

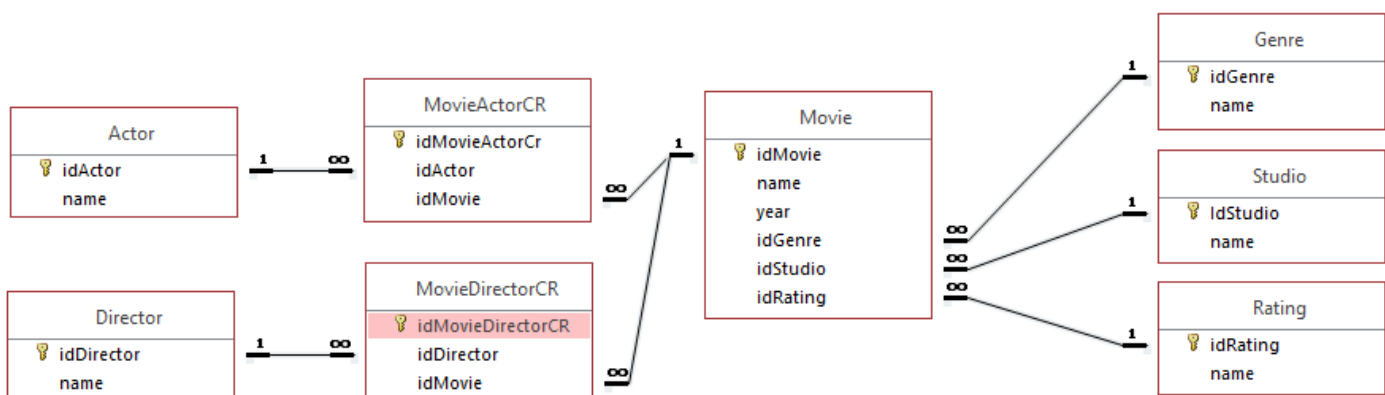
# 5 Data importeren en Database maken

Tot nu toe hebben we gezien dat we een database handmatig kunnen vullen. Een andere optie is dat we een database gaan vullen vanuit een programmeertaal.

In dit hoofdstuk gaan we de database vullen door middel van een bestand met gegevens. We gaan deze gegevens vervolgens weer in een relationele database zetten zodat we zo min mogelijk redundante (dubbele) data hebben.

We gaan een database maken met daarin heel veel Dvd's. We gebruiken hiervoor de gegevens die op een site te vinden zijn. Deze gegevens gaan we in onze eigen database model opslaan. We gaan in een volgend hoofdstuk deze data gebruiken in een python programma.

De uiteindelijke database zal er als volgt uitzien.



## 5.1 Data van internet halen

Op de site <http://www.hometheaterinfo.com/> (DEZE SITE BESTAAT INMUDELS NIET MEER) staan verschillende bestanden die informatie geven over de verschillende films die zij hebben. Deze lijsten gaan we in onze eigen database zetten en opslaan op de manier die voor ons handig is.

De lijst met dvd titels : <http://www.hometheaterinfo.com/download/dvdlist.zip>

De lijst met directors : [http://www.hometheaterinfo.com/download/DVD\\_Directors.zip](http://www.hometheaterinfo.com/download/DVD_Directors.zip)

De lijst met actueren : [http://www.hometheaterinfo.com/download/DVD\\_Actors.zip](http://www.hometheaterinfo.com/download/DVD_Actors.zip)

De lijst met DVD titels in CSV : [http://www.hometheaterinfo.com/download/dvd\\_csv.zip](http://www.hometheaterinfo.com/download/dvd_csv.zip)

En nog wat informatie over de velden die gebruikt zijn : <http://www.hometheaterinfo.com/keyto.htm>

This database is provided for non-commercial use only.

If you wish to incorporate the data in this list for any commercial use, please contact [doug@hometheaterinfo.com](mailto:doug@hometheaterinfo.com).

If you are using the list for a non-commercial site please give a link to:

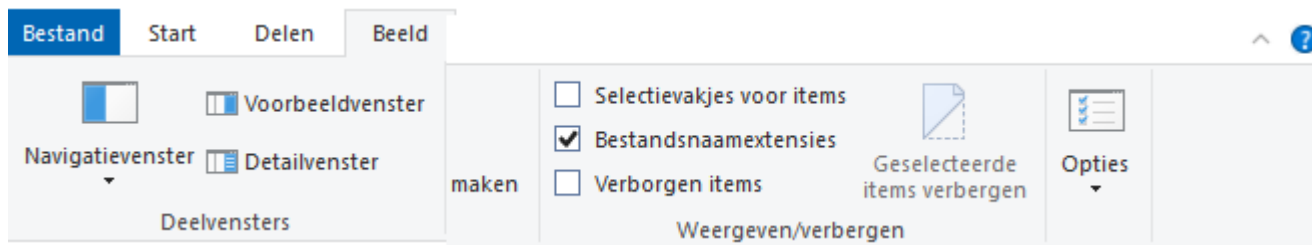
Home Theater Info <<http://www.hometheaterinfo.com/>> and Michael's Movie Mayhem <<http://dvdlist.kazart.com/>> acknowledging the effort that goes into this database.

Contact [doug@hometheaterinfo.com](mailto:doug@hometheaterinfo.com) to include your link to the Home Theater Info Link page.

Deze verschillende bestanden gaan we vervolgens uitpakken, en in dezelfde folder zetten. We zien in die folder dan de volgende bestanden. Je moet nooit naar het plaatje voor de bestanden kijken. Deze plaatjes geven aan met welk programma je dit bestand standaard zal openen, en heeft niets te maken wat voor bestand het is.

Voor deze opdracht gebruiken we Excel- en tekstbestanden. We kunnen Excel bestanden herkennen door de extensie (laatste letters achter de bestandsnaam). Ook in de kolom Type staat wat voor soort bestand het is.

Zie je niet de extensies van de bestanden dan kan het handig zijn om dat aan te zetten. Standaard staat dit in Windows uit, dit omdat de meeste Windows gebruikers niet technisch met de computer bezig zijn, maar op het niveau zitten dat ze een brief willen schrijven.



Door de bestandsextensie aan te zetten zie je wel de bestandstypen direct aan het bestand vast.

Excel bestanden zijn de bestanden met XLS (Excel 97) en XLXS achter de namen.

Tekstbestanden zijn de bestanden met bijvoorbeeld CSV en TXT achter de naam.

Drive - Da Vinci College > Boeken > Python > Van stroomdiagrammen naar Python code > H23 > DVDs

Naam	Status	Gewijzigd op	Type	Grootte
Actors_Index.csv	✓	5-4-2020 15:01	CSV-bestand	30.894 kB
Actors_Names.csv	✓	5-4-2020 15:01	CSV-bestand	6.316 kB
copyright.txt	✓	5-4-2020 14:56	TXT-bestand	1 kB
Directors_Index.csv	✓	5-4-2020 14:56	CSV-bestand	3.187 kB
Directors_Names.csv	✓	5-4-2020 14:56	CSV-bestand	807 kB
dvd_csv.txt	✓	5-4-2020 18:32	TXT-bestand	57.278 kB
dvdlist.xls	✓	13-4-2020 14:32	Microsoft Excel 97...	84.991 kB
dvdlist.xlsx	✓	18-5-2020 14:00	Microsoft Excel-w...	32.752 kB
IMDB-Top-250-Movies-Excel-Dashboard.xlsx	✓	5-4-2020 19:49	Microsoft Excel-w...	248 kB
IMDB-Top-250-Rev-1.xlsx	✓	5-4-2020 14:40	Microsoft Excel-w...	115 kB
Nieuw tekstdocument.txt	✓	5-4-2020 18:33	TXT-bestand	3 kB

Een tekstbestand kan je openen met een teksteditor zoals Notepad, Notepad++ of Sublime. We konden ook Python code openen met Notepad++. Wat deze bestanden kenmerkt is dat deze alleen tekst bevatten en geen tekst opmaak.

Een Excel-bestand heeft wel die tekst opmaak. Er staat in wat voor lettertype er wordt gebruikt, welke kolom of tabblad de gegevens moeten staan en hoe de auteur heet van het bestand. Dit zijn dingen die een tekstbestand (plain text) nooit zal hebben.

Open je een Excel-bestand in een tekst editor dan zal je allemaal vreemde dingen op je scherm zien.

```
dvdlst.xlsx
1 PKETXEOTDC4NULACKNULBSNULNULNUL!NULESC&Md'SOHNULNULµ NULNULDC3NULBSSTX[Content_Types].xml <EOT
2 δuĐÆİ+ñðøšδE DC4HÊkef+JlNULÁýøÄÖèq DC3SOHVTZöX%† (~•DC2èACKæÂ2Dδ<2VTÉ) áÇ4-QÖVT5BELy3ESË:xSTXOETXj5Äx¿
3 'q.ô*îðfX' `6mý:-NAK.uESBSjSOÄETBÁ@ðîxfµ»Ý}<.øEOTmZ$EM%Ä`EOT VTUq`fä=[`{ A.KP) GEB`È.»*Ü%BCÑ~dšq7-"
4 #
5 bÿD@) SUBDLEÜ,,*BELio*%ENQGESCêšö+48δSUBhDC2E) ÖóSUB$;«EMiKFô%´øùĐ[äýk«*jVTxVÂQÇÄVTDLEÖàðÜUSBSr@k@F~ãĐ:
6 *ñaðÝ< C:ei< æ'xes...?%S>...>ÄiT°O™Ö@L+tx8Ü,-íE@'pFÜ5/ÈÜ7è...µY»´yô:ýESC@;|Cm-Ü| íÓldèÜ~È~NULNULNULÿÿETXNUL
7 ;žE±ETBNAKjδETBIFC'kè4EOT
8 Ý&öµC(}...CAN`ŠÓÁ`DC2ktwEES÷SUBSO»^KK+vSUBMèHES*BS"%~dilZ*†NfŮièHX] DC3E!*EB#-~"3-!·ESCcCSDC4ŠÊ~NEGfZ%
9 WiEBèúVTèæSYNÄ™STXUS>·2`™ñ ...vIm,]})<`SN`Ñ(CSδx~jÄc%øa$Ä=5áHOÄt"!`ACKIE)TSOB=VTkSTXiø[ýÿÖ«Ä^T-°Ä%át.
10 [EB]b;S2%#SIžÊİ·°ÈçSUB\ :ù,,bK`+i,,dVT«DC4n0çq,E`BGSBeM%STXCóè-Ö%„ETXMSÄEBecGSëÑäb~"SOóâjéó^Èää<üCÓ@4
11 ÆvEETC~x1=8EETBÄ:EB
```

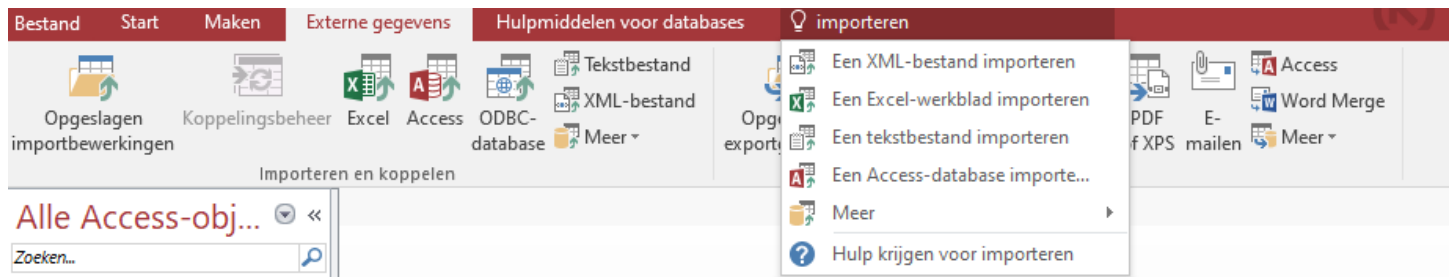
De tekst editor bekijkt het bestand en vertaalt de binaire code naar ASCII-code. Het is alleen geen ASCII maar een eigen opmaak van het programma.

## 5.2 Maken Access database

### 5.2.1 Importeren van Excel bestand

We beginnen met het opstarten van Microsoft Access, en maken een nieuwe lege database aan. Deze database noemen we Movies.

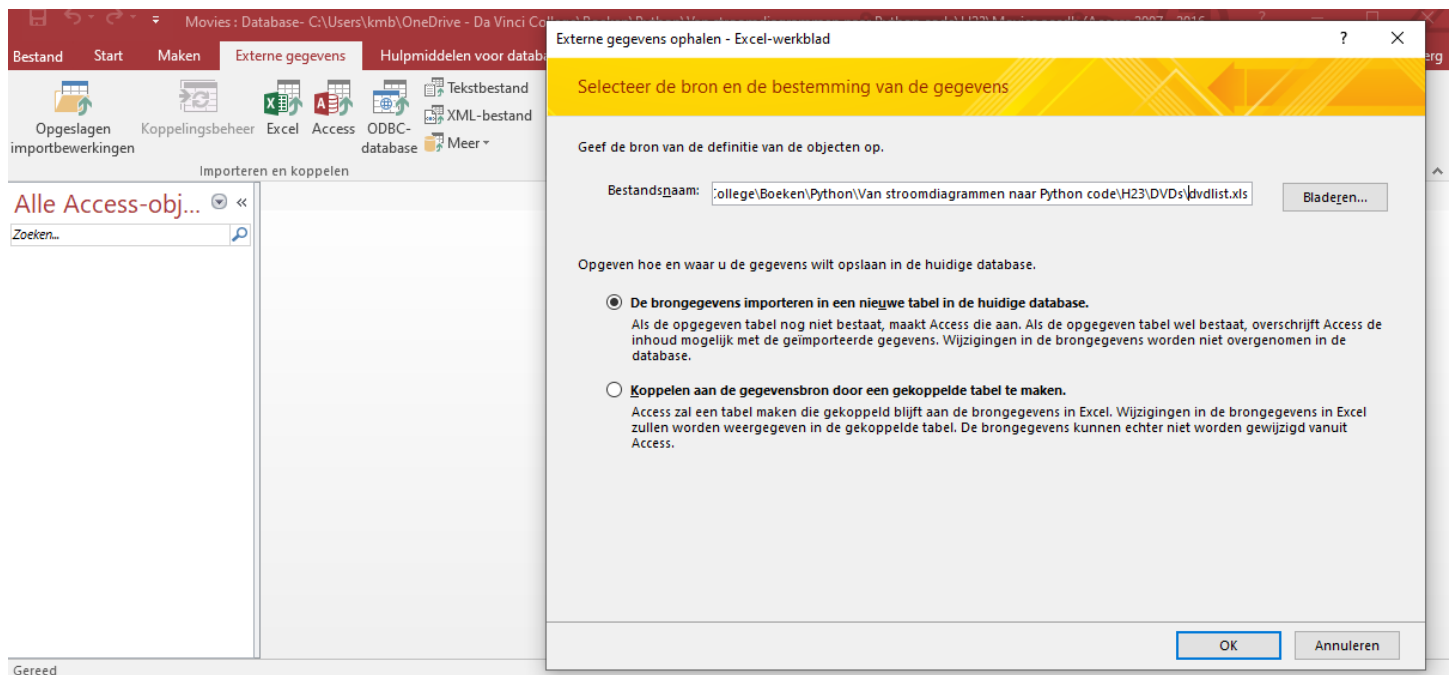
Op het tabblad “Externe gegevens” zien we een knop voor het importeren van Excel bestanden in de database. Als je deze knop niet hebt kan je ook zoeken op “importeren” zoals hieronder staat.



Let goed op dat je op **IMPORTEREN** drukt. Als je op EXPORTEREN drukt dan overschrijf je de huidige bestanden en ben je de gegevens kwijt.

We importeren de *dvdlist.xls*. De optie die eronder staat laten we ongewijzigd. We willen alle gegevens in de database brengen, en niet een koppeling naar Excel.

Nadat we op OK-drukken is de importeren enkele minuten bezig omdat het een groot bestand is.



Figuur 6: Importeren gegeven vanuit Excel

We willen alle werkbladen importeren van het Excel bestand, dus we kunnen bij het volgende scherm op {Volgende} drukken.

Wizard Werkblad importeren

Het werkbladbestand bevat meerdere werkbladen of bereiken. Welk werkblad of bereik wilt u gebruiken?

☒ Werkbladen weergeven  
☐ Benoemde bereiken weergeven

Voorbeeldgegevens voor werkblad '1'.

	DVD Title	Studio
1	!!!! Beat, Vol. 1: Shows 01 - 05	Bear Family
2	!!!! Beat, Vol. 2: Shows 06 - 09	Bear Family
3	!!!! Beat, Vol. 3: Shows 10 - 13	Bear Family
4	!!!! Beat, Vol. 4: Shows 14 - 17	Bear Family
5	!!!! Beat, Vol. 5: Shows 18 - 21	Bear Family
6	!!!! Beat, Vol. 6: Shows 22 - 26	Bear Family
7	!Hero: The Rock Opera: Live On Stage (2-Disc Collector's Edition)	ForeFront Records
8	!K7 150	!K7
9	!Women Art Revolution	Zeitgeist
10	#1 Latino Hits	BMG Music
11	#1 Serial Killer	Indican Pictures
12	#artoffline	IndiePix
13	#DigitalLivesMatter	Gravitas Ventures

< Annuleren < Vorige Volgende > Voltooien

### 5.2.1.1 Kolomkoppen

De bovenste rij van de Excel-sheet bevat de namen die we willen gebruiken als kolomnamen in de database. We kunnen de check-box dus aangevinkt dat de **eerste rij kolomnamen bevat** laten en gaan met {volgende} verder.

Wizard Werkblad importeren

Microsoft Access kan de kolomkoppen als veldnamen voor de tabel gebruiken. Bevat de eerste rij die is opgegeven kolomkoppen?

☒ Eerste rij bevat kolomkoppen

	DVD Title	Studio
1	!!!! Beat, Vol. 1: Shows 01 - 05	Bear Family
2	!!!! Beat, Vol. 2: Shows 06 - 09	Bear Family
3	!!!! Beat, Vol. 3: Shows 10 - 13	Bear Family
4	!!!! Beat, Vol. 4: Shows 14 - 17	Bear Family
5	!!!! Beat, Vol. 5: Shows 18 - 21	Bear Family
6	!!!! Beat, Vol. 6: Shows 22 - 26	Bear Family
7	!Hero: The Rock Opera: Live On Stage (2-Disc Collector's Edition)	ForeFront Records
8	!K7 150	!K7
9	!Women Art Revolution	Zeitgeist
10	#1 Latino Hits	BMG Music
11	#1 Serial Killer	Indican Pictures
12	#artoffline	IndiePix
13	#DigitalLivesMatter	Gravitas Ventures
14	#DigitalLivesMatter (Blu-ray)	Gravitas Ventures

< Annuleren < Vorige Volgende > Voltooien

### 5.2.1.2 Data typen

Het volgende scherm moeten we goed kijken of de data die we binnen halen van de datatype is dat Access als suggestie doet. Access doet de suggestie op basis van de eerste 24 regels. Dit is een punt waar het importeren fout kan gaan. Als we het verkeerde datatype kiezen, en later in het bestand wordt bijvoorbeeld een tekst gebruikt waar voor een getal als datatype gekozen is worden die regels niet geïmporteerd. Als het mis gaat moeten we de tabel weer verwijderen en het importeren overnieuw doen, Dit is dan het punt waar we goed moeten kijken of de gekozen datatypen toch niet anders waren.

Door op de kolommen te drukken kan je de data typen bekijken per kolom. Onderaan staat een horizontale scrollbar waarmee je naar de verschillende kolommen kunt scrollen.

We gaan verder door op {volgende} te drukken.

U kunt informatie opgeven voor elk veld dat u importeert. Selecteer velden in het gebied hieronder. U kunt de veldinformatie vervolgens wijzigen in het gebied Veldopties.

## Veldopties

Veldnaam:  Gegevenstype:    
 Geïndexeerd:   ☐ Veld niet importeren (Overslaan)

	Prj	Rating	Year	Genre	Aspect	UPC	DVD_ReleaseDate	ID	Timestamp	
1	5	NR	UNK	Music	1.33:1	4000127201263	10-mei-05	61689	07-jun-16	^
2	5	NR	UNK	Music	1.33:1	4000127201270	10-mei-05	61690	07-jun-16	
3	5	NR	UNK	Music	1.33:1	4000127201287	10-mei-05	61702	07-jun-16	
4	5	NR	UNK	Music	1.33:1	4000127201294	12-jul-05	65695	07-jun-16	
5	5	NR	UNK	Music	1.33:1	4000127201300	12-jul-05	65707	07-jun-16	
6	5	NR	UNK	Music	1.33:1	4000127201317	12-jul-05	65757	07-jun-16	
7	8	NR	UNK	Music	1.33:1	724359983592	29-mrt-05	60970	27-mei-06	
8	9	NR	UNK	Music	1.33:1	730003715099	12-aug-03	31992	10-feb-12	
9	9	NR	2010	Documentary	1.85:1	795975113939	20-mrt-12	220757	20-mrt-12	
10	8	NR	UNK	Music	1.33:1	743219736994	19-nov-02	22659	03-feb-05	
11	8	NR	2013	Horror	1.85:1	825284201888	23-jun-15	276862	23-jun-15	
12	5	NR	2015	Documentary	1.78:1	845637002634	13-feb-18	304832	13-feb-18	
13	9	NR	2016	Comedy	1.85:1	812034033998	26-mrt-19	315829	27-mrt-19	
14	9	NR	2016	Comedy	1.85:1	812034034001	26-mrt-19	315830	27-mrt-19	v

&lt;

&gt;

Annuleren

&lt; Vorige

Volgende &gt;

Voltooien

### 5.2.1.3 Primary-key

Omdat we de data nog niet kennen wat in de tabellen staat, weten we ook niet zeker of de kolom ID wel unieke nummers heeft. We kunnen dit uitproberen door deze gok te maken, we kunnen ook tijdelijk een id veld aanmaken. De tabel die we importeren gaan we toch als bron gebruiken voor andere tabellen, dus als we nu een extra id-veld aanmaken kan dat geen kwaad.

Wizard Werkblad importeren

Het is raadzaam een primaire sleutel te definiëren voor de nieuwe tabel. Met een primaire sleutel heeft elke record in de tabel een unieke aanduiding, zodat gegevens sneller kunnen worden opgehaald.

☒ Primaire sleutel van Access gebruiken  
☐ Zelf primaire sleutel bepalen   
☐ Geen primaire sleutel

Id1	DVD Title	Studio
1	!!!! Beat, Vol. 1: Shows 01 - 05	Bear Family
2	!!!! Beat, Vol. 2: Shows 06 - 09	Bear Family
3	!!!! Beat, Vol. 3: Shows 10 - 13	Bear Family
4	!!!! Beat, Vol. 4: Shows 14 - 17	Bear Family
5	!!!! Beat, Vol. 5: Shows 18 - 21	Bear Family
6	!!!! Beat, Vol. 6: Shows 22 - 26	Bear Family
7	!Hero: The Rock Opera: Live On Stage (2-Disc Collector's Edition)	ForeFront Record
8	!K7 150	!K7
9	!Women Art Revolution	Zeitgeist
10	#1 Latino Hits	BMG Music
11	#1 Serial Killer	Indican Pictures
12	#artoffline	IndiePix
13	#DigitalLivesMatter	Gravitas Venture
14	#DigitalLivesMatter (Blu-ray)	Gravitas Venture

Annuleren < Vorige Volgende > Voltooien

### 5.2.1.4 Naam tabel

Nadat we op {volgende} hebben gedrukt komen we bij het laatste scherm aan. Hier kunnen we een naam ingeven waar de gegevens die we importeren in komen te staan.

Wizard Tekst importeren

De wizard beschikt nu over alle informatie die nodig is om de gegevens te importeren.

Importeren in tabel:

☐ Mijn tabel door een wizard laten analyseren nadat de gegevens zijn geïmporteerd

Geavanceerd... Annuleren < Vorige Volgende > Voltooien

### 5.2.1.5 Controle import

We hebben nu de gegevens geïmporteerd vanuit het Excel bestand naar onze database. Nu gaan we controleren of alle gegevens zijn binnengehaald. Om dit snel te doen openen we de tabel DVD en kijken hoeveel rijen (records) erin zitten. Als we de tabel openen gaan we met <CTRL> <pijlte naar beneden> naar het laatste record. Zoals op de onderstaande afbeelding staat weergegeven zijn dat 47.704 rijen.

47697	'Til Death Do U	Gravitas Ventu	Out	2.0	LBX, 16:9, BLU-	\$16,99	PG-13
47698	'Til Death: The	Sony Pictures	Out	5.1	LBX, 16:9	\$38,99	NR
47699	'Til The River R	PBS	Out	2.0	4:3	\$14,99	NR
47700	'Til There Was	Paramount	Discontinued	5.1	LBX, 16:9	\$19,99	PG-13
47701	'Til There Was	Paramount	Discontinued	5.1	LBX, 16:9	\$29,99	PG-13
47702	'Tis Autumn: T	Outsider Pictu	Out	2.0	LBX	\$19,99	NR
47703	'Tis The Seaso	Hallmark Enter	Out	2.0	LBX, 16:9	\$9,95	NR
47704	'WOW' Factor	Healthy Learni	Discontinued	2.0	4:3	\$49,95	NR
*	(Nieuw)						

Record: 14 47704 van 47704 Geen filter Zoeken

Als we naar het Excel bestand kijken zien we dat we alle regels van het eerste tabblad hebben binnengehaald. We moeten dus deze stappen herhalen voor ieder tabblad.

Kan je uitleggen waarom er in Excel een rij meer is dan in Acces!

47701	'Til There	Paramou	Discontin	5.1	LBX, 16:9	\$19,99	PG-13	1997	Comedy	1.85:1	0973
47702	'Til There	Paramou	Discontin	5.1	LBX, 16:9	\$29,99	PG-13	1997	Comedy	1.85:1	0973
47703	'Tis	Outsider	Out	2.0	LBX	\$19,99	NR	2006	Documen	1.78:1	8978
47704	'Tis The	Hallmark	Out	2.0	LBX, 16:9	\$9,95	NR	2015	Comedy	1.78:1	8834
47705	'WOW'	Healthy	Discontin	2.0	4:3	\$49,95	NR	2007	Exercise	1.33:1	8270
47706											
47707											
47708											
47709											
47710											

#-b c-e f-i j-n o-s t-z + Gereed

We zien dat er 6 tabbladen zijn die op deze manier allemaal geïmporteerd zouden moeten worden. Dit is heel veel werk, en we komen er nu wel achter dat dit niet de handigste manier is om onze database te vullen. **We kunnen deze database verwijderen in de volgende paragraaf gaan verder met het importeren van een tekstbestand.**

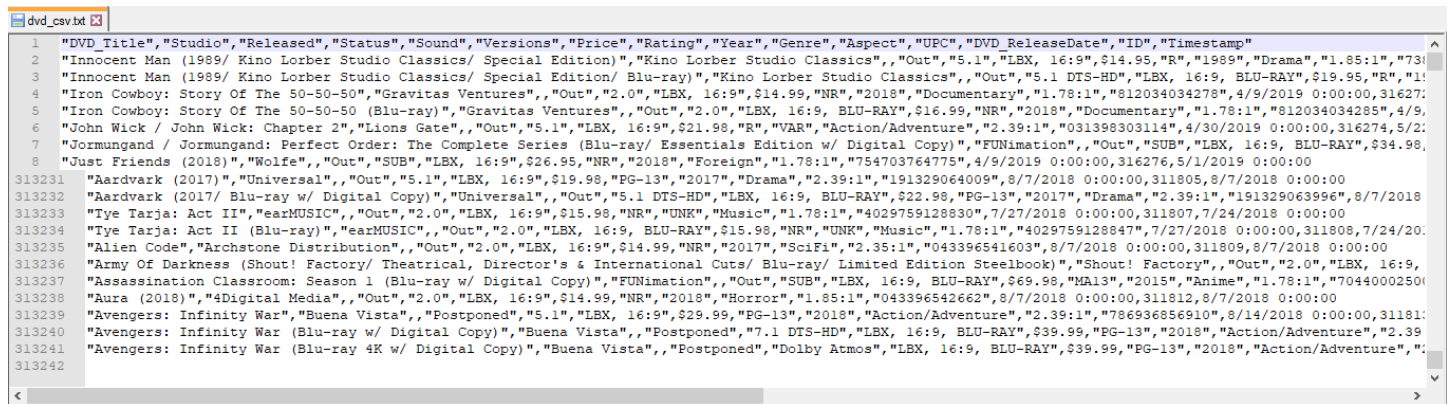


## 5.2.2 Importeren van CSV-bestanden

We beginnen met het opstarten van Microsoft Access, en maken een nieuwe lege database aan. Deze database noemen we Movies

### 5.2.2.1 CSV-bestanden

Een CSV oftewel (comma separated values) is een tekstbestand met daarin alle waarden gescheiden door een teken. Dit is volgens de naam een komma, maar heel vaak is het een puntkomma (<;> semi-column). Als we dit bestand bekijken zien we dat alle 313.240 films achter elkaar staan in een lijst. De eerste regel is ook weer de naam van de kolommen. Als het vorige importeren niet gelukt is, of je was zo slim om al wat vooruit te lezen dan kan je op de volgende manier de database in een keer inlezen. Als het al gelukt is met het Excel bestand, is het handig om deze lijst ook in te lezen onder een andere tabel-naam. Op de eerste regel kan je zien dat de verschillende velden inderdaad gescheiden zijn door een komma. Soms die je twee komma's achter elkaar staan zoals op regel 2. Dan is er voor die kolom geen gegevens opgenomen.



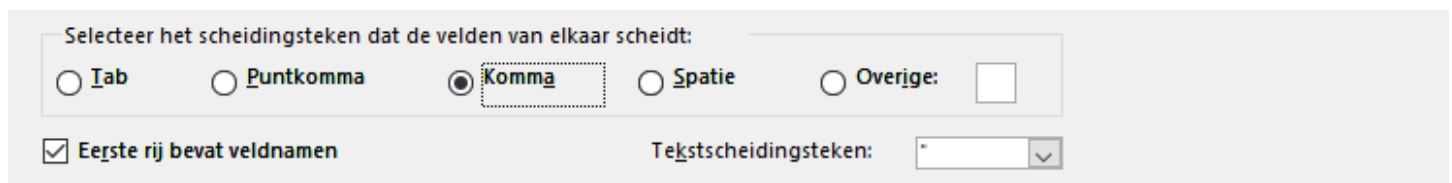
In Access staat bij {Externe gegevens} de knop {tekstbestand}, waar we nu op drukken.

### 5.2.2.2 Scheidingstekens

We krijgen nu een scherm waar nu het .csv bestand geselecteerd wordt. De schermen zien er een hetzelfde uit als bij het importeren van een Excel bestand, maar er zijn wel wat verschillen.



Bij een Excel bestand is het duidelijk waar de ene waarde eindigt, en de volgende begint. Bij een CSV-bestand moet dat expliciet verteld worden.



Ook hier kunnen we aangeven dat de eerste rij de veldnamen bevat, en we hebben in het bestand gezien dat om de teksten aanhalingstekens staan <">. Dit is niet altijd zo, maar dat kan in {Tekstscheidingsteken} ingevoerd worden.

### 5.2.2.3 Data typen

Bij een CSV-bestand moet je nog beter letten of de goede datatypen geselecteerd zijn voor de velden. Als Excel geïmporteerd wordt staat deze informatie meestal al in Excel, maar een CSV heeft geen extra informatie. Het ID-veld is een veld dat aangemerkt wordt als "Lange integer" en we hopen maar dat ergens in de 31.3240 regels niet een tekst staat.

Wizard Tekst importeren

U kunt informatie opgeven voor elk veld dat u importeert. Selecteer velden in het gebied hieronder. U kunt de veldinformatie vervolgens wijzigen in het gebied Veldopties.

Veldopties

Veldnaam:  Gegevenstype:

Geïndexeerd:  ☐ Veld niet importeren (Overslaan)

Rating	Year	Genre	Aspect	UPC	DVD ReleaseDate	ID	Timestamp
R	1989	Drama	1.85:1	738329235635	4/2/2019 0:00:00	316270	5/1/2019 0:00:00
R	1989	Drama	1.85:1	738329235642	4/2/2019 0:00:00	316271	5/1/2019 0:00:00
NR	2018	Documentary	1.78:1	812034034278	4/9/2019 0:00:00	316272	5/1/2019 0:00:00
NR	2018	Documentary	1.78:1	812034034285	4/9/2019 0:00:00	316273	5/1/2019 0:00:00
R	VAR	Action/Adventure	2.39:1	031398303114	4/30/2019 0:00:00	316274	5/22/2019 0:00:00
NR	2012	Anime	2.35:1	704400020384	5/7/2019 0:00:00	316275	5/8/2019 0:00:00
NR	2018	Foreign	1.78:1	754703764775	4/9/2019 0:00:00	316276	5/1/2019 0:00:00
MA13	2018	Anime	1.78:1	704400021046	5/7/2019 0:00:00	316277	5/8/2019 0:00:00
NR	2015	Anime	1.78:1	704400020391	5/7/2019 0:00:00	316278	5/8/2019 0:00:00
NR	1998	Foreign	1.37:1	760137233886	4/30/2019 0:00:00	316279	5/1/2019 0:00:00
MA13	1987	Anime	1.33:1	875707640022	3/26/2019 0:00:00	316280	5/1/2019 0:00:00
NR	1957	SciFi	2.35:1	738329237141	4/23/2019 0:00:00	316281	5/1/2019 0:00:00
NR	2009	TV Classics	1.78:1	883929532919	4/2/2019 0:00:00	316282	5/1/2019 0:00:00
PG	2019	Animation	2.35:1	883929645619	5/7/2019 0:00:00	316283	5/8/2019 0:00:00


Geavanceerd... Annuleren < Vorige Volgende > Voltooien

We geven de database een naam, en we kunnen gaan importeren.

### 5.2.2.4 Importeer fout oplossen

We kunnen onderstaande foutmelding kunnen krijgen.

Wizard Tekst importeren

 Het veldscheidingsteken uit de specificatie voor het tekstbestand komt overeen met het decimaalteken of het tekstscheidingsteken.

OK

Dit komt omdat we importeren met een komma <,> als scheidingsteken, en de komma ook wordt gebruikt als decimaal scheidingsteken in Nederland. Dit probleem heeft te maken met de kolom Price. Gelukkig zijn de prijzen allemaal in dollars, met een Amerikaans scheidingsteken met een punt <.>.

We gaan een paarschermen terug en drukken op de knop <Geavanceerd> waar we het decimaalsymbool kunnen wijzigen van een komma <,> naar een punt <.>.

Wizard Tekst importeren

U kunt informatie opgeven voor de import, die u vervolgens wijzigen in het gebied Veldopties.

**Veldopties**

Veldnaam: DVD\_Title

Geïndexeerd: Nee

**Dvd\_csv Importspecificatie**

Bestandsindeling: ☒ Gescheiden ☐ Vaste breedte

Scheidingsteken veld: ,

Tekstscheidingsteken: "

Taal: Nederlands

Codetabel: West-Europees (DOS)

**Datums, tijden en getallen**

Datumvolgorde: DMJ ☒ Jaar met vier cijfers

Datumscheidingsteken: - ☐ Voorloophulpnullen in datums

Tijdscheidingsteken: : Decimaalsymbool: .

**Veldgegevens:**

Veldnaam	Gegevenstype	Geïndexeerd	Oversla
DVD_Title	Korte tekst	Nee	<input type="checkbox"/>
Studio	Korte tekst	Nee	<input type="checkbox"/>
Released	Korte tekst	Nee	<input type="checkbox"/>
Status	Korte tekst	Nee	<input type="checkbox"/>
Sound	Korte tekst	Nee	<input type="checkbox"/>
Versions	Korte tekst	Nee	<input type="checkbox"/>
Price	Korte tekst	Nee	<input type="checkbox"/>
Rating	Korte tekst	Nee	<input type="checkbox"/>
Year	Korte tekst	Nee	<input type="checkbox"/>

Geavanceerd...

Annuleren < Vorige Volgende > Voltooien

We hadden er hier ook voor kunnen kiezen om bepaalde kolommen helemaal niet te gebruiken. Die keuze maken we nu later. Druk op {OK} en vervolgens op {Voltooien} om de import te starten.

#### 5.2.2.5 Controle import

Als we klaar zijn met importeren zien we het volgende resultaat. Naast de tabel die we wilden aanmaken is er nog een tabel bijgekomen met de naam DVD\_CSV\_IMPORTFOUTEN. Als we die tabel bekijken zien we waar er fouten zijn gemaakt bij het importeren. Als we de rijen bekijken zien we dat het fout gaat bij "Timestamp" en bij DVD\_ReleaseDate. Die "Timestamp" kunnen we wel negeren, maar DVD\_ReleaseDate willen we eigenlijk wel correct hebben. De vraag is waar het mis kan zijn gegaan.

Alle Access-obj...

Zoeken...

**Tabellen**

- DVD
- dvd\_csv\_importfouten

Fout	Rij	Veld
Typeconversie	5	DVD_ReleaseDate
Typeconversie	5	Timestamp
Typeconversie	10	DVD_ReleaseDate
Typeconversie	11	DVD_ReleaseDate
Typeconversie	12	DVD_ReleaseDate

Als we in de dvd-tabel kijken zien we inderdaad op rij 5, 10, 11 en 12 de datums niet zijn ingevuld.

Genre	Aspect	UPC	DVD_ReleaseDate	ID	Timestamp	klik om titel toe te voegen
Drama	1.85:1	738329235635	4-2-2019	316270	5-1-2019	
Drama	1.85:1	738329235642	4-2-2019	316271	5-1-2019	
Documentary	1.78:1	812034034278	4-9-2019	316272	5-1-2019	
Documentary	1.78:1	812034034285	4-9-2019	316273	5-1-2019	
Action/Adventure	2.39:1	031398303114		316274		
Anime	2.35:1	704400020384	5-7-2019	316275	5-8-2019	
Foreign	1.78:1	754703764775	4-9-2019	316276	5-1-2019	
Anime	1.78:1	704400021046	5-7-2019	316277	5-8-2019	
Anime	1.78:1	704400020391	5-7-2019	316278	5-8-2019	
Foreign	1.37:1	760137233886		316279	5-1-2019	
Anime	1.33:1	875707640022		316280	5-1-2019	

Als we naar het bestand kijken dan zien we op regel 6 (want de eerste regel was de header) dan zien we dat daar de datum 4/30/2019 staat. Dit is ook weer een Amerikaanse schrijfwijze van 30 april. We moeten de import weer overnieuw doen om dit ook te corrigeren.

```

1 :,"Aspect","UPC","DVD_ReleaseDate","ID","Timestamp"
2 :,"Out","5.1","LBX, 16:9","$14.95","R","1989","Drama","1.85:1","738329235635",4/2/2019 0:00:00,316270,5/1/2019 0:00:00
3 Studio Classics",,"Out","5.1 DTS-HD","LBX, 16:9, BLU-RAY","$19.95","R","1989","Drama","1.85:1","738329235642",4/2/2019 0:00:00,316271,5/1/2019 0:00:00
4 :NR","2018","Documentary","1.78:1","812034034278",4/9/2019 0:00:00,316272,5/1/2019 0:00:00
5 :,"BLU-RAY","$16.99","NR","2018","Documentary","1.78:1","812034034285",4/9/2019 0:00:00,316273,5/1/2019 0:00:00
6 :Action/Adventure","2.39:1","031398303114",4/30/2019 0:00:00,316274,5/22/2019 0:00:00
7 w/ Digital Copy)","FUNimation",,"Out","SUB","LBX, 16:9, BLU-RAY","$34.98","NR","2012","Anime","2.35:1","704400020384",5/7/2019 0:00:00,316275,5/8/2019 0:00:00
8 :,"754703764775",4/9/2019 0:00:00,316276,5/1/2019 0:00:00
9 :tion",,"Out","SUB","LBX, 16:9, BLU-RAY","$64.98","MAL3","2018","Anime","1.78:1","704400021046",5/7/2019 0:00:00,316277,5/8/2019 0:00:00
10 :,"SUB","LBX, 16:9, BLU-RAY","$29.98","NR","2015","Anime","1.78:1","704400020391",5/7/2019 0:00:00,316278,5/8/2019 0:00:00
11 :AY","$49.95","NR","1998","Foreign","1.37:1","760137233886",4/30/2019 0:00:00,316279,5/1/2019 0:00:00
12 :Out","SUB","4:3, BLU-RAY","$99.95","MAL3","1987","Anime","1.33:1","875707640022",3/26/2019 0:00:00,316280,5/1/2019 0:00:00
13 :,"SUB","16:9, BLU-RAY","$26.95","NR","1987","Anime","2.35:1","738329237141",4/22/2019 0:00:00,316281,5/1/2019 0:00:00

```

Deze keer bij Geavanceerd de datumvolgorde op de Amerikaanse manier. **Het veld TimeStamp kunnen we aanmerken dat we die willen overslaan.**

Datumvolgorde:

MDJ

☒ Jaar met vier cijfers
☐ Voorlooppnullen in datums

Datumscheidingsteken:

-

Tijdscheidingsteken:

:

Decimaalsymbool:

.

Als we nu importeren zien we dat alle regels goed geïmporteerd worden.

#### 5.2.2.6 Andere CSV-bestanden importeren

Dit moeten we ook doen voor de volgende bestanden die we alle in een aparte tabel zetten.

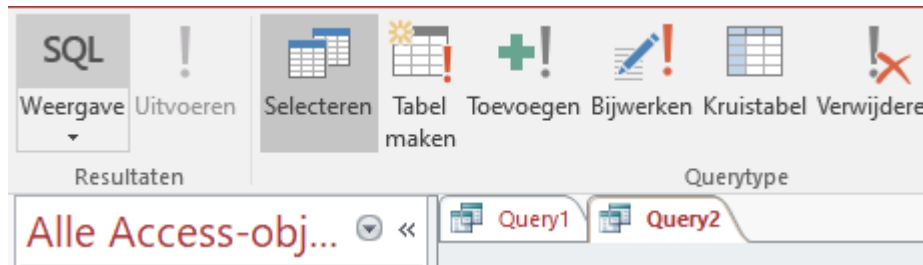
- Actors\_Index.csv
- Actors\_Names.csv
- Directors\_Index.csv
- Directors\_Names.csv

## 5.3 Corrigeren van de dvd-tabel

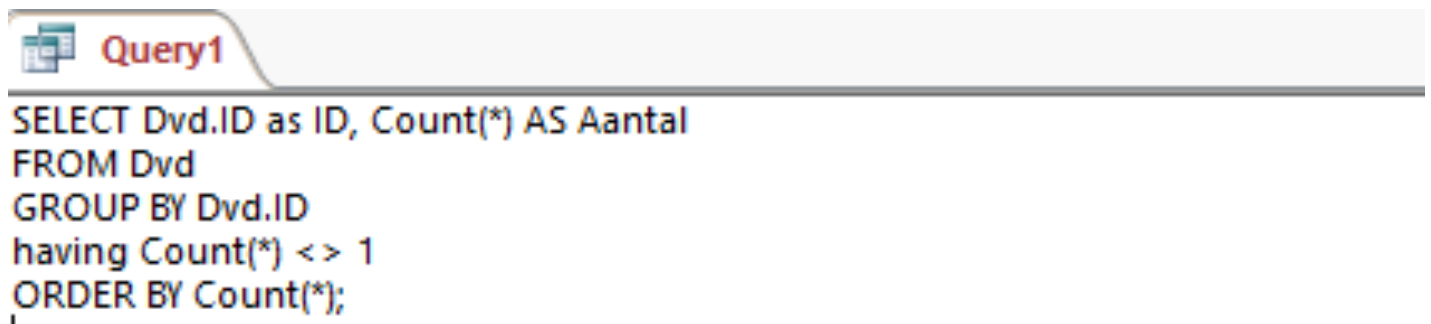
### 5.3.1.1 Controle ID-veld

We kunnen nu de tabellen bekijken in Access. Wat we als eerste willen weten is of het ID-veld dat in DVD staat ook een echt uniek veld is. Dit kunnen we met onderstaande SQL-Statement.

Ga naar {Maken} -> {QueryOntwerp}. Sluit het venster af met tabellen en druk op SQL-Weergave.



Dit is een veel moeilijker statement dan dat je tot nu toe hebt gehad. Het gaat er hier om dat je hem kan lezen, je hoeft hem (nog) niet zelf te kunnen maken.



Figuur 7: SQL om te kijken of elementen uniek zijn

Op de eerste regel staat SELECT, dat betekent dat die velden getoond moeten worden. Van de tabel DVD wil je het ID laten zien, en je wilt tellen hoe vaak dit ding voorkomt.

Op de tweede regel zie je welke tabel je wilt bekijken.

Op de derde regel zeg je waarop je wilt groeperen, zodat je ieder nummer maar 1x te zien krijgt. Dit is gelijk die COUNT voor de eerste regel, hoe vaak het ID voorkomt als je het samenvoegt.

De vierde regel zegt als je optel (count) wil ik alleen de resultaten zien die ongelijk zijn aan 1.

De laatste regel zegt dat je de resultaten wilt sorteren op aantal.

Doordat we niets te zien krijgen als resultaat betekent dat het ID-veld een echte ID is, en voor iedere film uniek is. We kunnen van dit veld dus de primary-key maken.

Als we naar het ontwerp van de DVD gaan kunnen we het ID-veld verplaatsen van de onderste plek naar de bovenste plek. Met de rechtermuisknop kunnen we het ID-veld als primary-key aanmaken.

Het veld Id1 kunnen we met de rechtermuisknop verwijderen. Als we nu de tabel opslaan hebben we de tabel weer wat beter gemaakt.

Alle Access-obj...

Zoeken...

Tabellen

- Actors\_
- Actors\_
- Directo
- Directo
- Dvd

Primaire sleutel

Knippen

Kopiëren

Plakken

Rijen invoegen

Rijen verwijderen

Eigenschappen

Veldnaam	Gegevenstype
Id1	AutoNummering
ID	Numeriek
DVD_Title	Korte tekst
Studio	Korte tekst
Released	Korte tekst
Status	Korte tekst
Sound	Korte tekst
Versions	Korte tekst
Price	Korte tekst
Rating	Korte tekst
Year	Korte tekst
Genre	Korte tekst

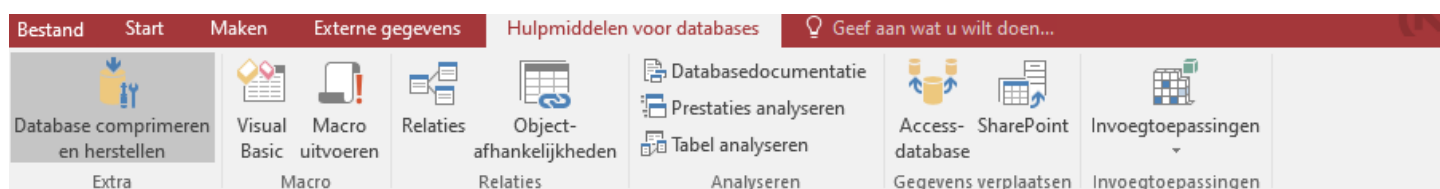
We kunnen de volgende kolommen weghalen:

- Id1
- Released
- Status
- Price
- Sound
- Versions
- Aspects
- UPC
- TimeStamp
- DVD\_Release

#### 5.3.1.2 Database comprimeren en herstellen

De database heeft nu heel wat toegevoegd, en verwijderd. Bij een Access database is het erg verstandig om dan regelmatig de database even op te schonen. Dan zijn alle dingen die nog ergens verdwaald zijn achter gebleven op te ruimen. Dit doe je bij Database Comprimeren en herstellen.

Je hoeft geen verdere stappen te ondernemen, als het comprimeren klaar is kan de database gewoon weer verder gebruikt worden.



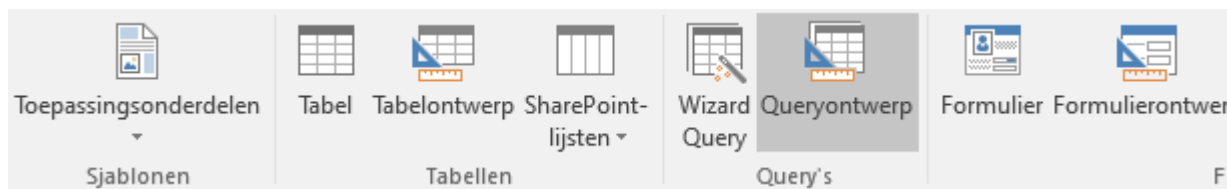


## 5.4 Maken van code-tabellen

Als we naar de tabellen in de database kijken zien we dat er een paar tabellen zijn die informatie bevatten die vaker terugkomen. Een studio is het bedrijf dat films maakt, en een studio maakt dus meerdere films. We willen dus een tabel maken met alle studio's en dan een Foreign-key in de dvd-tabel. Dit willen we doen we doen voor {Studio}, {Rating} en {Genre}.

ID	DVD_Title	Studio	Rating	Year	Genre	klik om titel toe te voegen
1	1-900 (1994)	Fox Lorber	R	1994	Foreign	
2	10 (Snapper Case)	Warner Brothe	R	1979	Comedy	
3	10 Things I Hate About You (1999)	Buena Vista	PG-13	1999	Comedy	
4	100 Girls By Bunny Yeager (CAV/ Special E	CAV	NR	1999	Late Night	
5	100 Years Of Horror (Passport Video/ Old	Passport Video	NR	1996	Documentary	
6	101 Dalmatians (1961)	Buena Vista	G	1961	Animation	
7	101 Dalmatians (1996)	Buena Vista	G	1996	Family	
8	10th Kingdom (Hallmark Entertainment/ C	Hallmark Enter	NR	2000	SciFi	
9	12 Monkeys (1995/ Universal/ DTS)	Universal	R	1995	SciFi	
10	12 Monkeys (1995/ Universal/ Snerial Edit	Universal	R	1995	SciFi	

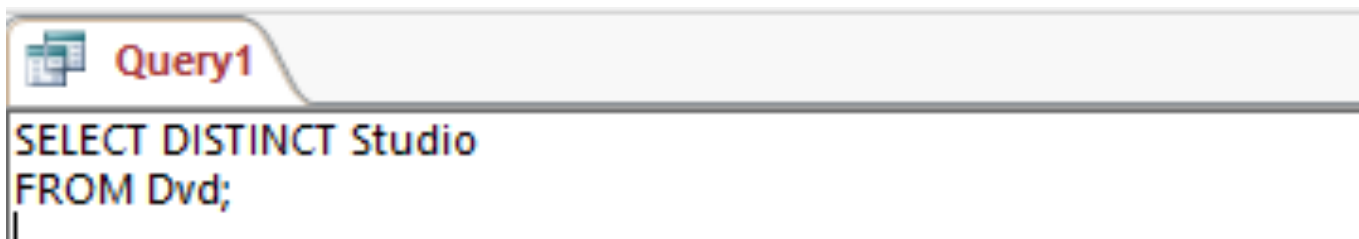
We gaan naar {Queryontwerp} in het tabblad Maken, we voegen DVD daartoe en sluiten het scherm. Vervolgens gaan we naar Weergave, en selecteren SQL-Weergaven.



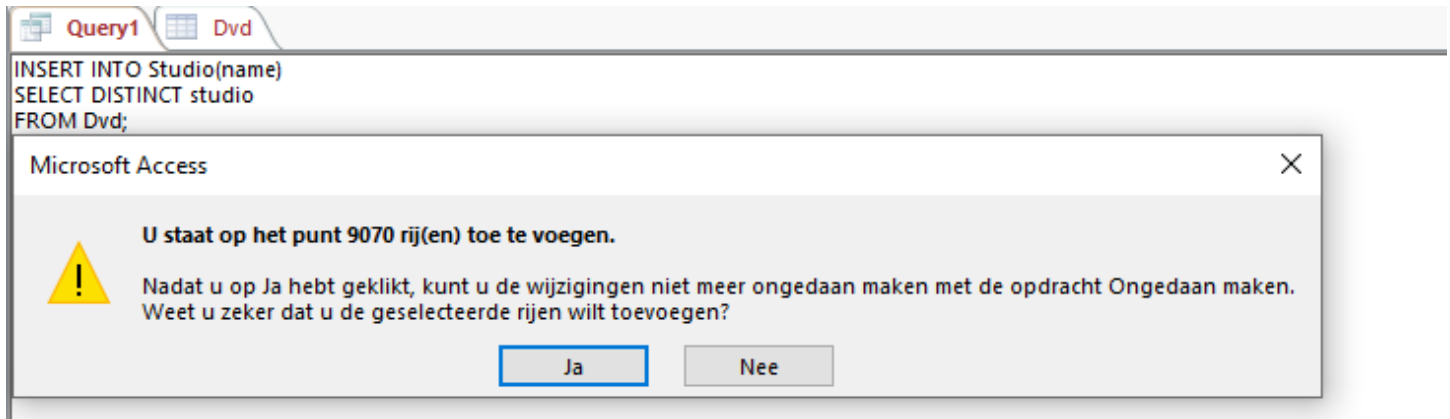
We willen nu alle unieke velden van de kolom Studio in een aparte tabel zetten. We maken hiervoor eerst een nieuwe tabel aan via {tabelontwerp} met de naam Studio. Je krijg dan onderstaand scherm te zien.

Alle Access-obj...		Query1 Dvd Studio	
Zoeken...		Veldnaam	Gegevenstype
Tabellen		IdStudio	AutoNummering
Actors_Index		name	Korte tekst
Actors_Names			
Directors_Index			
Directors_Names			
Dvd			
Studio			
Algemeen Opzoeken			
Veldlengte	255		
Notatie			
Invoermasker			
Bijschrift			
Standaardwaarde			
Validatieregels			
Validatietekst			
Vereist	Ja		
Lengte nul toestaan	Nee		
Geïndexeerd	Ja (Duplicaten OK)		
Unicode-compressie	Ja		
IME-modus	Geen besturingselement		
IME-zinmodus	Geen		
Tekstuitlijning	Algemeen		

Deze tabel gaan we vullen met gegevens. Eerst gaan we kijken welke studios er allemaal zijn. Dit doen we door de volgende SQL-statement. We hadden al gezien dat we met `SELECT * from DVD` alle elementen krijgen. Op de plek van de `*` zetten we de naam van de kolom die we willen zien. Het woord `DISTINCT` ervoor betekend dat we geen dubbeln willen zien.



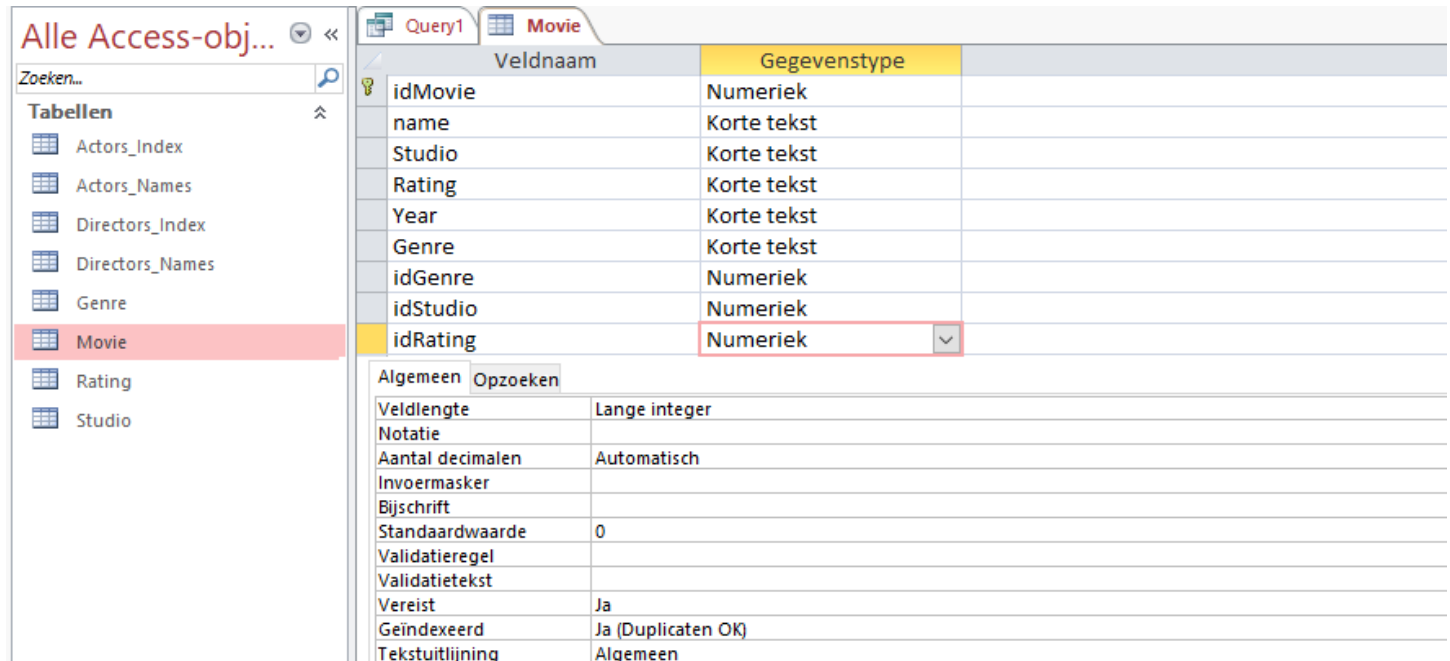
Het resultaat van die regel stoppen we in de tabel Studio die we net gemaakt hebben.



Figuur 8: INSERT-data van de ene tabel naar een andere

Dit doen we ook voor de tabellen Rating en Genre.

De volgende stap is dat we de tabel DVD hernoemen naar Movie, en de velden wijzigen zoals hieronder. Er komen dus 3 velden bij, id wordt gewijzigd naar idMovie. DVD\_Title hernoemen we naar name.





## 5.5 Foreign-keys aanmaken in de tabel

We gaan nu de id-velden die we in de vorige paragraaf hebben gemaakt invullen met de goede waarden. Die waarden zijn dus de primary-key waarden van de codetabellen.

```
SELECT Movie.Genre, Genre.name, Movie.idGenre, Genre.idGenre
FROM Movie, Genre
where Movie.Genre = Genre.Name
```

Query1			
Genre	name	Movie.idGenre	Genre.idGenre
Foreign	Foreign		15
Comedy	Comedy		6
Comedy	Comedy		6
Late Night	Late Night		19
Documentary	Documentary		9
Animation	Animation		3
Family	Family		12

We moeten dus de namen koppelen, en de daarbij behorende nummers in de tabel zetten. We voeren de volgende SQL-statement uit, en we zien dan dat de regels gevonden kunnen worden van de ene tabel in de andere. We moeten alleen nog de waarden updaten van de ene naar de andere. Als we onderstaande SQL-statement hebben uitgevoerd dan kunnen we zien dat in de Movies tabel het veld idGenre overal is ingevuld.

```
update Movie, Genre
set Movie.idGenre = Genre.idGenre
where Movie.Genre = Genre.name
```

Ditselfde moeten we nu ook doen voor de tabellen {Rating} en {Studio}.

Als je niet iedere keer 313.240 rijen bijwerkt heb je waarschijnlijk iets verkeerd gedaan.

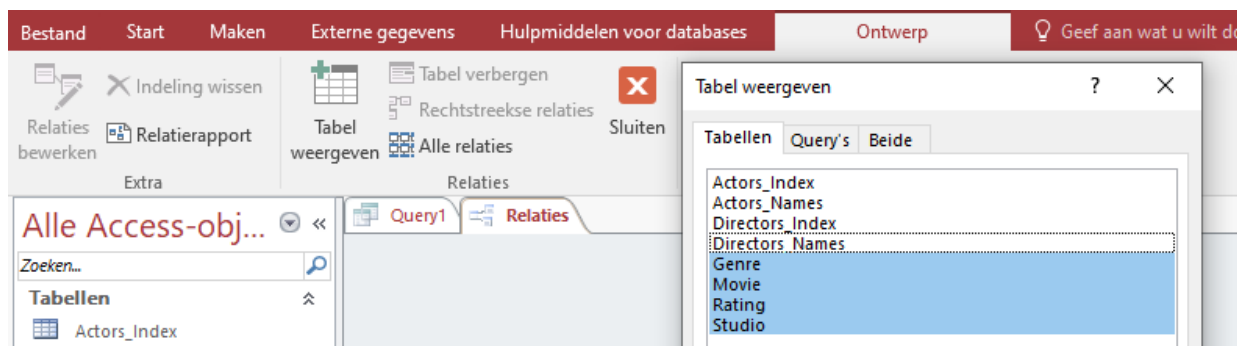
## 5.6 Opschonen van de database

We hebben nu de goede waarden in de id-velden gezet. Nu kunnen we de database een eind opruimen door de kolommen {Rating}, {Genre} en {Studio} uit de Movies tabel te halen. Dit doen we weer in de ontwerpweergave van de tabel. Op de kolommen die we willen verwijderen selecteren we met de rechtermuisknop: Rijen verwijderen.

De database is nu al een heel stuk kleiner, en we hebben nog steeds dezelfde informatie in de database staan.

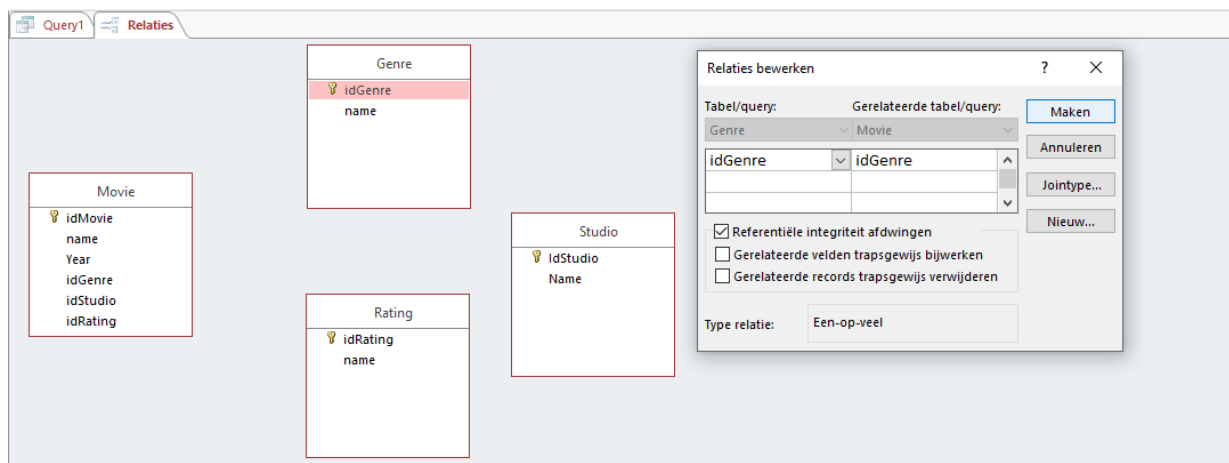
Query1		Movie	
Veldnaam		Gegevenstype	
name		Korte tekst	
Year		Korte tekst	
idGenre		Numeriek	
idStudio		Numeriek	
idRating		Numeriek	

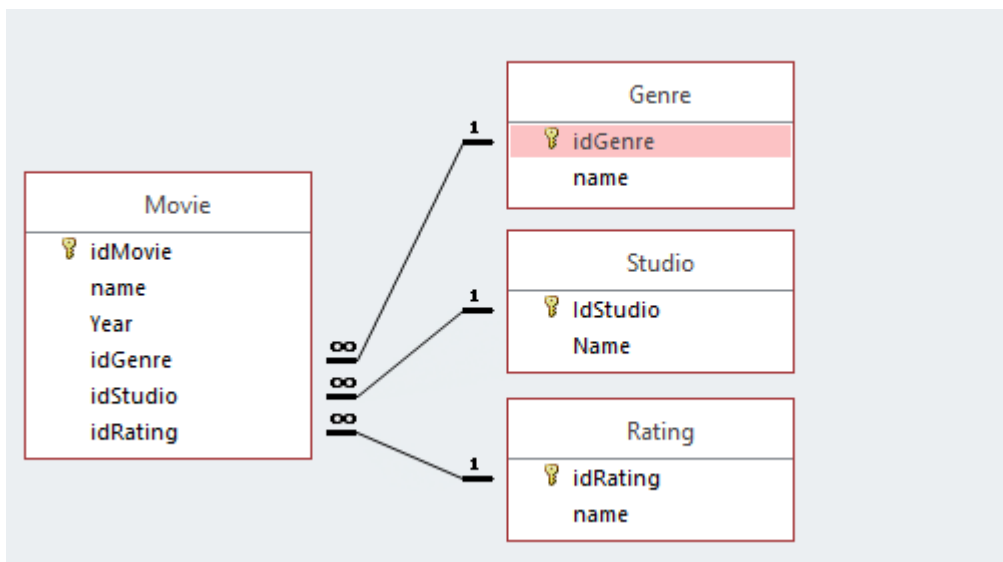
We gaan nu naar het tabblad Ontwerp en gaan de relaties aanmaken voor de database. We selecteren de tabellen {Genre}, {Movies}, {Rating} en {Studio}, drukken op {Toevoegen} en sluiten het scherm.



We slepen de idGenre van de tabel {Genre} naar {Movies} op de plek idGenre en dan zien we de pop-up tevoorschijn komen om de relaties te bewerken. We vinken weer de {Referentie integriteit afdwingen} aan. Dit doen we ook voor de andere tabellen.

LET OP: je begint met slepen van de code tabel naar de {Movies} tabel toe.





Figuur 9: Relatiediagram Movies database

### 5.7 {year} veld correct maken

Als we naar het jaar van de films kijken dan zien we de volgende jaren.

Select distinct year  
FROM Movie  
Order By Year

year			
UNK			
211			
VAR			
199			
1939			
1947			
name	Korte tekst		
Year	Korte tekst		
idGenre	Numeriek		

We kunnen zien dat {year} een korte tekst is, en geen getal. We gaan ervan uit dat 211 en 199 fouten zijn in de database. In die twee jaartallen zijn er niet heel veel films uitgekomen.

Als we het datatype omzetten naar numeriek, dan gaat dat niet goed met de waarden voor VAR en UNK. We kiezen ervoor dat dit niet erg is, en zien wel wat er gaat gebeuren met die waarden.

Als we toch het datatype omzetten, kunnen we gelijk de hoofdletter Y van Year wijzigen in een kleine letter, zodat deze naam een correcte schrijfwijze heeft.

Als we nu naar de tabel kijken dan zien we dat waar eerst UNK en VAR stond nu niets in ingevuld.

3376	Land Before Time (Old Version/ 1999 Rele	1988	3	8342	2
3377	Landlady	1998	17	8165	12
3378	Landmarks Of Early Film		9	3900	8
3379	Landmarks Of Early Film #2: The Magic Of I		9	3900	8
3380	Standard Deviants: Language Basics (2-Pac	2000	28	1526	8
3381	Lansky	1999	22	3598	12
3382	Ian Dancin	1995	19	3900	8

Er zijn twee soorten van niets in een database. Niets is wat anders dan bijvoorbeeld een lege string. Niets wordt ook wel {NULL} genoemd en betekent eigenlijk onbekend. De velden waar niets is ingevuld voor jaar betekent in een database dat het jaar onbekend is.

We hebben ook de {Directors\_names} en {Actors\_Names} tabellen geïmporteerd. Deze twee tabellen gaan we eerst hernoemen naar Director en Actor. Ook al staan er meerdere elementen in een tabel we gebruiken altijd de enkelvoudige naam. De velden in de tabellen hernoemen we ook. Actor\_id wordt idActor en Actor wordt name. Van de Director tabel hernoemen we het veld Director naar Name en Director in idDirector.

Comprimeer je database nogmaals zodat deze weer opgeruimd is.

## 5.8 Directors en Actors, Cross-references

We willen directors en actors aan de films koppelen. Maar hoeveel acteurs mogen maximaal meedoen. We hebben een tabel met alle acteurs, en kunnen 10 acteur velden in de Movies tabel bijvoegen. Maar wat doen we met de 11<sup>e</sup> acteur. Als we dit in python zouden oplossen dan zouden we een lijst maken waar acteurs ingezet kunnen worden. Met een database doe je dit met een cross-reference tabel.

We hebben een tabel met Movies en een tabel met Actors. Nu willen we een tabel die opslaat welke acteurs er in welke films gespeeld hebben. We gaan ervan uit dat 1 film meerdere acteurs heeft en we noemen de tabel MovieActorsCR, waar de CR staat voor cross-reference. Zo kunnen we gelijk aan de tabel naam zien wat de functie is van de tabel. De Actors\_index tabel hernoemen we dan naar die naam. Dit doen we ook zo voor de Directors\_index tabel, deze noemen we MovieDirectorCR.

De velden in de tabellen passen we ook gelijk aan naar de nieuwe namen die we nu hebben.

MovieDirectorCR		
Veldnaam	Gegevenstype	
ID	Numeriek	
idDirector	Numeriek	
idMovie	Numeriek	

MovieDirectorCR			MovieActorCR		
Veldnaam	Gegevenstype		Veldnaam	Gegevenstype	
idActor	Numeriek				
idMovie	Numeriek				

Als we naar de film Iron man kijken uit 2008 dan zien we de volgende gegevens.

142334	Iron Man (2008/ Paramount)	2008	1	6027	11
142335	Vinny Sincere 2		10	7101	15

Met de primary-key van de film kunnen we nu opzoeken in de cross-reference tabel welke acteurs erin hebben gespeeld.

Movie	Query1
select * from MovieActorCR where idMOvie = 142334	

Movie		Query1	
	ID	idActor	idMovie
	4716	83	142334
	23188	494	142334
	96765	2017	142334
	120135	2479	142334
	164516	3588	142334
	270908	6527	142334
	307463	7876	142334
	837036	50160	142334
	884437	58275	142334
	1133543	124798	142334

Deze lijst is alleen niet heel duidelijk, wie is acteur 83? Dit moeten we dan in de Actor tabel opzoeken.

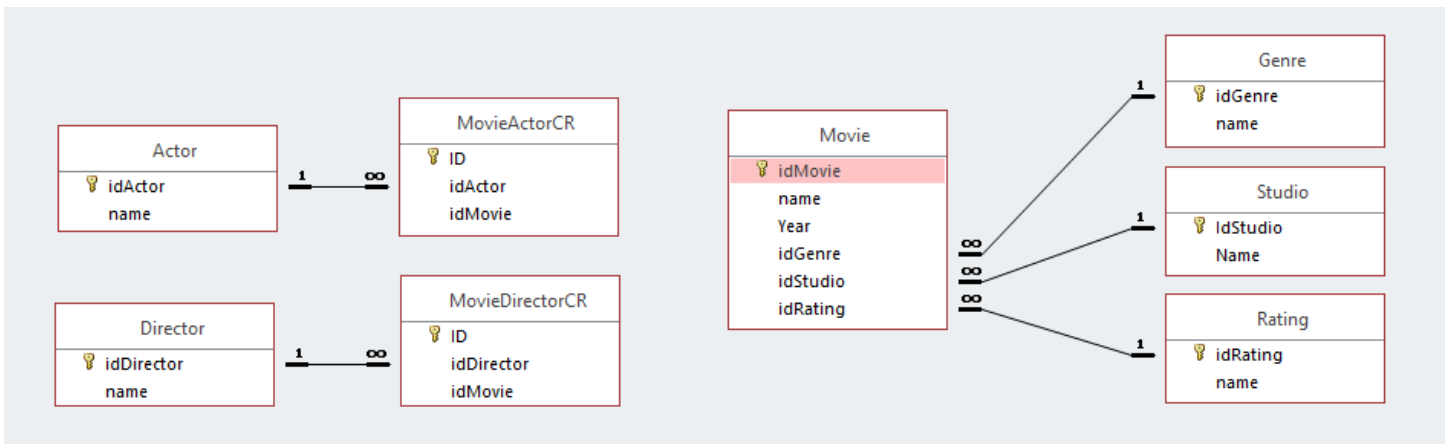
We kunnen die in een keer doen met een SQL-Statement.

Movie	Query1
select Actor.* from MovieActorCR, Actor where idMOvie = 142334 and actor.idActor = MovieActorCR.idActor	

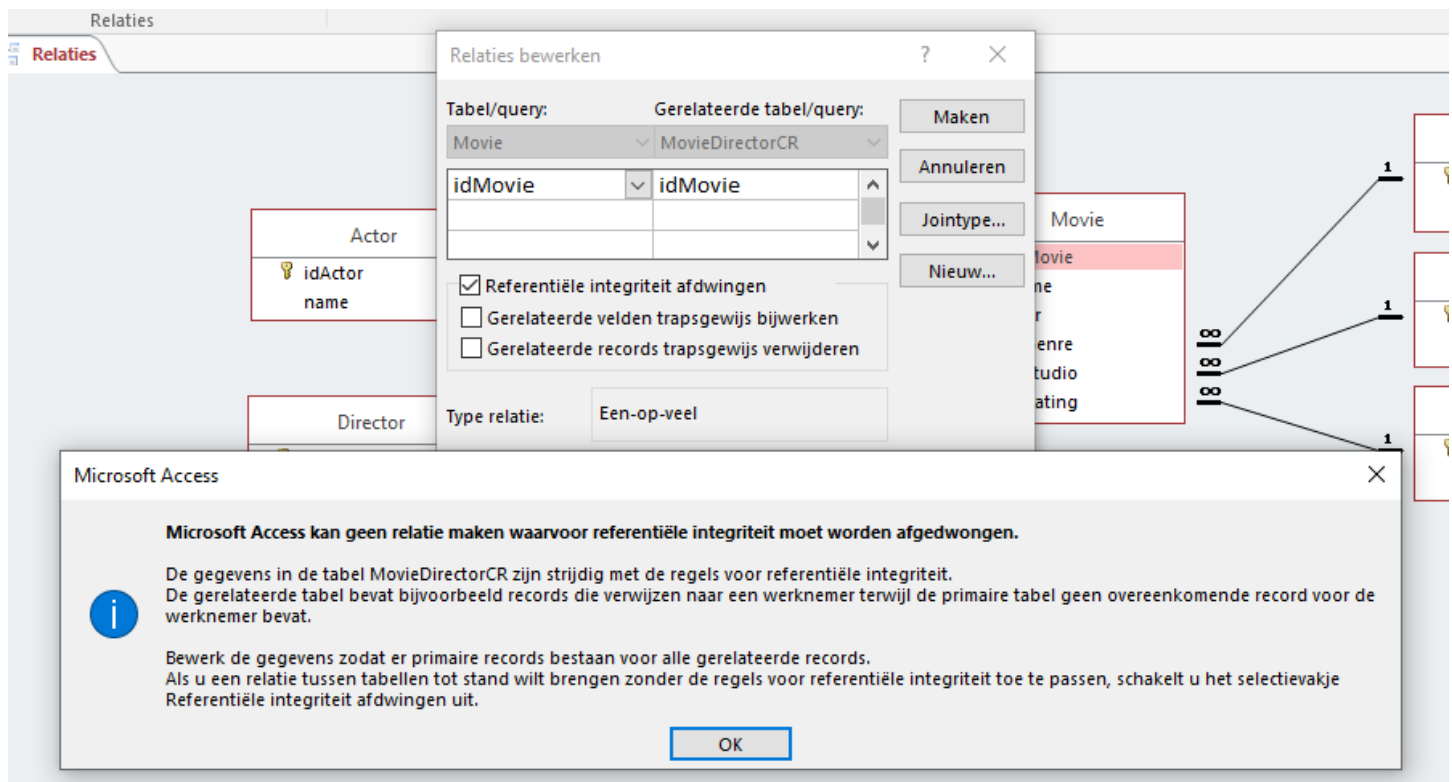
Movie	Query1
idActor	name
83	Gregg, Clark
494	Downey, Robert, Jr.
2017	Paltrow, Gwyneth
2479	Tahir, Faran
3588	Bridges, Jeff
6527	Howard, Terrence
7876	Smitrovich, Bill
50160	Toub, Shaun
58275	Bibb, Leslie
124798	Badreya, Sayed

We zien dat de nummers kloppen, alleen hebben we nog niet de relaties in de database gelegd. Zolang je dat niet doet kunnen we in de cross-reference tabel ook waarden zetten waar helemaal geen acteurs bij horen.

We doen dit weer bij het tabblad ontwerp en dan bij de relaties. Let op dat we slepen van de Actor tabel naar de CR-tabel. Dit heeft te maken met het soort relatie, idActor komt maar 1x voor



Als we deze relatie willen leggen tussen idMovie en de CR-tabellen zien we dat er iets niet goed gaat. Dit heeft niets te maken dat wij het niet goed doen, maar dat er een fout in de database zit. In “Figuur 10: Fout bij relatie idMovie.” zie je ook gelijk de kracht van een goede database. Want je wilt dat alles goed gekoppeld is, en dat er geen fouten in mogen zitten. Een goede database is zo gemaakt dat er gewoon geen fouten in kunnen zitten.



Figuur 10: Fout bij relatie idMovie.

Er staan idMovies in de MovieActorCR tabel die niet in de Movie tabel staan. We kunnen dit achterhalen met de volgende SQL-Statement, en zien dat er 44 idMovies niet bekend zijn.

```

Query2 Query3
Select distinct MovieActorCR.idMovie
from MovieActorCR left join MOVie
on MovieActorCR.idMovie = MOVie.idMovie
where MOVie.idMovie is null
  
```

De 44 idMovie die niet bekend zijn verwijderen we gewoon in de CR-tabel. Dit zullen wel films zijn die eens verwijderd zijn, en niet alle tabellen zijn bijgewerkt.

```

Query2 Query3
delete * from MovieActorCR where MovieActorCR.idMovie in
(
  Select distinct MovieActorCR.idMovie
  from MovieActorCR left join MOVie
  on MovieActorCR.idMovie = MOVie.idMOVie
  where MOVie.idMOVie is null
)

```

Als we dit hebben uitgevoerd dan kunnen we wel de relatie leggen tussen Movie en MovieActorCR. We moeten ditzelfde doen met de Director tabel.

```

Query2 Relaties
delete * from MovieDirectorCR where MovieDirectorCR.idMovie in
(
  Select distinct MovieDirectorCR.idMovie
  from MovieDirectorCR left join MOVie
  on MovieDirectorCR.idMovie = MOVie.idMOVie
  where MOVie.idMOVie is null
)

```

Tijdens deze stappen hebben we wat foutjes gemaakt. Zo hebben we wat namen van kolommen op de verkeerde manier geschreven. Zoek de foutjes op en wijzig de namen.

Als we helemaal klaar zijn hebben we de onderstaande database gekregen.

