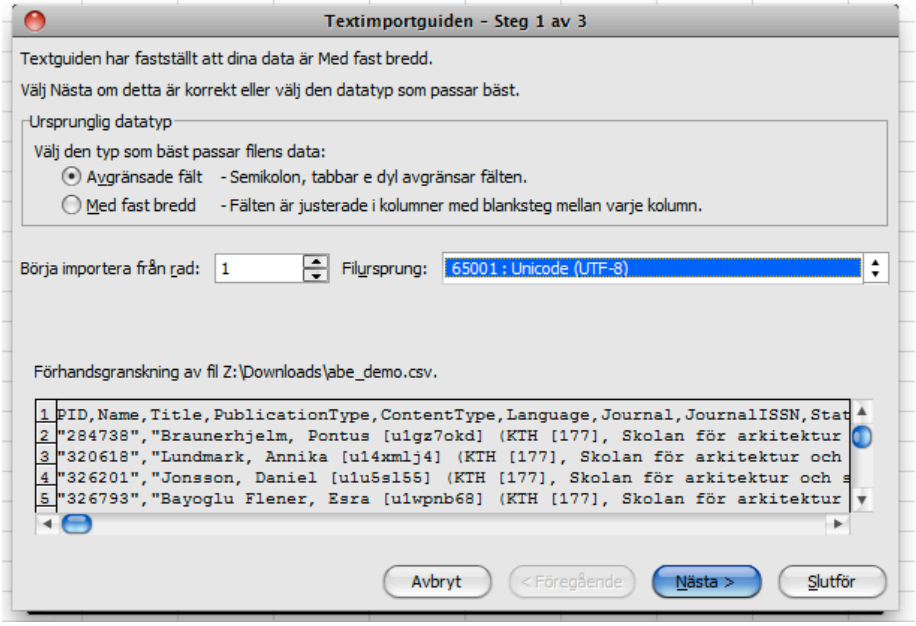
När man är klar med sina justeringar av URL:en kopierar man den från sitt redigeringsprogram och klistrar in den genererade URL:en i adressfönstret på sin webbläsare. När man trycker på <retur> kommer webbläsaren att öppna en dialogruta för att öppna eller spara en fil med namn **csvAll**. Man bör inte välja att öppna filen (eftersom den inte har något filsuffix öppnas den oftast med fel program), utan istället välja att spara den på hårddisken. När man sparar filen bör man passa på att byta namnet csvAll till något mer signifikant och samtidigt lägga till fil-suffixet .csv.

Inläsning av CSV-fil i Excel

För att använda den sparade CSV-filen i Excel 2010 startar man Excel med ett tomt kalkylblad och väljer *Data > Från text*

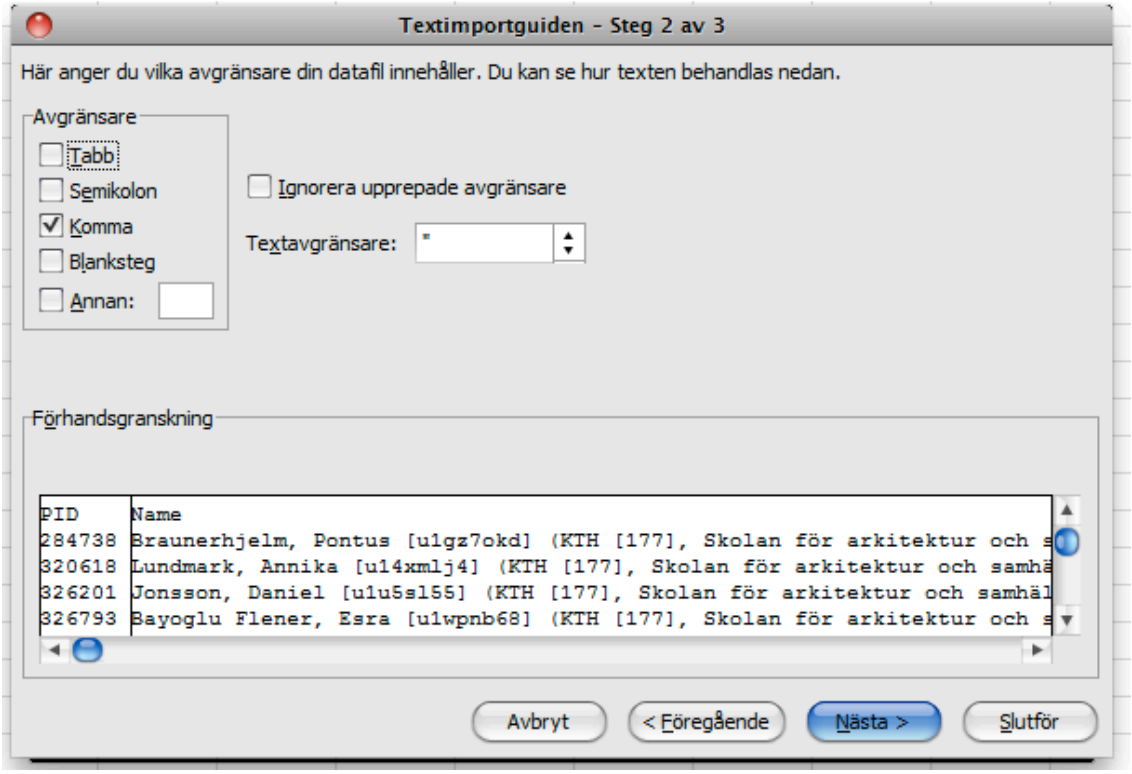


I dialogrutan som öppnas när man klickar på *Från text* väljer man sin CSV-fil och får sedan se *Textimportguiden* med följande dialogruta:

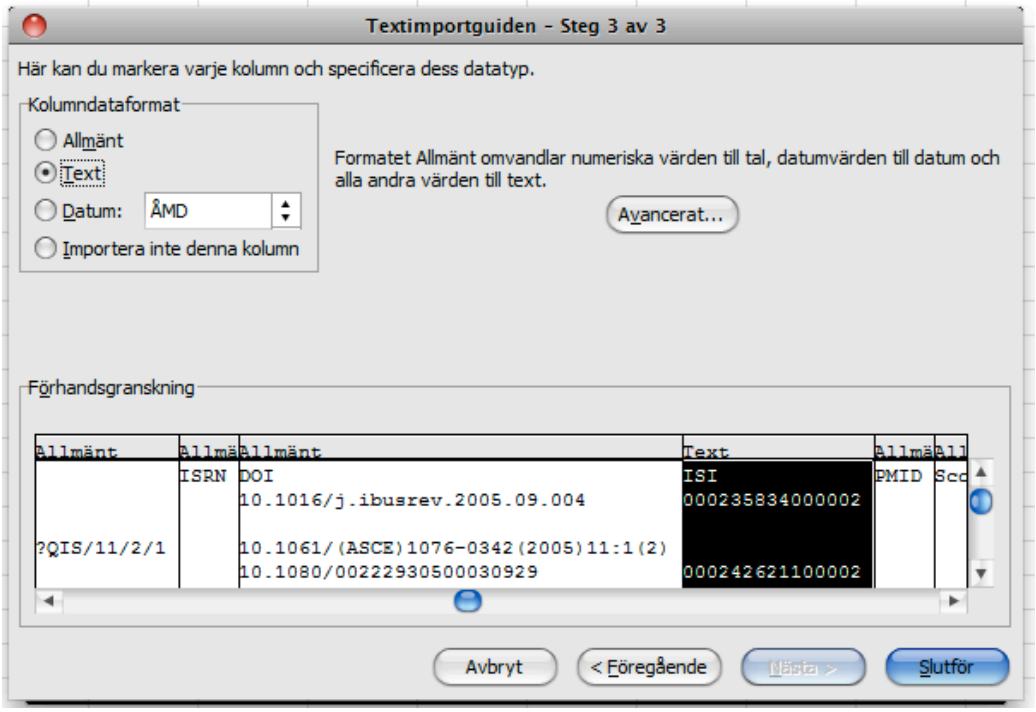


I denna dialogruta väljer man *Avgänsade fält* och *Filursprung 65001: Unicode (UTF-8)*, som är det format och den teckenkodning som DiVA använt för sina data. Om man inte väljer Unicode UTF-8 kommer inte svenska ÅÄÖ och andra diakritiska tecken att presenteras på ett korrekt sätt i Excel.

Sedan klickar man på *Nästa >* och i efterföljande dialogruta väljer man *Komma* som avgränsare.

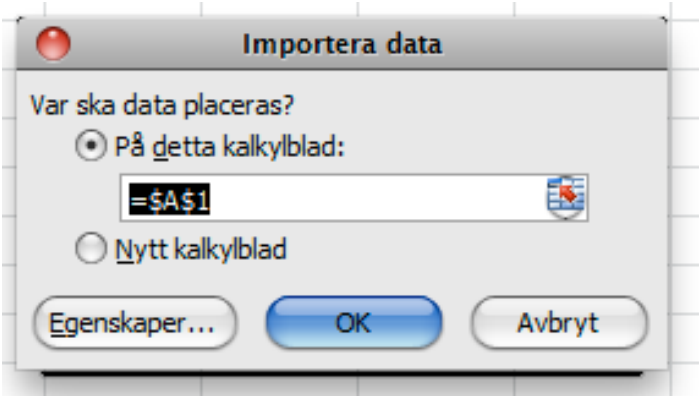


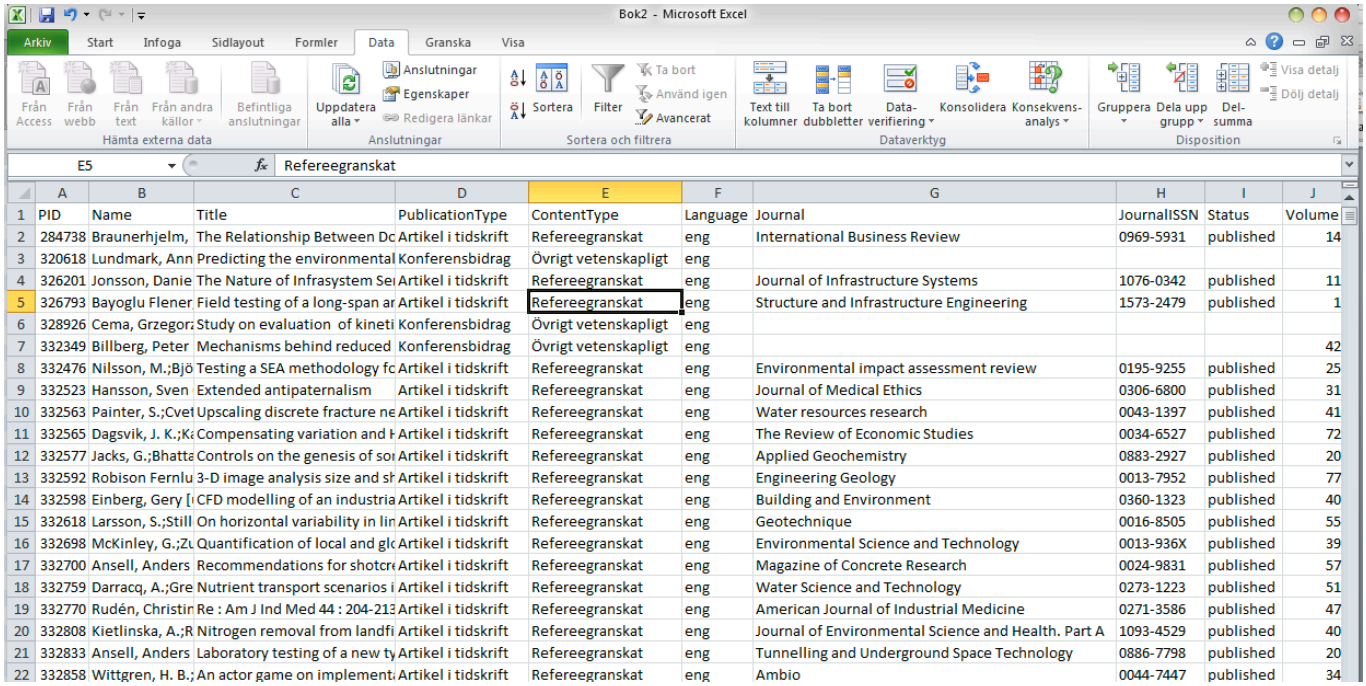
Man klickar åter igen på *Nästa >* och i dialogrutan som följer letar man upp kolumnen som innehåller ISI ID (UT-nummer), markerar den och väljer att kolumnen ska innehålla dataformatet *Text*.



Om man inte gör detta kommer Excel att behandla Thomsons UT-nummer som stora numeriska tal, vilket leder till att Excel städar bort alla inledande nollor ur numret och sedan representerar det som blir kvar av UT-numret som ett exponentiellt tal på detta vis: 2.26904E+11.

När man är klar med att sätta dataformatet på ISI ID (Thomsons UT-nummer) klickar man på *Slutför* och väljer sedan var data ska placeras i kalkylarket. Föreslagen position duger i de allra flesta fall, så man kan här klicka på *OK*.





	A	B	C	D	E	F	G	H	I	J
1	PID	Name	Title	PublicationType	ContentType	Language	Journal	JournalISSN	Status	Volume
2	284738	Braunerhjelm,	The Relationship Between Dc	Artikel i tidskrift	Refereegranskat	eng	International Business Review	0969-5931	published	14
3	320618	Lundmark, Ann	Predicting the environmental	Konferensbidrag	Övrigt vetenskapligt	eng				
4	326201	Jonsson, Danie	The Nature of Infrasystem Sei	Artikel i tidskrift	Refereegranskat	eng	Journal of Infrastructure Systems	1076-0342	published	11
5	326793	Bayoglu Flener	Field testing of a long-span ar	Artikel i tidskrift	Refereegranskat	eng	Structure and Infrastructure Engineering	1573-2479	published	1
6	328926	Cema, Grzegor	Study on evaluation of kineti	Konferensbidrag	Övrigt vetenskapligt	eng				
7	332349	Billberg, Peter	Mechanisms behind reduced	Konferensbidrag	Övrigt vetenskapligt	eng				42
8	332476	Nilsson, M.;Bjö	Testing a SEA methodology fo	Artikel i tidskrift	Refereegranskat	eng	Environmental impact assessment review	0195-9255	published	25
9	332523	Hansson, Sven	Extended antipaternalism	Artikel i tidskrift	Refereegranskat	eng	Journal of Medical Ethics	0306-6800	published	31
10	332563	Painter, S.;Cvet	Upscaling discrete fracture ne	Artikel i tidskrift	Refereegranskat	eng	Water resources research	0043-1397	published	41
11	332565	Dagsvik, J. K.;K	Compensating variation and H	Artikel i tidskrift	Refereegranskat	eng	The Review of Economic Studies	0034-6527	published	72
12	332577	Jacks, G.;Bhatta	Controls on the genesis of soi	Artikel i tidskrift	Refereegranskat	eng	Applied Geochemistry	0883-2927	published	20
13	332592	Robison Fernlu	3-D image analysis size and sh	Artikel i tidskrift	Refereegranskat	eng	Engineering Geology	0013-7952	published	77
14	332598	Einberg, Gery [	CFD modelling of an industria	Artikel i tidskrift	Refereegranskat	eng	Building and Environment	0360-1323	published	40
15	332618	Larsson, S.;Still	On horizontal variability in lin	Artikel i tidskrift	Refereegranskat	eng	Geotechnique	0016-8505	published	55
16	332698	McKinley, G.;Zu	Quantification of local and glc	Artikel i tidskrift	Refereegranskat	eng	Environmental Science and Technology	0013-936X	published	39
17	332700	Ansell, Anders	Recommendations for shotcre	Artikel i tidskrift	Refereegranskat	eng	Magazine of Concrete Research	0024-9831	published	57
18	332759	Darracq, A.;Gre	Nutrient transport scenarios i	Artikel i tidskrift	Refereegranskat	eng	Water Science and Technology	0273-1223	published	51
19	332770	Rudén, Christir	Re : Am J Ind Med 44 : 204-213	Artikel i tidskrift	Refereegranskat	eng	American Journal of Industrial Medicine	0271-3586	published	47
20	332808	Kietlinska, A.;R	Nitrogen removal from landfi	Artikel i tidskrift	Refereegranskat	eng	Journal of Environmental Science and Health. Part A	1093-4529	published	40
21	332833	Ansell, Anders	Laboratory testing of a new ty	Artikel i tidskrift	Refereegranskat	eng	Tunnelling and Underground Space Technology	0886-7798	published	20
22	332858	Wittgren, H. B.;	An actor game on implementi	Artikel i tidskrift	Refereegranskat	eng	Ambio	0044-7447	published	34

Avslutningsvis bör man notera att DiVA:s CSV-output innehåller en bugg, så att det kan förekomma "ohanterade" citationstecken i Abstract-fältet, varför detta fält kan delas upp i flera delar. Det påverkar i så fall alla fält som kommer efter Abstract-fältet för posten i fråga. Buggen är rapporterad till DiVA-supporten i juli 2011. Då fälten efter Abstract bara innehåller administrativa DiVA-data är detta inte något större problem, under förutsättning att man inte vill använda innehållet i abstract-fältet.