

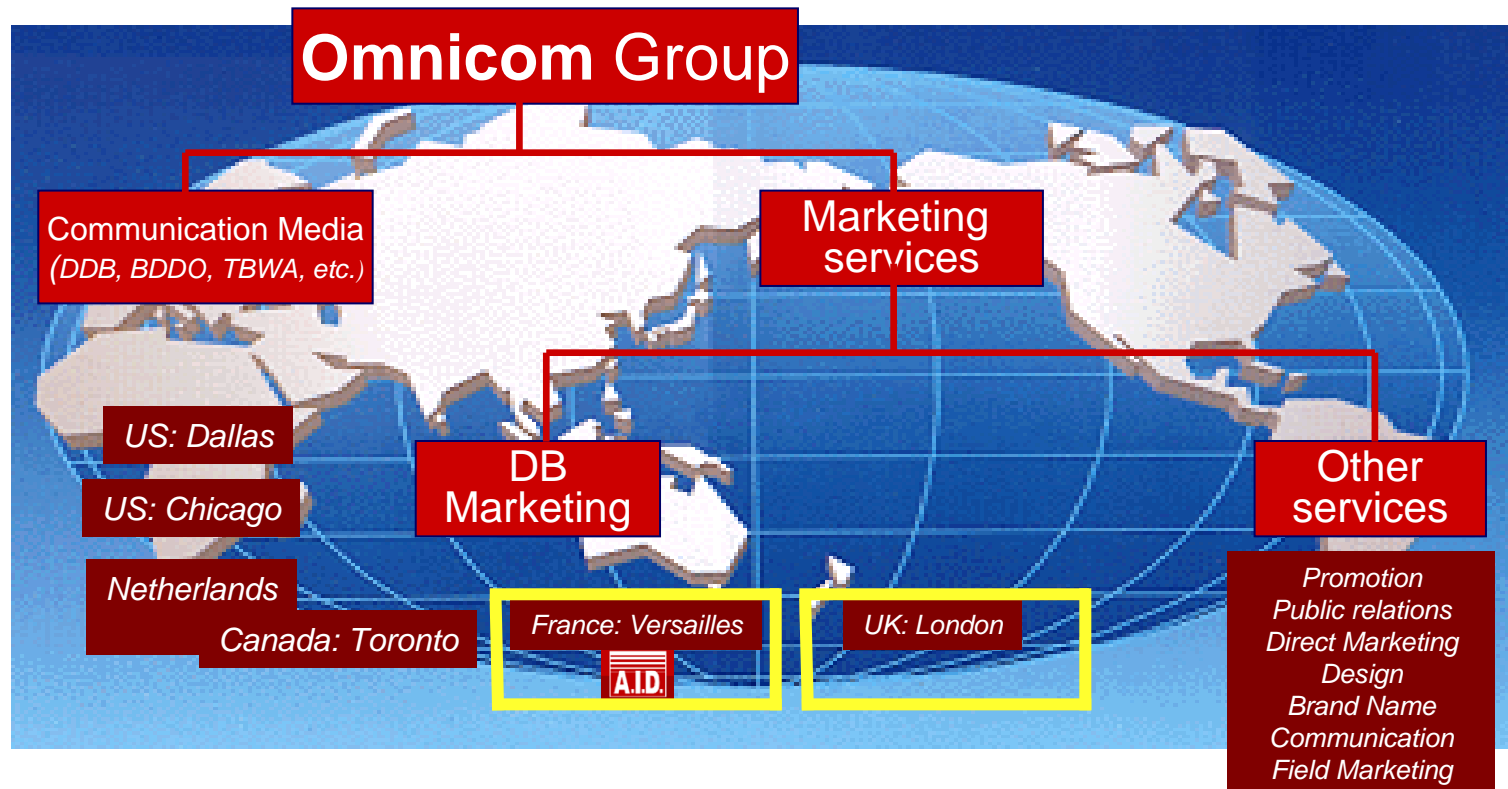


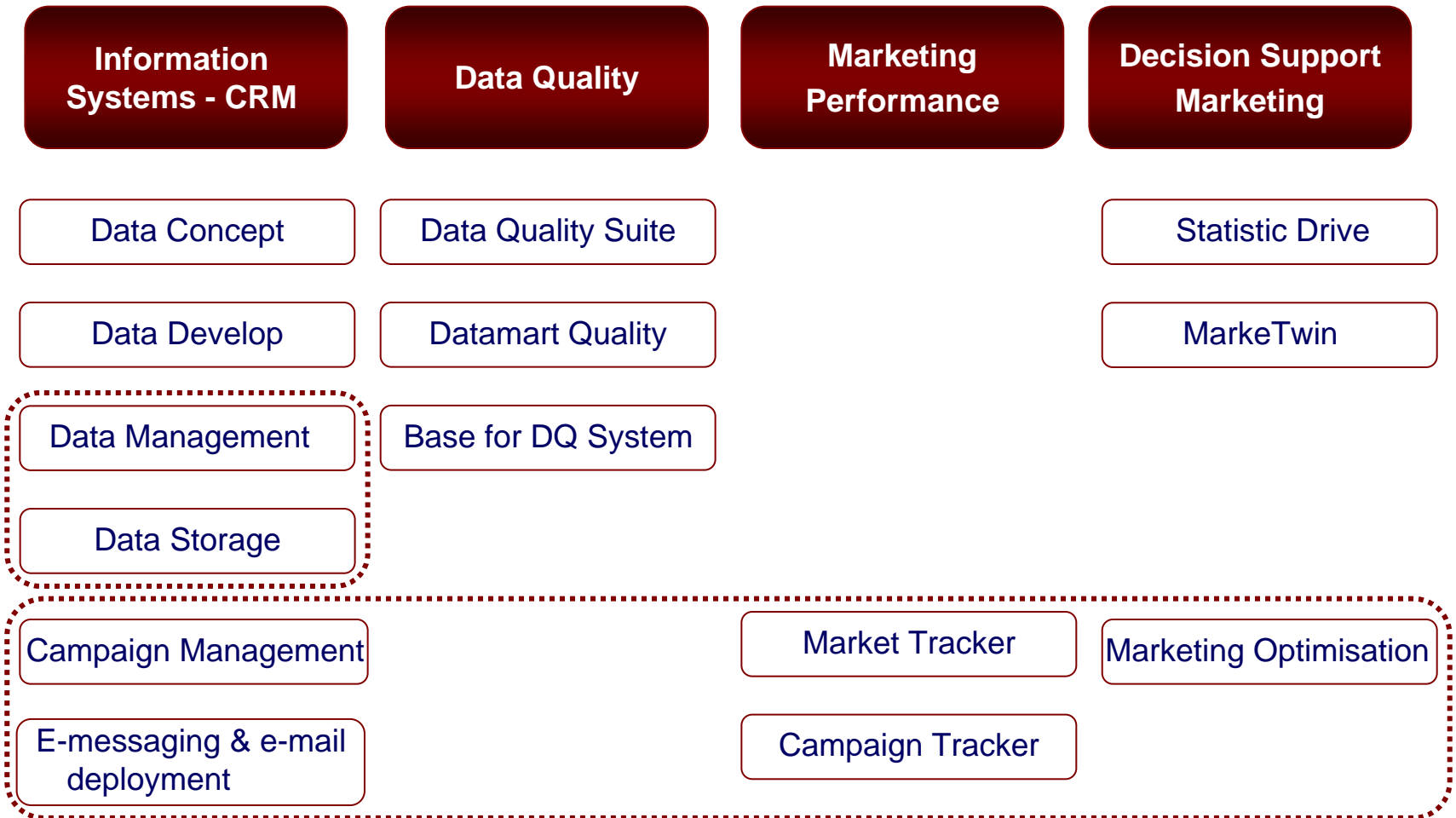
B.D.Q.S.

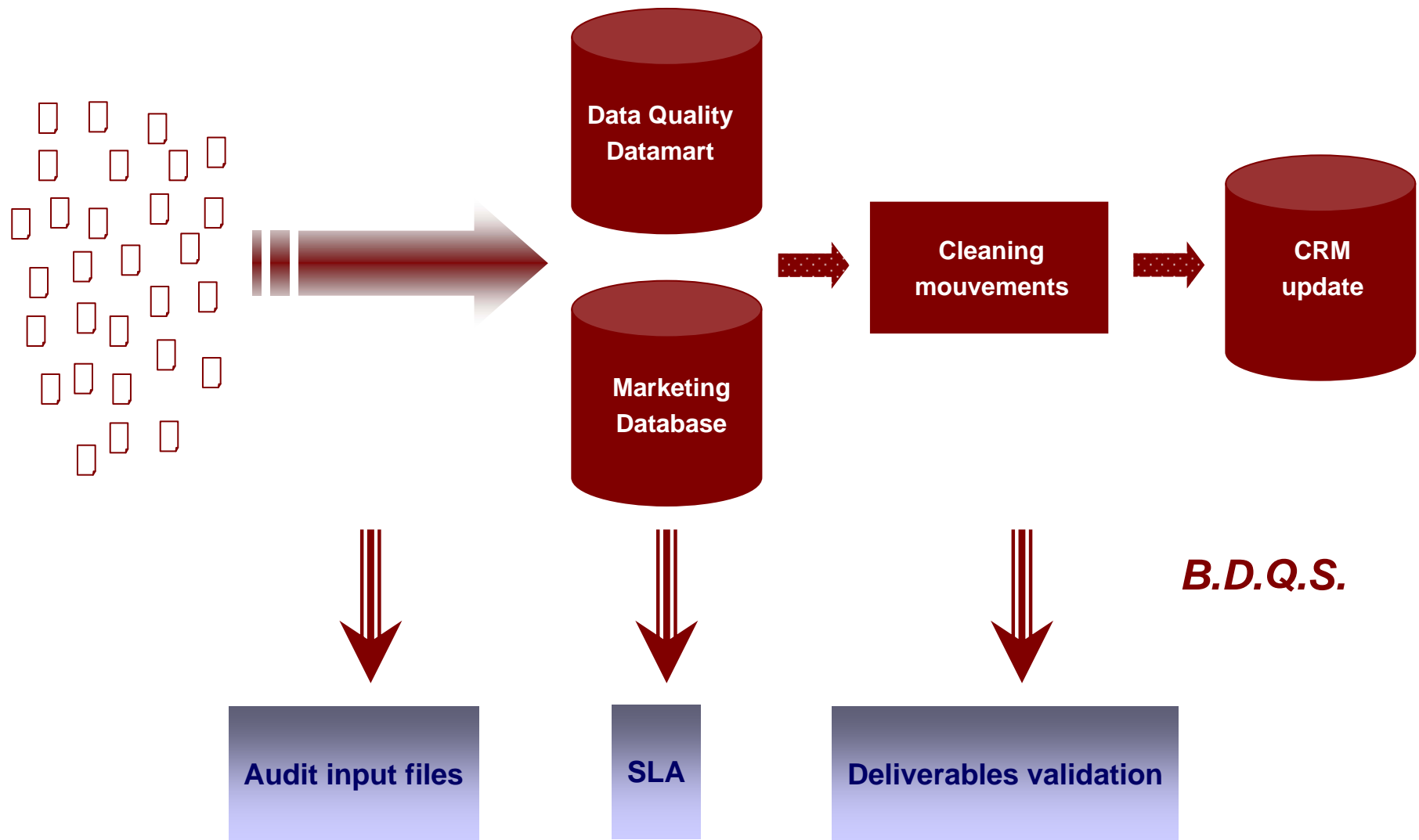
Why we use it at A.I.D.

GIVING MEANING TO YOUR DATA

- Omnicom Group is one of the leading international groups in the communications industry.
- A holding company at the forefront in the fields of communications, public relations, marketing and database services, Omnicom is present in over 100 countries.







⇒ **To audit** an input file:

- A file is received for deduplication, direct marketing operation or integration into a marketing database
- The contain, the parsing, LOVs, patterns are controlled thanks to BDQS

→ **To list the cleanings to put in place before to use the file**

⇒ **To measure on a regular basis** a system: marketing database, CRM,..hosted by A.I.D:

- To publish a data quality scorecard
- To validate our **SLA** engagement using a **tool different from the production tool**
- To give to the production tracks of improvement

⇒ **To validate** a file before delivery:

- After a deduplication operation or a CRM cleaning, the files are controlled before delivery
- Type of control: LOVs, uniqueness, integrity, fake values,...

- **List of values, integrity, uniqueness**
- **Pattern analysis**
- **Fake values: characters, words, repetitive characters**
- **Emails check, phone number check**
- **Duplicates: sophisticated tool based on a statistical method and taken into account data quality indicators**
- **Statistical analysis implementation in progress and specific algorithms can be implemented by the client**

condition C7

```
{  
  compare_not_null last_name { editdist <= 1 }  
  and compare_not_null first_name { letters >= 80 , skel > 50 }  
  and compare address { letters >= 80 , skel > 75 }  
  and compare_not_null country { equals }  
}
```

→ All the attributes can be combined thanks to indexes as Edit Distance

```
group email {  
  conditions {  
    C1 (11)  
    or C2 (12)  
    or C3 (13)  
    or C4 (14)  
    or C5 (10)  
    or C6 (15)  
  }  
}
```

→ Compare all the records with the same email

```
group tel {  
  conditions {  
    C7 (22)  
    or C8 (23)  
    or C9 (24)  
    or C10 (5)  
  }  
}
```

→ Compare all the records with the same phone number

Example :

					Level of confidency
1	Smith	Albert	salbert@hotmail.com		
2	Smitt	A	salbert@hotmail.com		12
3	Smithe	Alberd	salbert@hotmail.com	0139239345	14
4	Smiss	Albert		0139239345	21
5	Smiss	Albert		0139239345	5

2 dedicated queries :

- **Number of masters ?**
 - **Depending on the confidency level accepted**
- **Number of duplicates ?**
 - **To analyze the group of duplicates : origin, ...**

BDQS : Control Center

File Edit DataBase Statistics Publication Window Help

BDQS Base for Data Quality Systems

DEFINE COLUMN CATEGORY

Table : mit_demo Count not Null: 742 237
 Column : Column 9 Count of Null : 12 026
 Datatype : String

10 MIN		10 MAX		10 RANDOM	
28	¿0160	5	0894		
1	¿0159	1	1007		
4	¿0158	1	14002		
2	¿0155	1	30855		
171	oct.-60	1	3201 AB		
4	mars-87	1	7800		
1	mars-83	1	85757		
-145	mars-79	1	CO16		
262	mars-71	1	LA9 6NT		
4	mars-60	1	9914 3BY		

Select Category :

10 LEAST FREQUENT		10 MOST FREQUENT	
1	79008	6	713
1	75002	2	275
1	75009	1	926
1	75001	1	888
1	75017	1	642
1	75010	1	610
1	1000	1	605
1	75011	1	576
1	75015	1	469
1	75003	1	450

ALGORITHM NAME

Overview of the field :
 min, max, random
 modalities, most frequent,
 last frequent

PATTERNS	
[0-9]{5}	388 791
[0-9]{4}	86 835
[1-9][0-9]{3}[][A-EG-HJ-NPR-RV-XZ]{2}	47 042
[A-PR-LWVY][A-BD-HK-Y][0-9][A-BD-H][NP-LWV-Z]{2}	43 726

Most frequent patterns found on the field

Edit Dictionary

FREQUENCIES

- Targeting the entire population, not only samples

A query tool
user-friendly and
very fast

*Only few seconds for
complex queries on
millions and more
records*

To find specific
cases

An operational
measurement:
populations to
clean can be
directly extracted

BDQS : Control Center

File Edit DataBase Statistics Publication Window Help

Current table : mit_demo

Columns / Attributes

- Column 9
- Column 10
- Column 11
- Column 12
- Column 13

Column Queries Table Queries

		- Column 11							
		GB	FR	DE	IT	ES	NL	CH	BE
Completeness : Column 1		25.21%	19.29%	15.28%	10.22%	8.42%	6.81%	6.28%	3.66%
Completeness : Column 2		26.08%	19.95%	15.80%	10.57%	8.70%	7.04%	6.49%	3.79%
Completeness : Column 3		25.31%	18.29%	15.67%	10.44%	8.41%	6.93%	6.35%	3.69%
Completeness : Column 5		75.53%	20.28%	0.10%	0.88%	2.91%	0.10%	0.09%	0.09%
Completeness : Column 6		0.00%	2.52%	0.00%	0.42%	46.54%	14.05%	35.64%	0.00%
Completeness : Column 8		26.76%	20.74%	14.95%	11.34%	9.31%	5.10%	6.67%	4.06%
Completeness : Column 21		15.01%	14.02%	25.58%	6.69%	9.11%	8.57%	6.59%	3.69%
Completeness : Column 22		0.04%	56.23%	0.07%	0.82%	0.00%	1.55%	0.00%	1.42%
Completeness : Column 23		0.00%	88.65%	0.02%	0.15%	7.39%	0.04%	0.00%	0.05%
\$Integer [Compleat..	1	2	133,377	107,336	71,452	58,804	1	7	7
		0%	35%	28%	19%	15%	0%	0%	0%
	Total	2	133,377	107,336	71,452	58,804	1	7	7
		0%	35%	28%	19%	15%	0%	0%	0%

Current Query Column Modalities Column S...

MODALITY	COUNT
<input type="checkbox"/> AT	5,324
<input type="checkbox"/> BE	25,746
<input type="checkbox"/> CH	44,121
<input type="checkbox"/> CZ	13,978
<input type="checkbox"/> DE	107,363
<input type="checkbox"/> DK	4,991
<input type="checkbox"/> ES	59,144
<input type="checkbox"/> FI	3,630
<input type="checkbox"/> FR	135,573
<input type="checkbox"/> GB	177,191
<input type="checkbox"/> IE	12,074
<input type="checkbox"/> IT	71,852
<input type="checkbox"/> KNUTSFORD	1
<input type="checkbox"/> LONDON	1
<input type="checkbox"/> LU	1,396
<input type="checkbox"/> NL	47,867
<input type="checkbox"/> NO	3,247
<input type="checkbox"/> PL	19,898
<input type="checkbox"/> SE	8,836
<input type="checkbox"/> WETHERBY, LEEDS	1
<input type="checkbox"/> pl	3

- To contact us:

Olivier Coppet
Brigitte Labois

A.I.D.

00 33 1 39 23 93 00

<http://www.aid.fr>