



**InfoZoom<sup>®</sup>**  
**ACADEMY**

**Best Practice Day 2011**  
**DQ-Workshop**



© 2011 humanIT Software GmbH  
Brühler Straße 9 | 53119 Bonn | Telefon +49 228 90954 –0 | Telefax +49 228 90954 –11  
Ingo Lenzen, academy@infozoom.com

Alle enthaltenen Inhalte und Abbildungen sind urheberrechtlich geschützt.  
Kopieren oder Nachdruck verboten.  
Ausnahmen sind nur mit ausdrücklicher, schriftlicher Genehmigung gestattet.

1. Auflage

---

Inhalt

---

1	Ausgangslage .....	4
2	Analytische Fallbeispiele .....	6
2.1.	Datenstrukturen .....	6
2.2.	Metadaten.....	9
2.3.	Fehlende Hausnummer .....	10
2.4.	Dublettenanalyse .....	12
2.5.	Cluster-Bildung.....	16
2.6.	Automatisierung.....	18

## 1 Ausgangslage

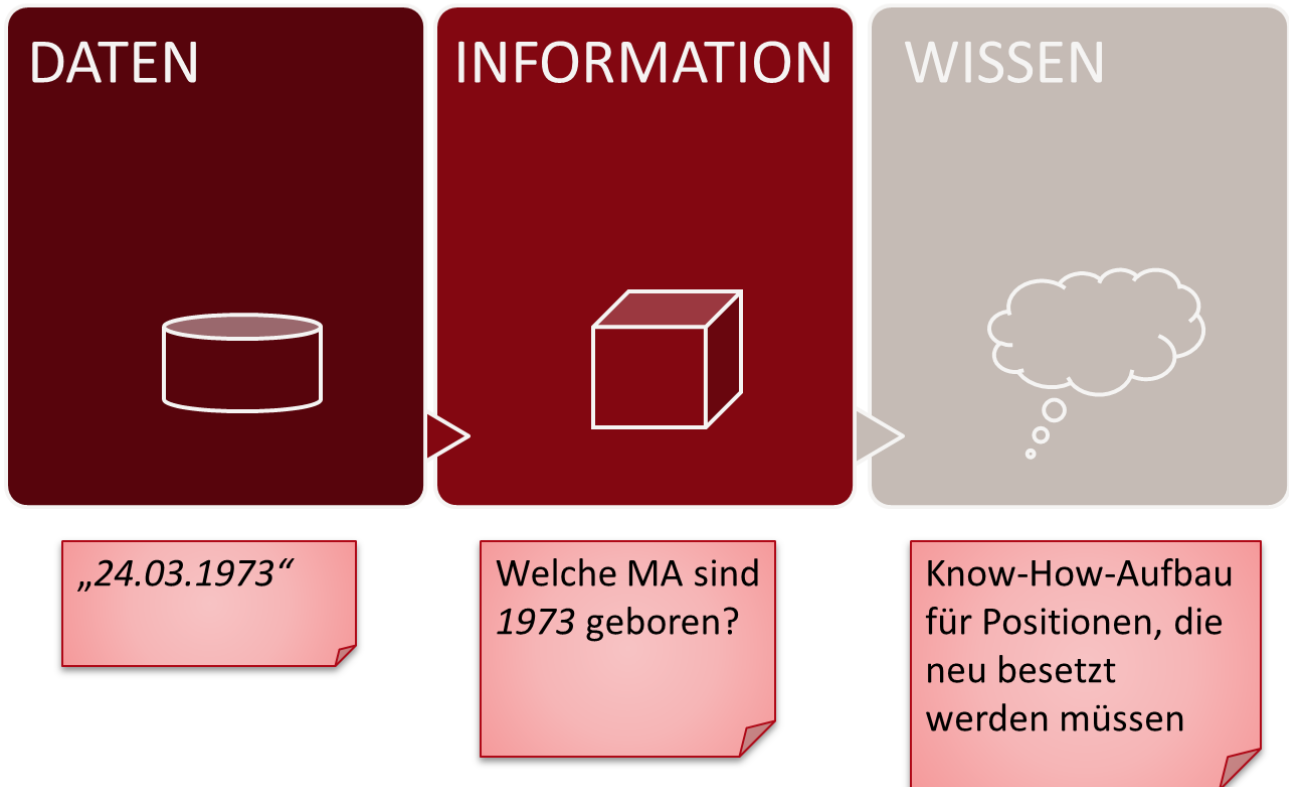
### Situation

- Situation
  - Die Erkennungs- und Steuersysteme sind nur so gut wie die Informationen, auf denen sie beruhen:
    - bei schlechter DQ werden Risiken nicht erfasst!
  - Schlechte Datenqualität hat direkte Auswirkungen
    - auf die Geschäftsprozesse
    - auf die Erfüllung von Richtlinien
    - als Entscheidungsgrundlage
    - und somit auf die Gewinnsituation
  - Mangelhafte Datenqualität und ungeeignete Informationssysteme sind ein unternehmerisches Risiko

### Handlungsempfehlung

- Handlungsempfehlung
  - Chancen und Risiken frühzeitig erkennen durch stetige Steigerung von Transparenz und Informationsqualität in der Unternehmensorganisation
    - Prozesse
    - Daten- und Dokumentenfluss
    - IT-Systeme
    - Aufgabenstellung u. -verteilung
  - »Transparenz schafft Handlungsalternativen!«

## Daten – Informationen – Wissen



## Die richtigen Fragen stellen

- Fragen
  - Stimmen alle Daten, die ich zum Rechnen benutze?
  - Weiß ich eigentlich genau, was in meinen Reports berechnet wird?
  - Weiß ich, was es mich kostet, 50% der Kosten zu reduzieren?
  - Weiß ich überhaupt genau, was ich wissen will?



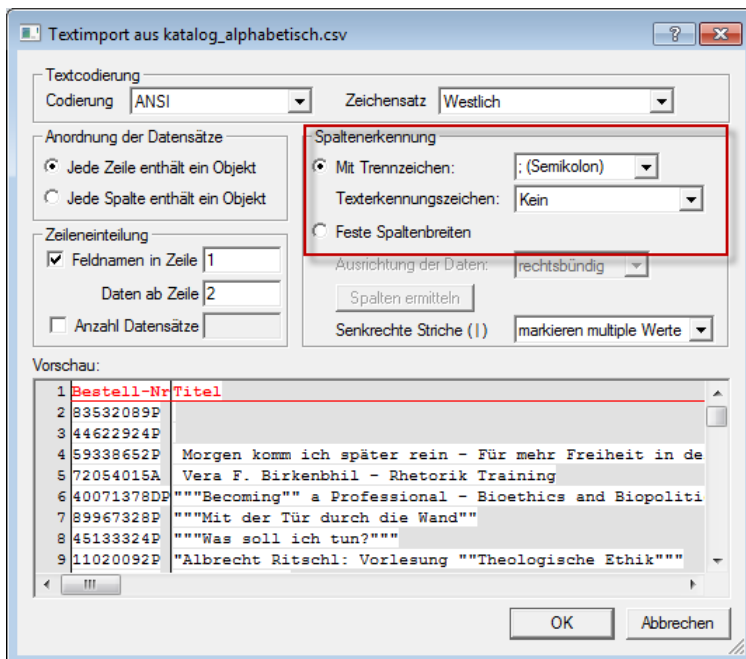
## 2 Analytische Fallbeispiele

### 2.1. Datenstrukturen

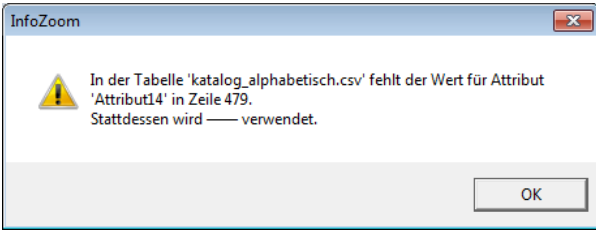
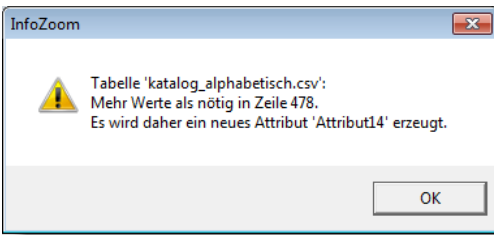
#### Datenstrukturen

- Aufgabenstellung:
  - Laden Sie alle Daten aus der Datei »katalog\_alphabetisch.csv« nach InfoZoom ein.
    - Achten Sie darauf, dass Sie das richtige Trennzeichen angeben
    - Achten Sie darauf, welches Texterkennungszeichen Sie angeben.
  - Wie viele Datensätze zeigen nach dem Import eine falsche Datenstruktur (Werte werden den Feldern falsch zugeordnet)?
  - Korrigieren Sie diese Datensätze und löschen Sie anschließend die überflüssigen Attribute.
  - Überprüfen Sie die Datenformate. Welche Unstimmigkeiten gibt es in den Daten?
  - Speichern Sie die korrigierten Daten als InfoZoom-Tabelle mit dem Dateinamen »katalog\_alphabetisch.fox« ab.

Öffnen Sie die Datei »katalog\_alphabetisch.csv« aus dem Übungsverzeichnis in InfoZoom. Im Import-Dialogfenster sollten die Einstellungen »;« für Trennzeichen und »Kein« für Texterkennungszeichen eingestellt werden.

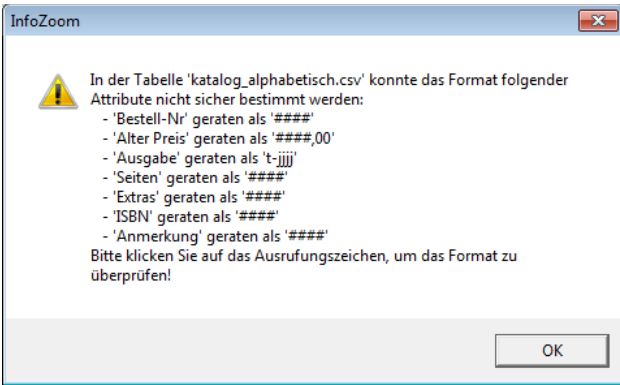


Beim Einlesen der Daten erhalten Sie den Hinweis, dass die Datenstruktur in Zeile 478 (inkl. Überschrift) nicht korrekt eingehalten werden konnte:



Bestätigen Sie beide Hinweise.

Ebenso erhalten Sie einen Hinweis, dass das Format einiger Attribute nicht eindeutig bestimmt werden konnte. Dieses weist darauf hin, dass ggfs. Dateninhalte nicht korrekt gepflegt wurden:



Bestätigen Sie auch diesen Hinweis.

Öffnen Sie anschließend die Werteliste des Attributs »Anmerkung« und zoomen Sie in die beiden Werte ein.

2 von 1.418 Objekten 14 Attribute		46437174	86248193P
Bestell-Nr	⚠	46437174	86248193P
Titel	🔍	Abi-Profi Englisch	Alles, was Sie über Rohstoffe wissen
Artikel Url	🔍	Ausgabe Nordrhein-Westfal	Co.
Alter Preis	⚠	http://www.terrasl	http://www.terrasl
Neuer Preis	🔍	9,95	34,90
Rubrik	🔍	2,99	29,99
Autor(en)	🔍	Bücher	E-Books
Verlag	🔍	—	Udo Rettberg
Ausgabe	⚠	Cornelsen	FinanzBuch Verlag
Seiten	⚠	2001	1-2007
Extras	⚠	82	451
ISBN	⚠	—	—
Anmerkung	⚠	3-464-37174-3	3898793095
Attribut14	🔍	—	—

**Anmerkung**

▼ 2 Werte

- 3-464-37174-3
- 3898793095

Nun können Sie erkennen, dass die Datenstruktur nicht korrekt eingelesen wurde. Im Attribut »Alter Preis« steht beispielsweise der Inhalt des Attributs »Artikel Url«.

In zwei Datensätzen wurden die Inhalte falsch zugeordnet.

Korrigieren Sie beide Datensätze, indem Sie die jeweiligen Inhalte anpassen.

2 von 1.418 Objekten 14 Attribute	46437174	86248193P
Bestell-Nr	46437174	86248193P
Titel	Abi-Profi Englisch; Ausgabe Nordrhein-Westfalen	Alles, was Sie über Rohstoffe wissen müssen - Erfolgreich mit Kaffee, Gold & Co.
Artikel Url	http://www.terrashop.de/Buch/Abi-Profi-Englisch-Ausgabe-Nordrhein-Westfalen-ISBN-3464371743/art/4	http://www.terrashop.de/E-Book/Alles-was-Sie-ueber-Rohstoffe-wissen-muessen-Erfolgreich-mit-Kaffee-G
Alter Preis	9,95	34,90
Neuer Preis	2,99	29,99
Rubrik	Bücher	E-Books
Autor(en)	—	Udo Rettberg
Verlag	Cornelsen	FinanzBuch Verlag
Ausgabe	2001	1-2007
Seiten	82	451
Extras	—	451
ISBN	3-464-37174-3	3898793095
Anmerkung	3-464-37174-3	3898793095
Attribut14	—	—

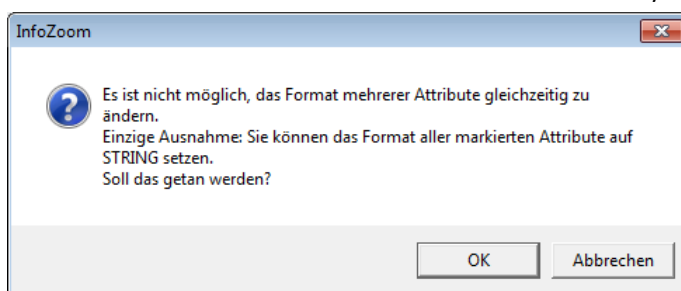
Löschen Sie die Werte aus dem Attribut »Anmerkung« und führen Sie einen Reset durch.

Löschen Sie abschließend das Attribut »Attribut14«, falls es keine Werte enthält.

Im nächsten Schritt überprüfen Sie die Formate.

Setzen Sie das Format der Attribute »Bestell-Nr«, »Extras«, »ISBN« und »Anmerkung« auf »String«.

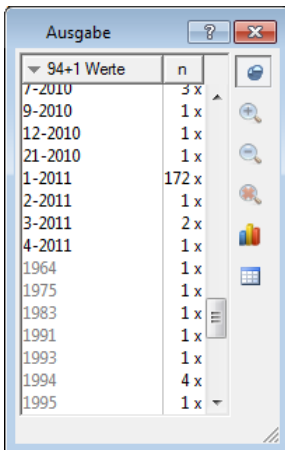
Markieren Sie hierzu alle Attribute und klicken Sie auf das Symbol »Format« im Registerreiter »Analyse«.



Bestätigen Sie den Hinweis mit »OK«.

Öffnen Sie den Formatdialog des Attributs »Alter Preis« und überprüfen Sie alle Werte durch Bestätigung des eingestellten Formats. Führen Sie dieses auch mit dem Attribut »Seiten« durch.

Analysieren Sie die Inhalte des Attributs »Ausgabe« über die Werteliste.



Wert	n
7-2010	3 x
9-2010	1 x
12-2010	1 x
21-2010	1 x
1-2011	172 x
2-2011	1 x
3-2011	2 x
4-2011	1 x
1964	1 x
1975	1 x
1983	1 x
1991	1 x
1993	1 x
1994	4 x
1995	1 x

Es ist kein eindeutiges Format zu erkennen. Daher muss auch hier das Format »String« angewendet werden.

Öffnen Sie den Formatdialog des Attributs »Ausgabe« und passen Sie das Format an.

Speichern Sie die Datei unter dem Namen »katalog\_alphabetisch.fox« ab.

## 2.2. Metadaten

### Metadaten

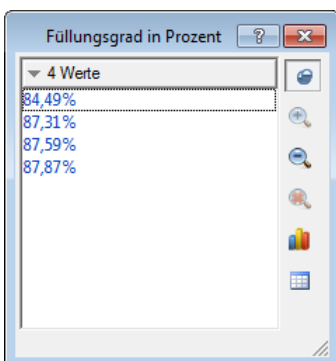
- Aufgabenstellung:
  - Erstellen Sie aus der Datei »katalog\_alphabetisch.fox« eine Metatabelle.
  - Wie viele Attribute weisen einen Füllungsgrad zwischen 80% und 90% aus?

Wechseln Sie auf die Registerkarte »Überprüfen« und klicken Sie auf das Symbol »Metadaten«.



Öffnen Sie die Werteliste des Attributs »Füllungsgrad in Prozent« in der Metadaten-Tabelle.

Zoomen Sie in die Werte zwischen 80% und 90% ein.



Wert
84,49%
87,31%
87,59%
87,87%

Die Attribute »Autor(en)«, »Ausgabe«, »Seiten« und »ISBN« weisen einen Füllungsgrad zwischen 80% und 90% aus.

## 2.3. Fehlende Hausnummer

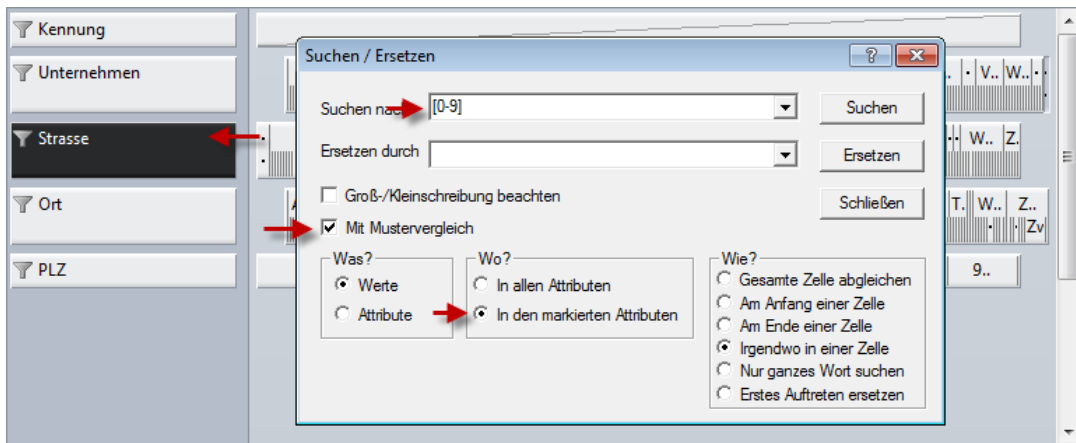
### Fehlende Hausnummer

- Aufgabenstellung:
  - Öffnen Sie die Datei »Adressen.fox«.
  - Finden Sie alle Adressen, die keine Hausnummer im Attribut »Strasse« haben und speichern Sie die gefundenen Adressen als Excel-Datei ab.
  - Lösungsweg:
    - Suchen nach Hausnummern im Attribut »Strasse« über den Suchen-Dialog
      - Suchen nach einer Ziffer mit Musterwert »[0-9]«
      - Daten werden festgehalten
    - Zoom-Out des Attributs »Strasse«
    - Ausschließen der gefundenen Werte (mit Hausnummer); d.h. Umkehrung der Selektion
    - Ausgabe des Ergebnisses als Excel-Datei

Öffnen Sie die Datei »Adressen.fox« aus dem Übungsverzeichnis.

Öffnen Sie den Suchen-Dialog und geben Sie im Suchfeld die Bereichssuche »[0-9]« (Ziffern) ein (als Mustervergleich).

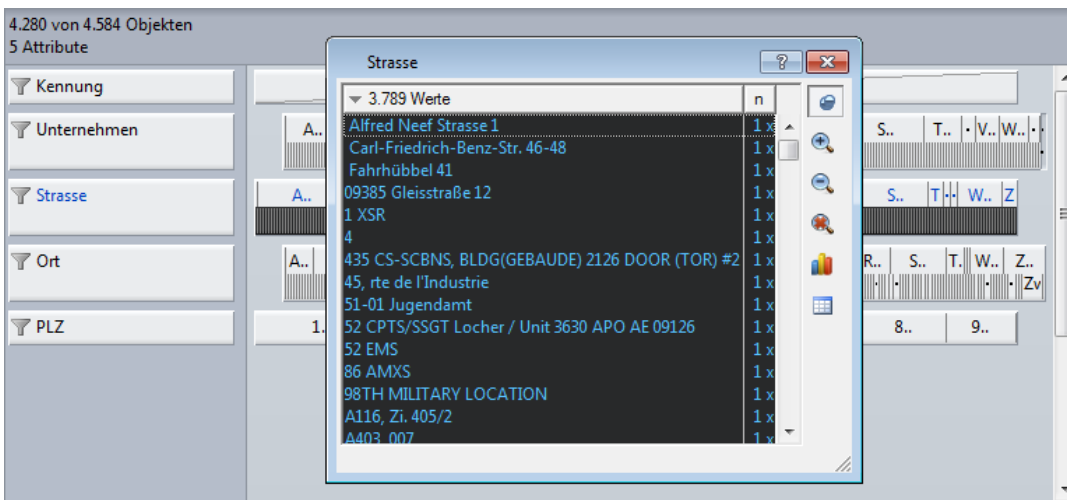
Selektieren Sie die Checkbox »Mit Mustervergleich«. Markieren Sie das Attribut »Strasse« und im Suchen-Dialog den Radio-Knopf »In den markierten Attributen«.



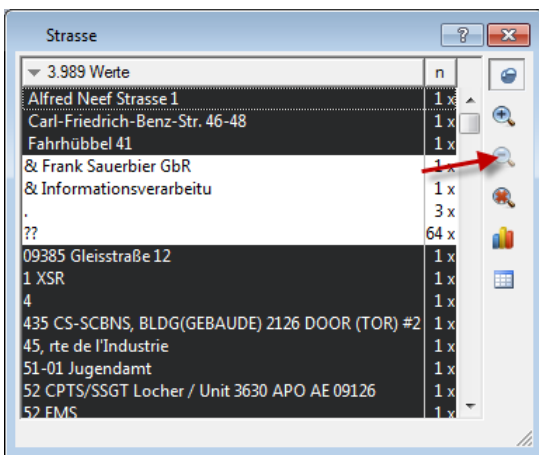
Nun sind alle Strassen fokussiert, die eine Ziffer als Inhalt enthalten.

Klicken Sie auf »Suchen« und öffnen Sie die Werteliste des Attributs »Strasse«.

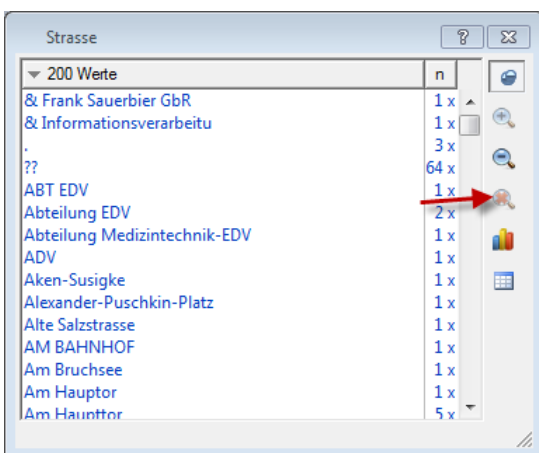
Der Suchen-Dialog kann geschlossen werden.



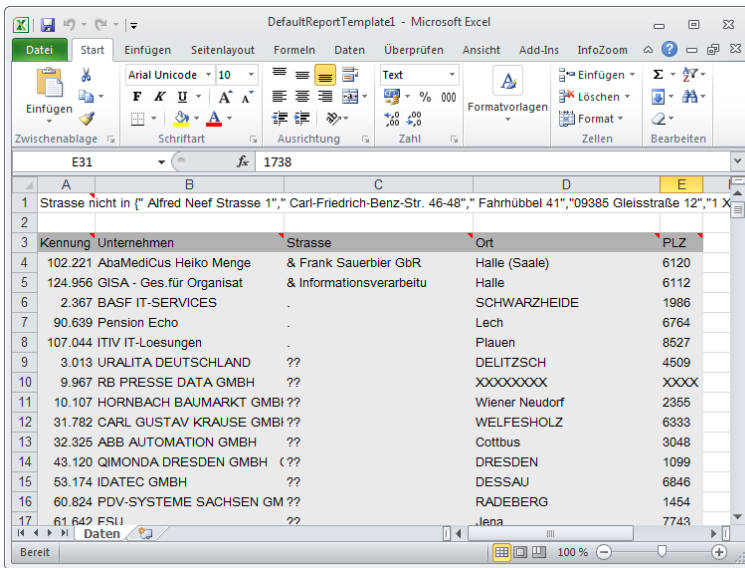
Führen Sie nun einen »Zoom-Out« auf das Attribut »Strasse« aus, ohne die Markierung der Werte zu verändern.



Schließen Sie die zuvor gefundenen Werte aus. Als Ergebnis erhalten Sie alle Strassen, welche keine Ziffer (Hausnummer) als Inhalt haben.



Markieren Sie alle Attribute und erstellen Sie über die Registerkarte »Ergebnisse« eine Excel-Datei mit den gefundenen Werten. Speichern Sie diese Datei ab.



	A	B	C	D	E
1	Strasse nicht in (" Alfred Neef Strasse 1"," Carl-Friedrich-Benz-Str. 46-48"," Fahrhübel 41","09385 Gleisstraße 12","1 X				
2					
3	Kennung	Unternehmen	Strasse	Ort	PLZ
4	102.221	AbaMediCus Heiko Menge	& Frank Sauerbier GbR	Halle (Saale)	6120
5	124.966	GISA - Ges.für Organisat	& Informationsverarbeitu	Halle	6112
6	2.367	BASF IT-SERVICES	.	SCHWARZHEIDE	1986
7	90.639	Pension Echo	.	Lech	8764
8	107.044	ITIV IT-Loesungen	.	Plauen	8527
9	3.013	URALITA DEUTSCHLAND	??	DELITZSCH	4509
10	9.967	RB PRESSE DATA GMBH	??	XXXXXXXX	XXXX
11	10.107	HORNBACH BAUMARKT GMBH	??	Wiener Neudorf	2355
12	31.782	CARL GUSTAV KRAUSE GMBH	??	WELFESHOLZ	6333
13	32.325	ABB AUTOMATION GMBH	??	Cottbus	3048
14	43.120	QIMONDA DRESDEN GMBH	(??	DRESDEN	1099
15	53.174	IDATEC GMBH	??	DESSAU	6846
16	60.824	PDV-SYSTEME SACHSEN GM	??	RADEBERG	1454
17	61.642	FSU	??	.lena	7743

## 2.4. Dublettenanalyse

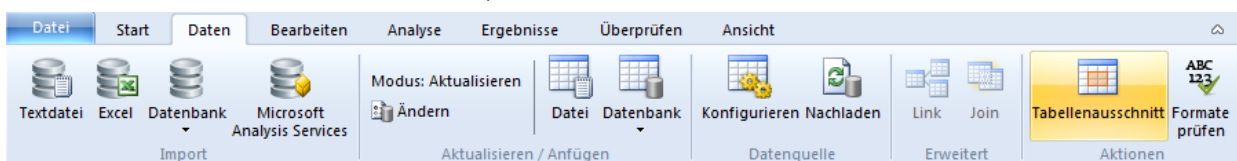
### Dublettenanalyse

- Aufgabenstellung:
  - Öffnen Sie die Datei »Adressen.fox«.
  - Finden Sie möglichst alle doppelten Adressen und speichern Sie nach Erstellung der Analyseattribute die Datei unter dem Namen »Adressen\_Analyse.fox« ab.
  - Lösungsweg:
    - Erstellen Sie den Phonetischen Code der Attribute »Ort«, »Strasse« und »Unternehmen« (Formelattribut)
    - Erstellen Sie einen Analysewürfel und zählen Sie die Adressen
      - Als Kennzahl nutzen Sie die Aggregation »Anzahl«
      - Als Dimensionen nutzen Sie die oben erzeugten Attribute
    - Öffnen Sie die Werteliste der Kennzahl und zoomen Sie in die größten Werte ein.
      - Öffnen Sie gleichzeitig zur Kontrolle die Wertelisten »Ort«, »Strasse« und »Unternehmen«

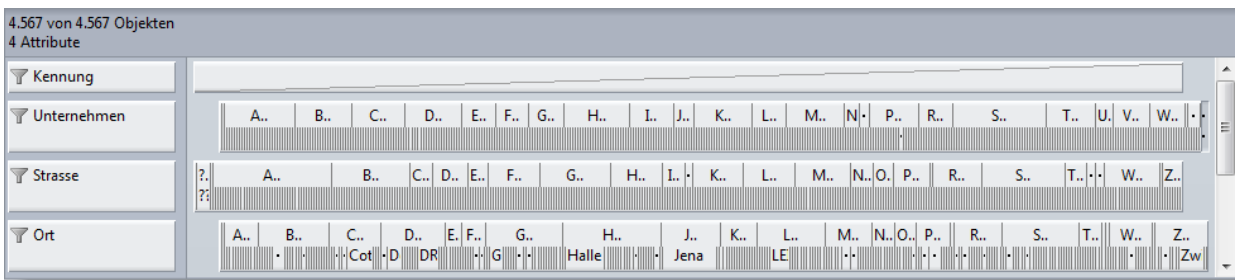
Öffnen Sie die Datei »Adressen.fox« aus dem Übungsverzeichnis.

Führen Sie einen Reset durch.

Markieren Sie die Attribute »Unternehmen«, »Strasse« und »Ort« und erstellen Sie einen Tabellenausschnitt.

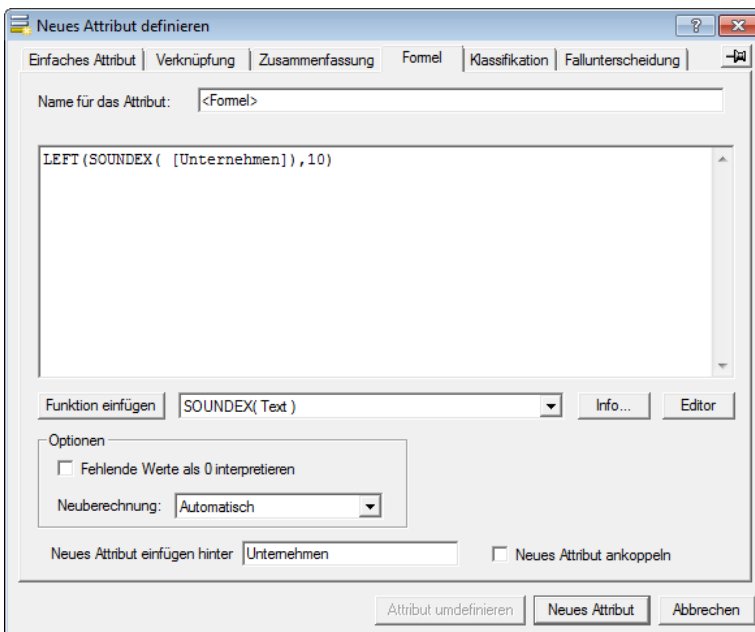


Eine neue Tabelle mit 4.567 Objekten wird erzeugt.

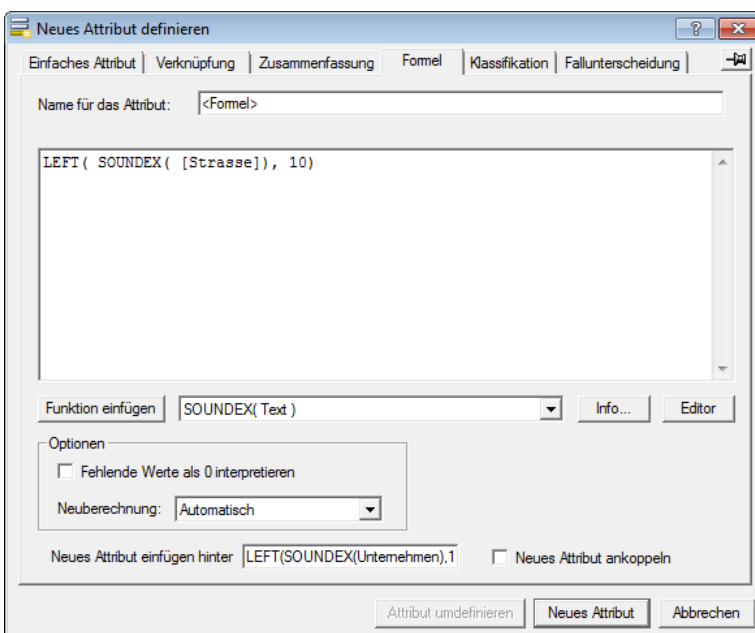


Erstellen Sie drei neue abgeleitete Attribute (Formel), die jeweils den Phonetischen Code der Attribute »Unternehmen«, »Strasse« und »Ort« errechnen. Beim »Unternehmen« und bei der »Strasse« sollen die ersten zehn Stellen des Code berücksichtigt werden. Beim Attribut »Ort« werden die ersten sechs Stellen berücksichtigt.

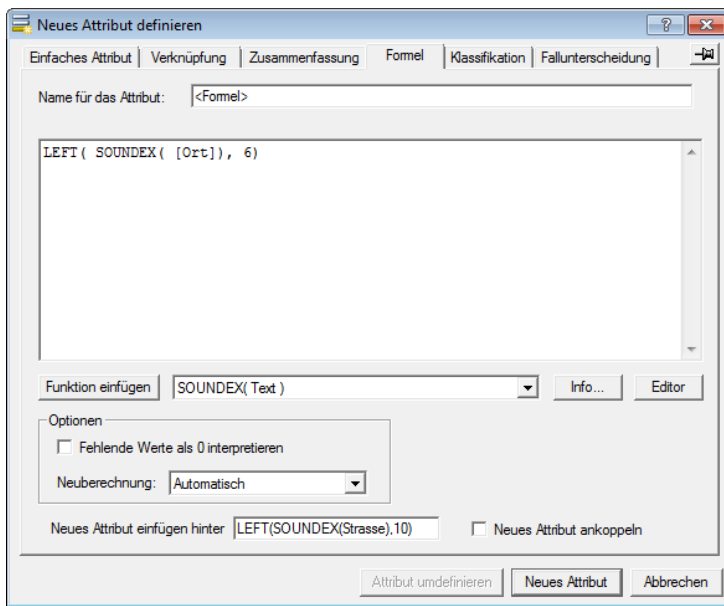
Unternehmen:



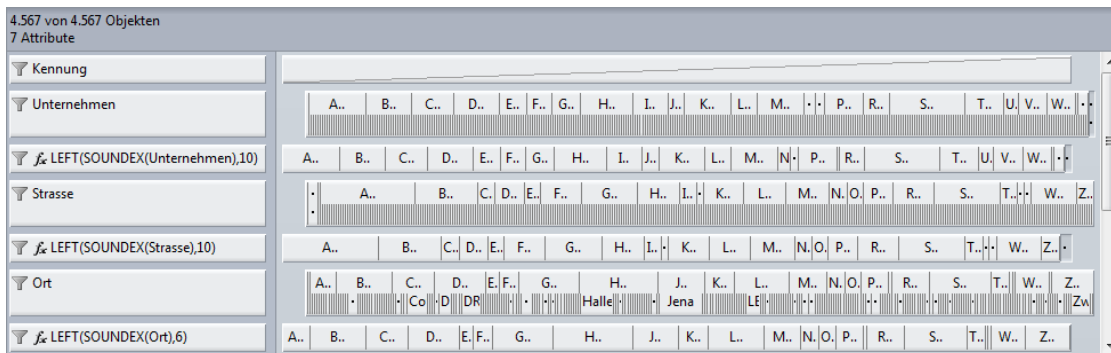
Strasse:



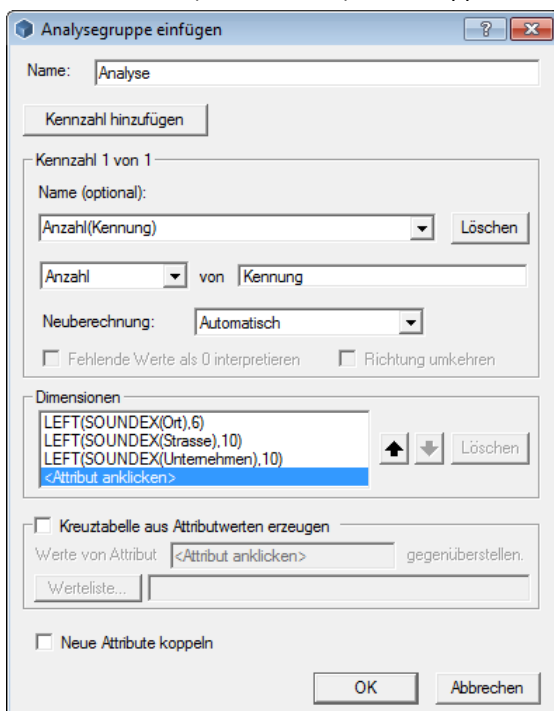
Ort:



Als Ergebnis erhalten Sie drei neue abgeleitete Attribute.



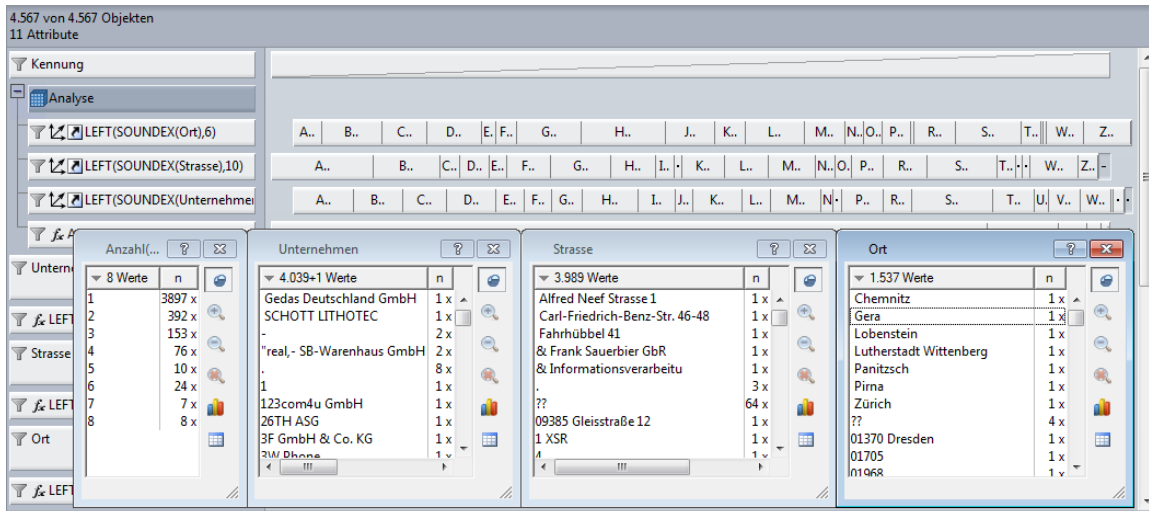
Erstellen Sie nun einen Analysewürfel, der die einzelnen Objekte je Phonetischem Code von »Ort«, »Strasse« und »Unternehmen« (Dimensionen) zählt. Koppeln Sie die Attribute dabei nicht aneinander.



Als Ergebnis erhalten Sie eine Analysegruppe mit einer Kennzahl und drei Dimensionen.

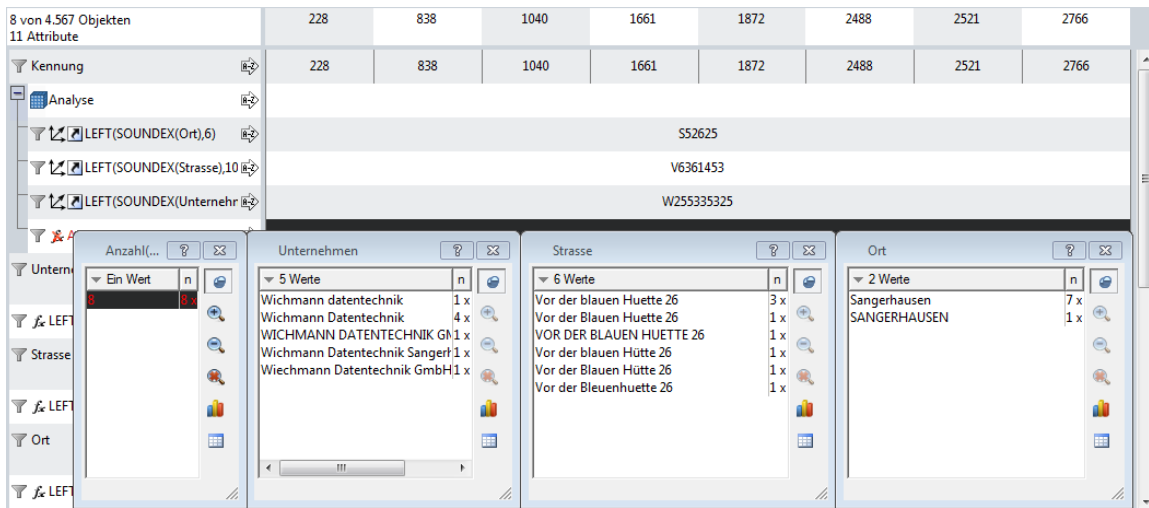


Öffnen Sie die Wertelisten der Attribute »Ort«, »Strasse«, »Unternehmen« und »Anzahl(Kennung)«. Ordnen Sie die Wertelisten auf dem InfoScope an.



Die »kritischen« Werte sind nun die Werte größer eins in der Werteliste »Anzahl(Kennung)«.

Zoomen Sie beispielsweise in den Wert »8« ein und Sie erhalten eine Adresse, die in 8 verschiedenen Varianten im Adressbestand vorhanden ist.



## 2.5. Cluster-Bildung

Neben der manuellen Datenbereinigung kann auch eine halbautomatische Datenbereinigung durchgeführt werden.

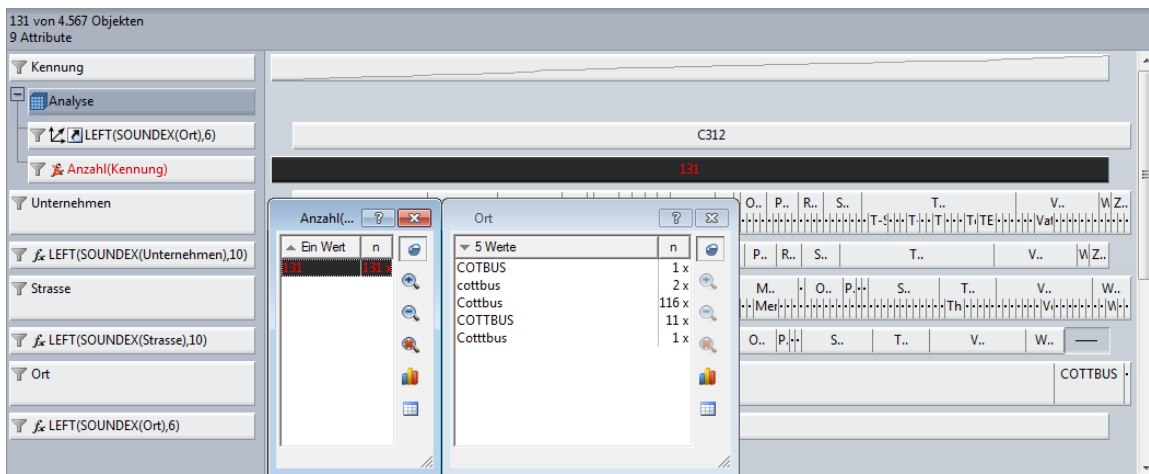
### Cluster-Bildung

- Aufgabenstellung:
  - Öffnen Sie die Datei »Adressen\_Analyse.fox«.
  - Definieren Sie Cluster, welche das Attribut »Ort« bereinigen.
  - Lösungsweg:
    - Erstellen Sie eine Klassifikation und fassen Sie Cluster-Werte in Wertemengen zusammen.

Die Lösung wird beispielhaft an einem Ort durchgeführt.

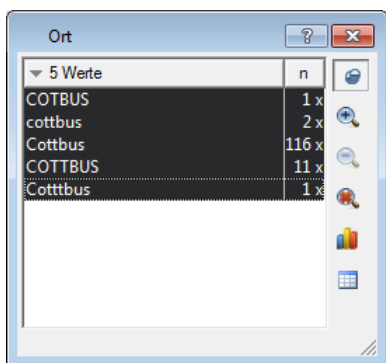
Nutzen Sie zur Clusterfindung bei einem Ort die oben aufgeführte Phonetische Analyse. Ziehen Sie hierzu die Phonetischen Dimensionen zu »Unternehmen« und »Strasse« aus dem Analysewürfel.

Öffnen Sie die Wertelisten der Attribute »Anzahl(Kennung)« und »Ort«. Zoomen Sie in einen Wert in der Werteliste »Anzahl(Kennung)« ein (hier: »131«).



In der Werteliste Ort werden die verschiedenen Schreibweisen angezeigt.

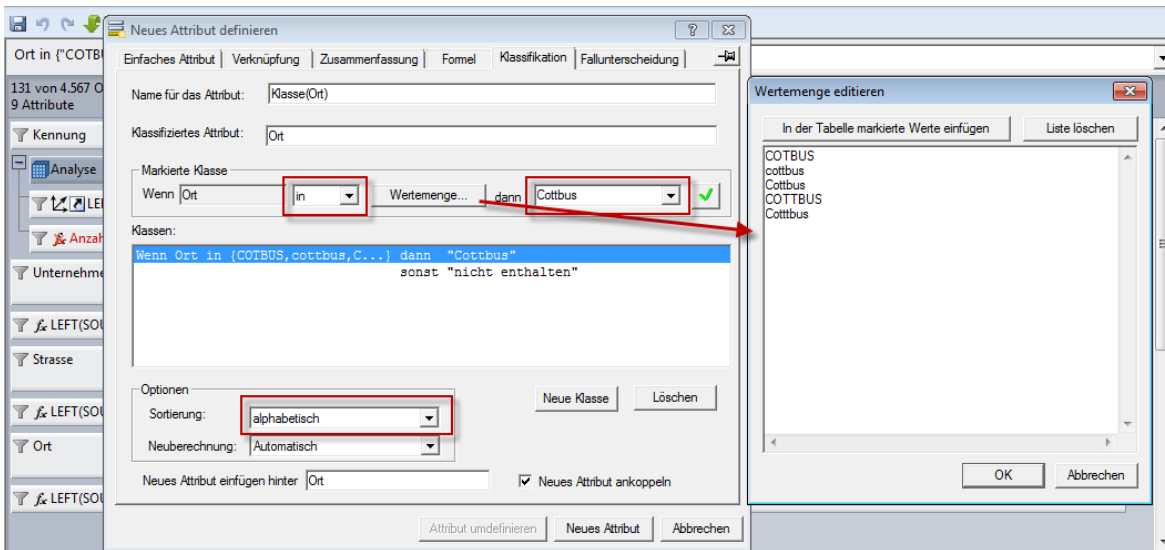
Selektieren Sie alle Werte in der Werteliste des Attributs »Ort«.



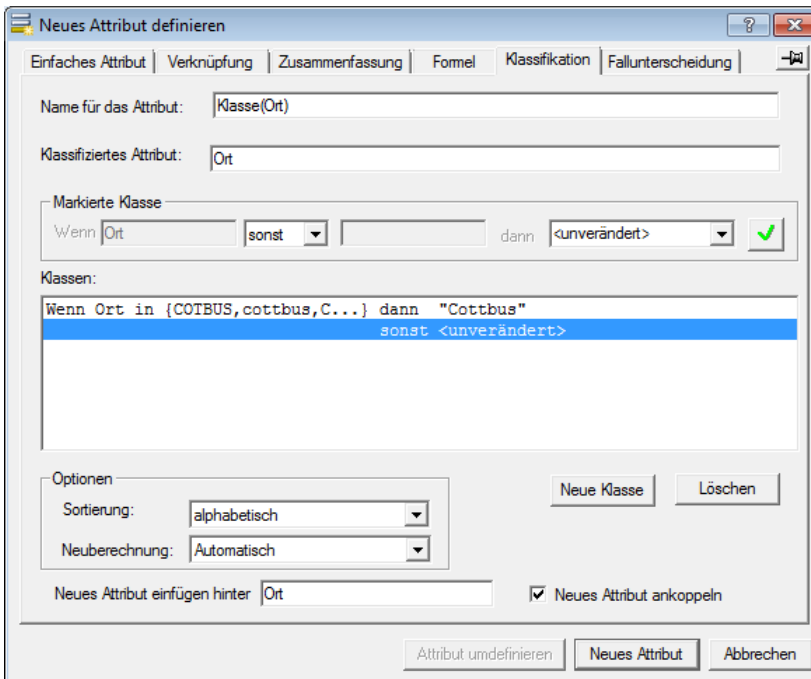
Erstellen Sie ein neues abgeleitetes Attribut (Klassifikation) und wählen Sie als Operator »in«.

In der zugehörigen Wertemenge werden die zuvor selektierten Werte eingetragen.

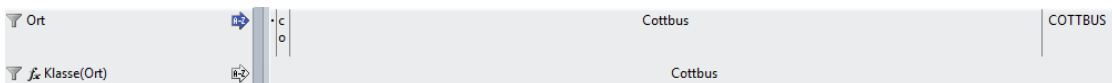
Wählen Sie als Ersetzungstext »Cottbus« und reihen Sie den Ersetzungstext alphabetisch ein.



In der zweiten Klasse mit dem Operator »sonst« wählen Sie den Ersetzungstext »<unverändert>«.



Sie haben nun ein Cluster für den Ort »Cottbus« definiert.



## 2.6. Automatisierung

### Automatisierung

- Typische Anwendungsgebiete sind
  - Konvertierung von Daten
  - Bereinigung von Daten
  - Erzeugung von InfoZoom-Tabellen aus beliebigen Quellen (auch ODBC)
  - Zusammenfügen von Daten aus verschiedenen Quellen (Join)
  - Aktualisieren von Daten (update, insert, delete)
  - Ausführen von vordefinierten Anfragen
  - Anzeigen von tabellarischen oder graphischen Reports
  - Erzeugung von spezifischen InfoZoom Tabellen für unterschiedliche Benutzer (-Gruppen)

### Automatisierung

