Victoria Gaunitz | Dataharvest 2023

Bring structure in unstructured documents

- Without coding

sverigeSRradio

Bring structure in unstructured documents

SOURCES:

1. People - find and talk to those who have knowledge about the subject.

sverigeSRradio

Bring structure in unstructured documents

SOURCES:

1. People - find and talk to those who have knowledge about the subject.
2. Documents - find the relevant papers that may exist.

sverigeSRradio

Bring structure in unstructured documents

SOURCES:

1. People - find and talk to those who have knowledge about the subject.
2. Documents - find the relevant papers that may exist.
3. Information - that has not been summarized or processed in any way.

sverigeSRradio

Bring structure in unstructured documents

STRUCTURE:

1. When you build a database, YOU are the architect.

There will be a lot of decisions to make!

sverigeSRradio

Bring structure in unstructured documents

STRUCTURE:

1. When you build a database, you are the architect.
2. Think in rows and columns.

sverigeSRradio

Bring structure in unstructured documents

STRUCTURE:

1. When you build a database, you are the architect.
2. Think in rows and columns.
3. You decide what each row should contain.

sverigeSRradio

Bring structure in unstructured documents

STRUCTURE:

1. When you build a database, you are the architect.
2. Think in rows and columns.
3. You decide what each row should contain.
4. And then you decide what you want to capture in the columns.

sverigeSRradio

Bring structure in unstructured documents

	A	B	C	D	E
1	Id	Namn	Kommun	Favoritmat	
2	1	Sandra	Karlstad	Pasta	
3	2	Victoria	Kalmar	Pizza	
4	3	Anna	Göteborg	Hamburgare	
5	4	Markus	Östersund	Sallad	
6	5	Fredrik	Växjö	Fiskgryta	
7					
8					

Bring structure in unstructured documents

STRUCTURE:

1. When you build a database, you are the architect.
2. Think in rows and columns.
3. You decide what each line should contain.
4. And then you decide what you want to capture in the columns.
5. Be aware that it will take time.

sverigeSRradio

Bring structure in unstructured documents

STRUCTURE:

1. When you build a database, you are the architect.
2. Think in rows and columns.
3. You decide what each line should contain.
4. And then you decide what you want to capture in the columns.
5. Be aware that it will take time.
6. And that you will make changes.

sverigeSRradio

Bring structure in unstructured documents

ARRANGEMENT:

1. What kinds of questions can we answer from these columns?

sverigeSRradio

Bring structure in unstructured documents

ARRANGEMENT:

1. What kinds of questions can we answer from these columns?
2. What anomalies do we need to consider?

sverigeSRradio

Bring structure in unstructured documents

ARRANGEMENT:

1. What kinds of questions can we answer from these columns?
2. What anomalies do we need to consider?
3. What is missing?

sverigeSRradio

Bring structure in unstructured documents

To think about

1. What is the source?

sverigeSRradio

Bring structure in unstructured documents

To think about

1. What is the source?
2. What kind of material is it? Verdicts? Incident reports? Decisions?

sverigeSRradio

Bring structure in unstructured documents

To think about

1. What is the source?
2. What kind of material is it? Verdicts? Incident reports? Decisions?
3. How do you get the material? Paper? Pdf? Is OCR needed?

sverigeSRradio

Bring structure in unstructured documents

To think about

1. What is the source?
2. What kind of material is it? Verdicts? Incident reports? Decisions?
3. How do you get the material? Paper? Pdf? Is OCR needed?
4. What information is in the material?

sverigeSRradio

Bring structure in unstructured documents

To think about

1. What is the source?
2. What kind of material is it? Verdicts? Incident reports? Decisions?
3. How do you get the material? Paper? Pdf? Is OCR needed?
4. What information is in the material?
5. Which unique identifier should you use to link the material and the data.

sverigeSRradio

Bring structure in unstructured documents

To think about

1. What is the source?
2. What kind of material is it? Verdicts? Incident reports? Decisions?
3. How do you get the material? Paper? Pdf? Is OCR needed?
4. What information is in the material?
5. Which unique identifier should you use to link the material and the data.
6. How to deal with anomalies? If you do it on an ongoing basis, you will see the problems as they arise

sverigeSRradio

Bring structure in unstructured documents

Make a test

1. Request a smaller amount of material, such as from a few district courts, for a shorter period of time or the five most recent documents.

sverigeSRradio

Bring structure in unstructured documents

Make a test

1. Request a smaller amount of material, such as from a few district courts, for a shorter period of time or the five most recent documents.
2. What do you need from the documents? Which variables are relevant?

sverigeSRradio

Bring structure in unstructured documents

Make a test

1. Request a smaller amount of material, such as from a few district courts, for a shorter period of time or the five most recent documents.
2. What do you need from the documents? Which variables are relevant?
3. What type of variable? Yes/no, x, date, free text.

sverigeSRradio

Bring structure in unstructured documents

One column for each answer

This makes it quick to fill in and (I think) easier to get an overview and make a summary

CTG	Cytotec	Snitt	Dödsfall	Förstföderska	Tidigare snitt	S
x		x	Barnet	x	Nej	S
x			Barnet	x	Nej	S
	x	x		Nej	x	K
				Nej	?	Y
x		x	Barnet	x	Nej	S
				Nej	?	S
	?			Nej	x	U
x		x		Nej	x	V
x		x		x	Nej	S
			Barnet	?	?	S
x				x	Nej	S
x		x	Barnen	Nej	?	S
x	?		Barnet	Nej	x	A
		x		?	?	H
x		x	Barnet	Nej	?	K
				?	?	M
x		x	Barnet	?	?	C
				?	?	H
x			Barnet	x	Nej	D
			Barnet	Nej	?	A
x		x		x	Nej	A

sverigesradio

Bring structure in unstructured documents

Make a test

1. Request a smaller amount of material, such as from a few district courts, for a shorter period of time or the five most recent documents.
2. What do you need from the documents? Which variables are relevant?
3. What type of variable? Yes/no, x, date, free text.
4. How much information do you need to enter?

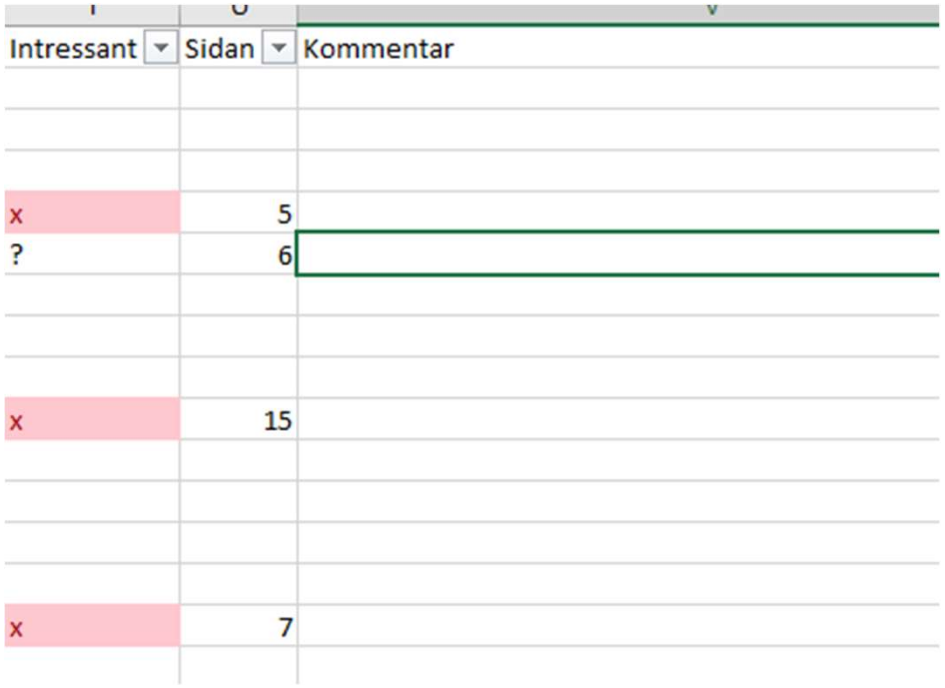
sverigeSRradio

Bring structure in unstructured documents

Longer texts do not need to be entered

- If the free text is an interesting case/case or in some other way special, put an x.
- You can add page number and a comment
- Since you still have the basic material, you can go back and read after you have gone through the documents.

- But sometimes, if there are a short final decision it can make it easier to free text search



Intressant	Sidan	Kommentar
x	5	
?	6	
x	15	
x	7	

sverigeSRradio

Bring structure in unstructured documents

Make a test

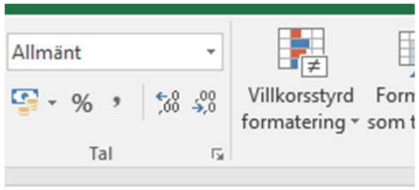
1. Request a smaller amount of material, such as from a few district courts, for a shorter period of time or the five most recent documents.
2. What do you need from the documents? Which variables are relevant?
3. What type of variable? Yes/no, x, date, free text.
4. How much information do you need to enter?
5. Remember that you still have the material.

sverigesradio

Bring structure in unstructured documents

Utilize excel/spreadsheet

1. Conditional formatting for date checking of the request



X	Y	Z
Begärt ut	Ensam	Gemensa
2021-09-07		x
2021-09-01	x	
2021-08-25	x	
2021-08-22		
2021-08-30		
2021-09-02		x
2021-08-16		
2021-09-01	x	
2021-08-28		

sverigesradio

Bring structure in unstructured documents

Utilize excel/spreadsheet

1. Conditional formatting for date checking of the request
2. Conditional formatting for checks
   a) Highest values

sverigeSRradio

Bring structure in unstructured documents

	A	B	C	D
1	Namn	Personnr	Förvärvsinkomst	
2	Anna Johansson	1933xxxx-xxxx	1 465 100	
3	Bertil Olsson	1955xxxx-xxxx	1 215 200	
4	Carl Carlsson	1950xxxx-xxxx	1 697 000	
5	David Andersson	1955xxxx-xxxx	1 130 100	
6	Erik Magnusson	1955xxxx-xxxx	1 276 900	
7	Fredrik Nilsson	1946xxxx-xxxx	1 317 700	
8	Gunnar Larsson	1956xxxx-xxxx	1 229 700	
9	Hanna Svensson	1948xxxx-xxxx	1 110 700	
10	Ingvar Karlsson	1951xxxx-xxxx	1 001 300	
11	Johan Gustavsson	1942xxxx-xxxx	1 748 200	
12				
13				

sverigesradio

Bring structure in unstructured documents

Utilize excel/spreadsheet

1. Conditional formatting for date checking of the request
2. Conditional formatting for checks
   a) Highest values
   b) Empty cells

sverigeSRradio

Bring structure in unstructured documents

Dom3	Stämning3	Utredning3	Dom4	Stämning4	Utredning4	Dom5	Stämning5	Utredning5	Antal
									0
T 613-23	x	x	T 46-23	x	x	T 82-22	x	x	15
T 163-23	x								6
T 976- 18	x	x	-	-	-	-	-	-	15
T 1-23		x	T 964-23		x	T 28-23		x	10
T 35-23	x	x	T 18-22	x	x	T 408-23	x	x	15
T 676-23	x	x	-	-	-	-	-	-	15
T 84-23	x	x	T 6-22	x	x	-	-	-	15
T 8-21	x	x	T 0-21	x	x	-	-	-	15
									0
T 892-22	x	x	T 205-22	x	x	T 7398-23	x	x	15
									0

sverigesradio

Bring structure in unstructured documents

Utilize excel/spreadsheet

1. Conditional formatting for date checking of the request
2. Conditional formatting for checks
   a) Highest values
   b) Empty cells
   c) Duplicates

sverigeSRradio

Bring structure in unstructured documents

	A	B	C
1	Namn	Parti	
2	Åsa Lindestam	S	
3	Lotta Johnsson Fornarve	V	
4	Kerstin Lundgren	C	
5	Lars Jilmstad	M	
6	Karin Enström	M	
7	Anders Ygeman	S	
8	Sultan Kayhan	S	
9	Sultan Kayhan	S	
10	Mattias Vepsä	S	
11	Barbro Westerholm	L	
12	Betty Malmberg	M	
13	Mikael Oscarsson	KD	
14	Elsemarie Bjellqvist	S	
15	Jan R Andersson	M	
16	Anders Åkesson	C	
17	Harald Hjalmarsson	M	

sverigeSRradio

Bring structure in unstructured documents

Utilize excel/spreadsheet

1. Conditional formatting for date checking of the request
2. Conditional formatting for checks
3. Compilations/Overview

sverigeSRradio

Bring structure in unstructured documents

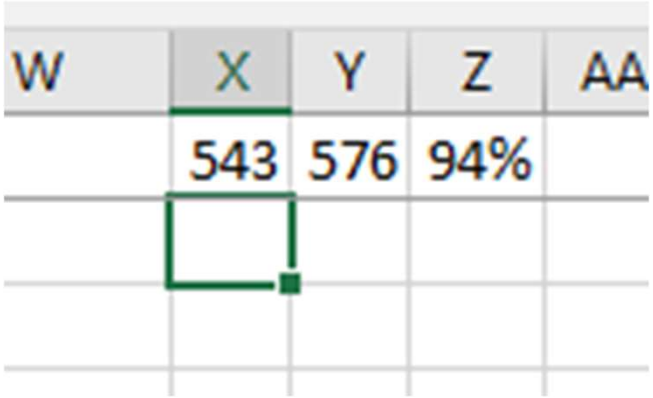
	CTG	41
	Cytotec	3
	Snitt	35
	Dödsfall	42
	Barnet	40
0		
1	**Förstföderska**	88
2	Ja	27
3	Nej	31
4		
5	**Tidigare snitt**	88
6	Ja	15
7	Nej	30
8		

sverigeSRradio

Bring structure in unstructured documents

Utilize excel/spreadsheet

1. Conditional formatting for date checking of the request
2. Conditional formatting for checks
3. Compilations/Overview
4. Percentage for motivation



sverigesradio

Bring structure in unstructured documents

IMPORTANT

1. Remember that you don't remember as much as you think.

sverigeSRradio

Bring structure in unstructured documents

IMPORTANT

1. Remember that you don't remember as much as you think.
2. Write explanations for what each column contains..

sverigeSRradio

Bring structure in unstructured documents

IMPORTANT

1. Remember that you don't remember as much as you think.
2. Write explanations for what each column contains.
3. Write down why you make the trade-offs/limitations that you do.

sverigeSRradio

Bring structure in unstructured documents

IMPORTANT

1. Remember that you don't remember as much as you think.
2. Write explanations for what each column contains.
3. Write down why you make the trade-offs/limitations that you do.
4. There is no right solution.

sverigeSRradio

Time for questions!

sverigeSRradio