Demonstrations of linguistic data matching, consolidation, and cleansing

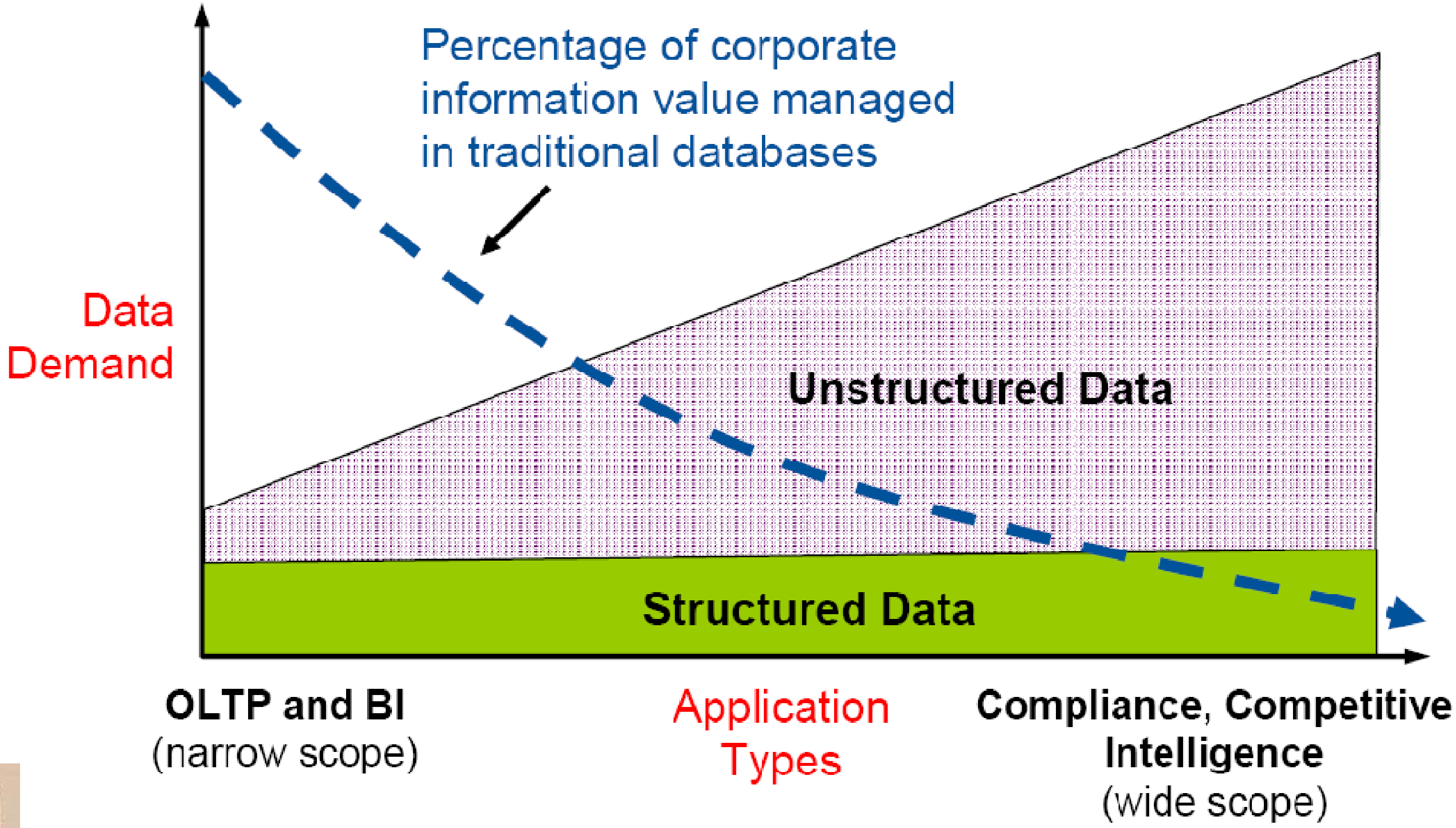
Jeff Fried
VP Advanced Solutions
Jeff.Fried@fastsearch.com

fast

Data Quality

- Still people entering the data
  - – We do typos
  - – We duplicate
  - – We mess up

Unstructured Information in the Warehouse



Gartner

Mark Beyer

Data Consolidation and ETL Studio

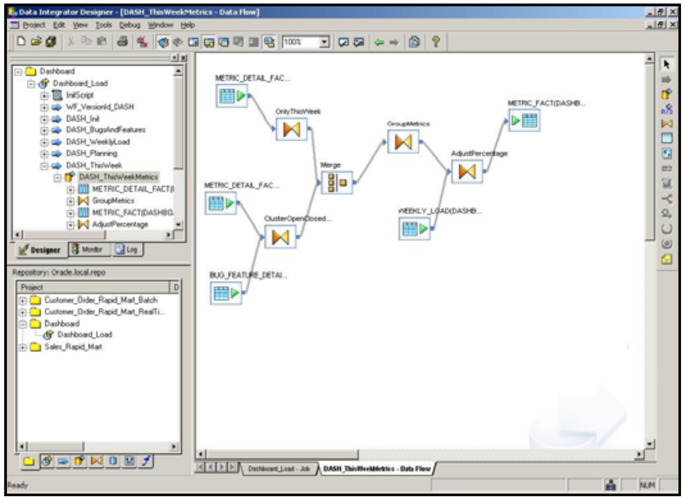
Direct access to RDBMs
for info from some Telco's

XML feed from
other Telco's

XML

Flat files (CSV or fixed)
from the 'laggards'

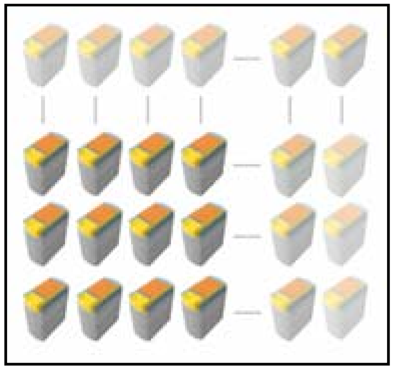
Logic for matching
and cleansing



Logic for ETL

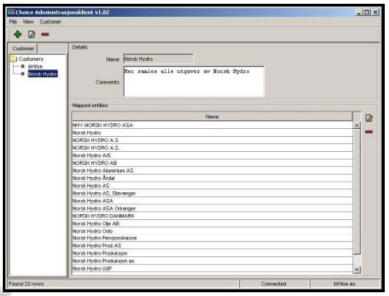
Lookup to ESP for
Matching

Ordered hits (by quality)



Cleansed data
Back to ESP for
next round
Cleansing
or for 'search'

Ambigous
data
(close hits or
unidentified)



Name: Norsk Hydro

Comments: Her samles alle utgaver av Norsk Hydro

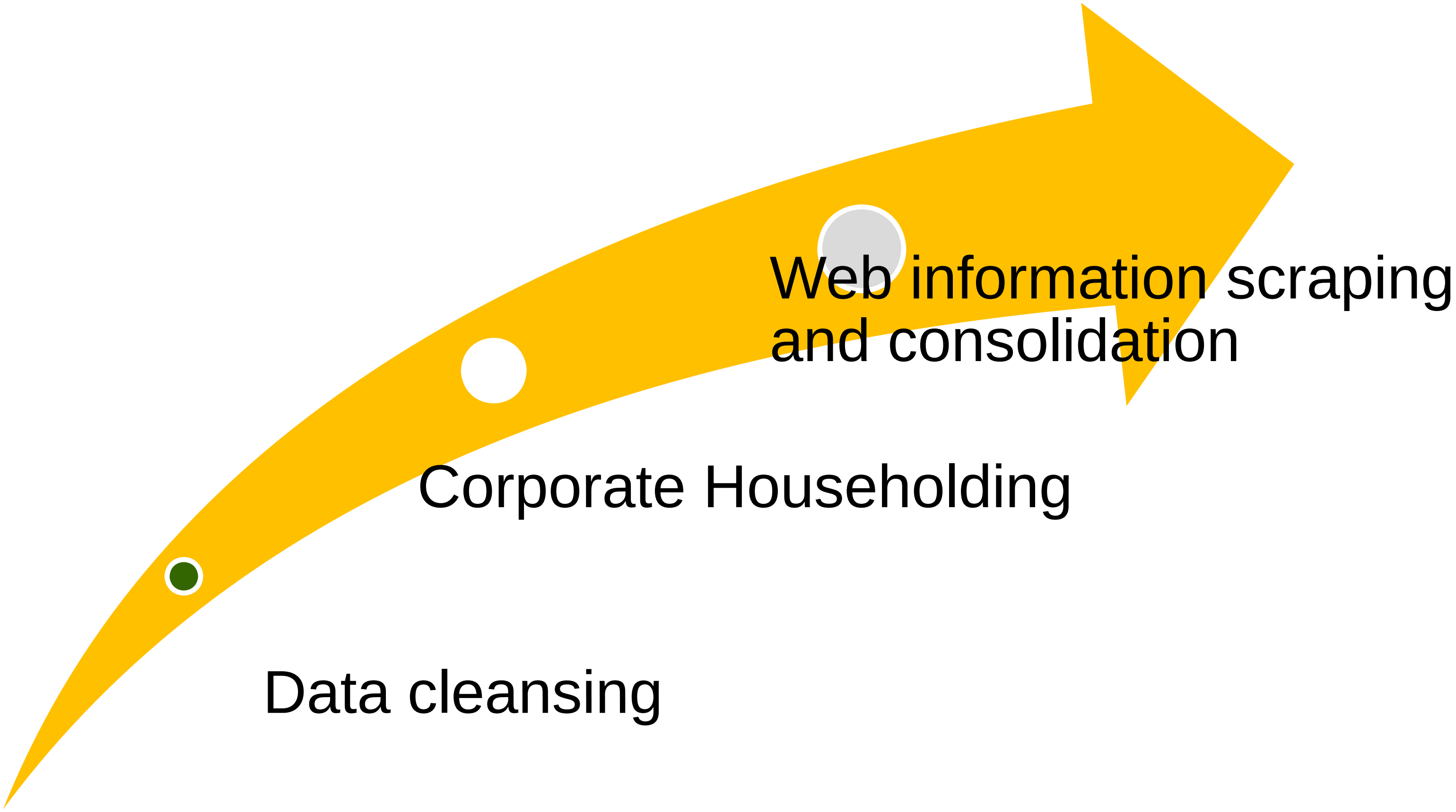
Name
NHY-NORSK HYDRO ASA
Norsk Hydro
NORSK HYDRO A.S
NORSK HYDRO A.S.
Norsk Hydro A/S
NORSK HYDRO AB
Norsk Hydro Aluminium AS
Norsk Hydro Årdal
Norsk Hydro AS
Norsk Hydro AS, Stavanger
Norsk Hydro ASA
Norsk Hydro ASA Orkanger
NORSK HYDRO DANMARK
Norsk Hydro Olje AB
Norsk Hydro Oslo
Norsk Hydro Pensjonskasse
Norsk Hydro Prod.AS
Norsk Hydro Produksjon
Norsk Hydro Produksjon as
Norsk Hydro USP

Found 22 rows | Connected | bWise as

'Error' database for manual inspection,
correction, storage/learning

clean data

Master database for
persistant storage

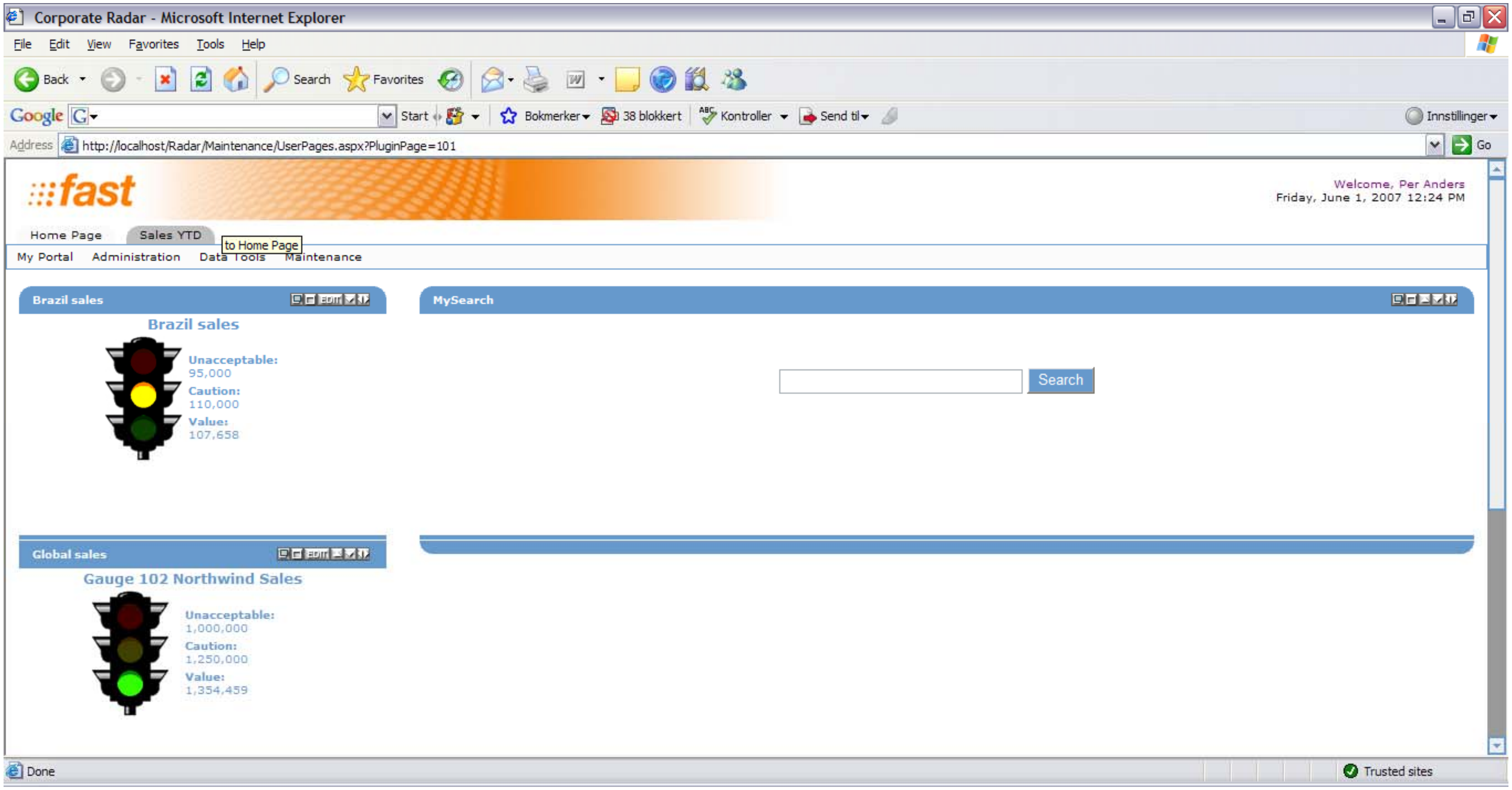


Web information scraping and consolidation

Corporate Householding

Data cleansing

Viewing data, we notice issues



Corporate Radar - Microsoft Internet Explorer

Address http://localhost/Radar/Maintenance/UserPages.aspx?PluginPage=101

Welcome, Per Anders
Friday, June 1, 2007 12:24 PM

Brazil sales

Brazil sales

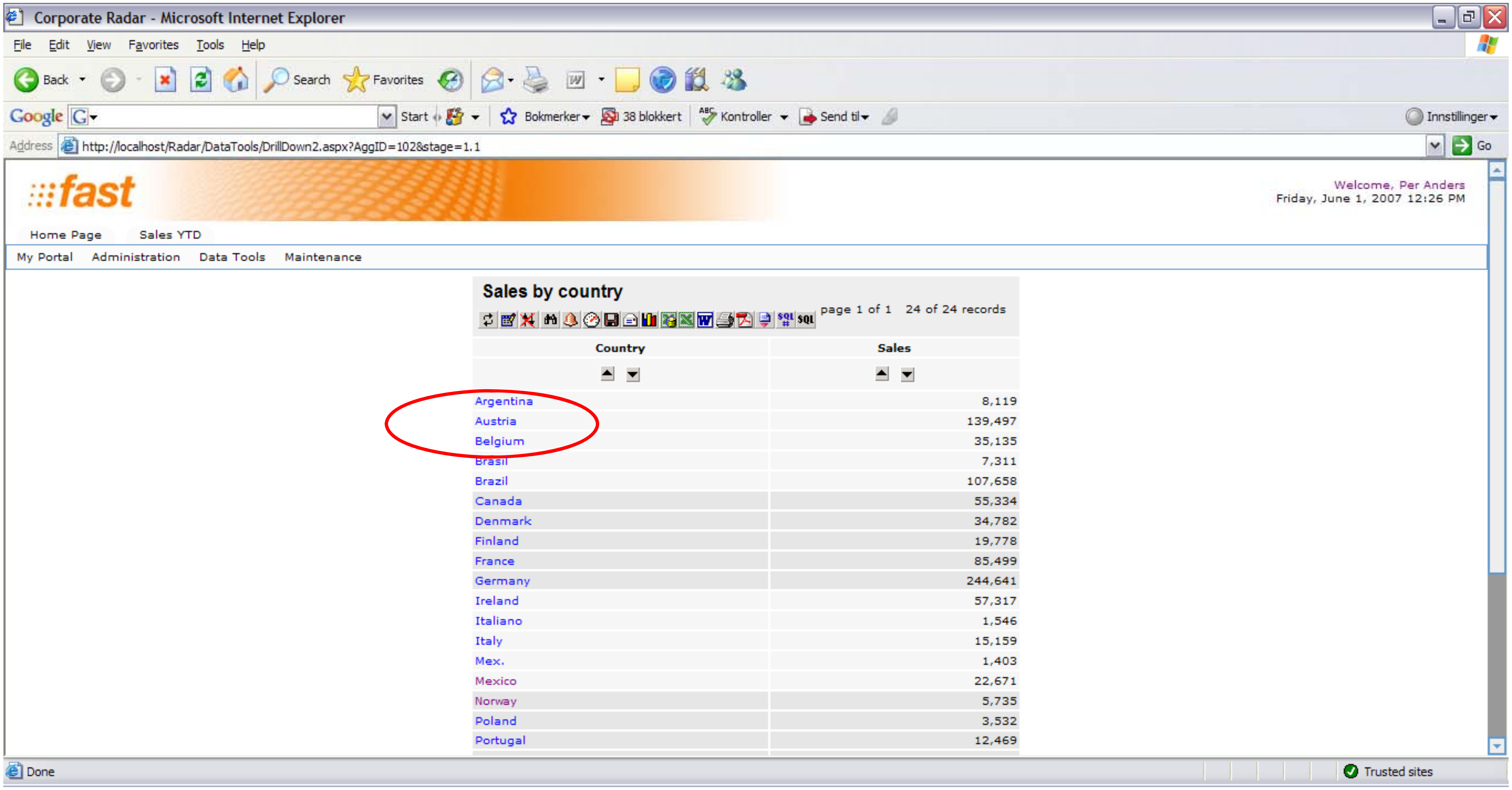
Unacceptable:
95,000
Caution:
110,000
Value:
107,658

Global sales

Gauge 102 Northwind Sales

Unacceptable:
1,000,000
Caution:
1,250,000
Value:
1,354,459

We click on the gauge


Corporate Radar - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Google Start Bokmerker 38 blokkert Kontroller Send til Innstillinger
Address http://localhost/Radar/DataTools/DrillDown2.aspx?AggID=102&stage=1.1
Go
:::fast
Home Page Sales YTD
My Portal Administration Data Tools Maintenance
Done
Trusted sites

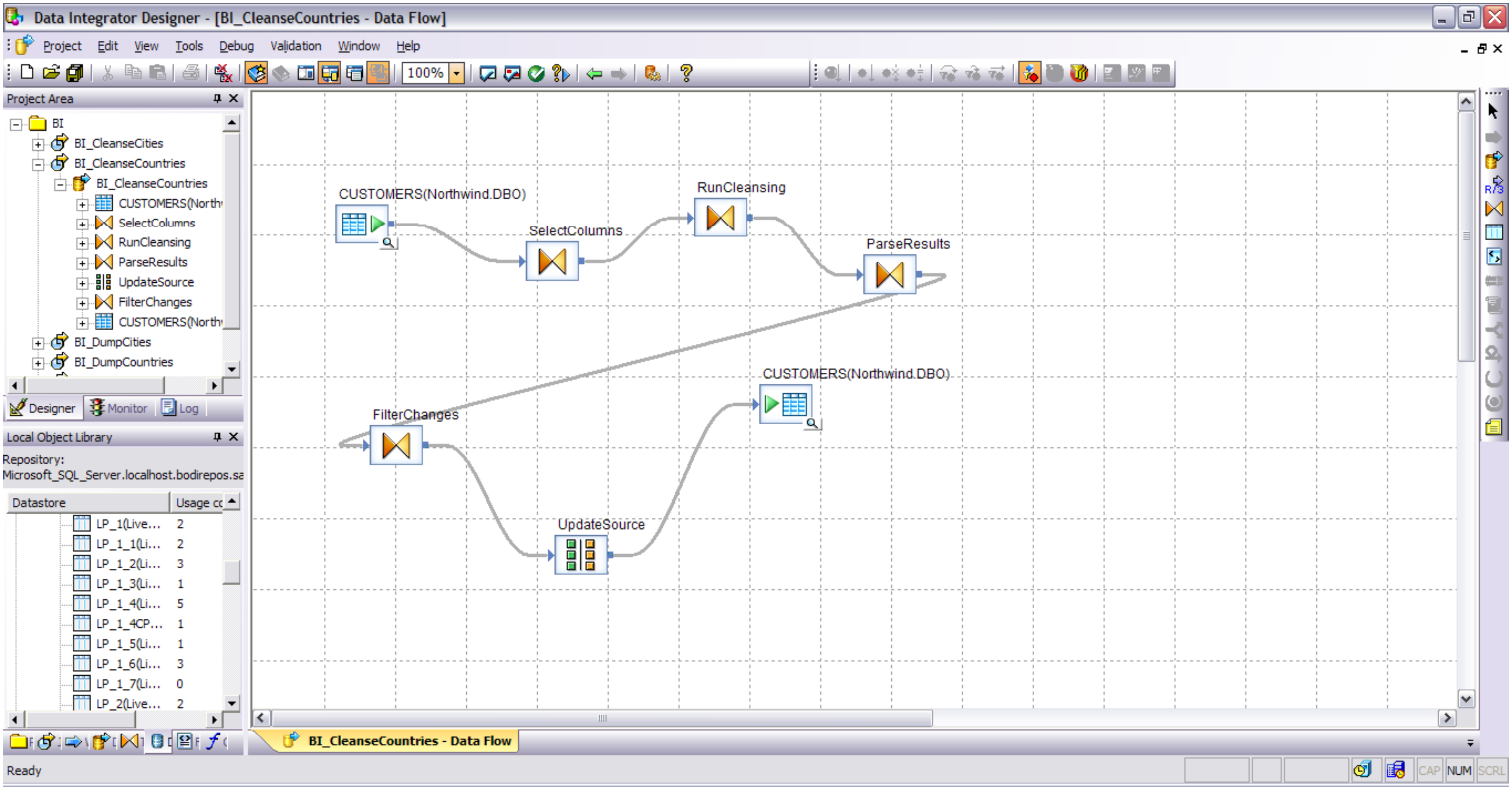
Welcome, Per Anders
Friday, June 1, 2007 12:26 PM

Sales by country

page 1 of 1 24 of 24 records

Country	Sales
Argentina	8,119
Austria	139,497
Belgium	35,135
Brasil	7,311
Brazil	107,658
Canada	55,334
Denmark	34,782
Finland	19,778
France	85,499
Germany	244,641
Ireland	57,317
Italiano	1,546
Italy	15,159
Mex.	1,403
Mexico	22,671
Norway	5,735
Poland	3,532
Portugal	12,469

We cleanse the data

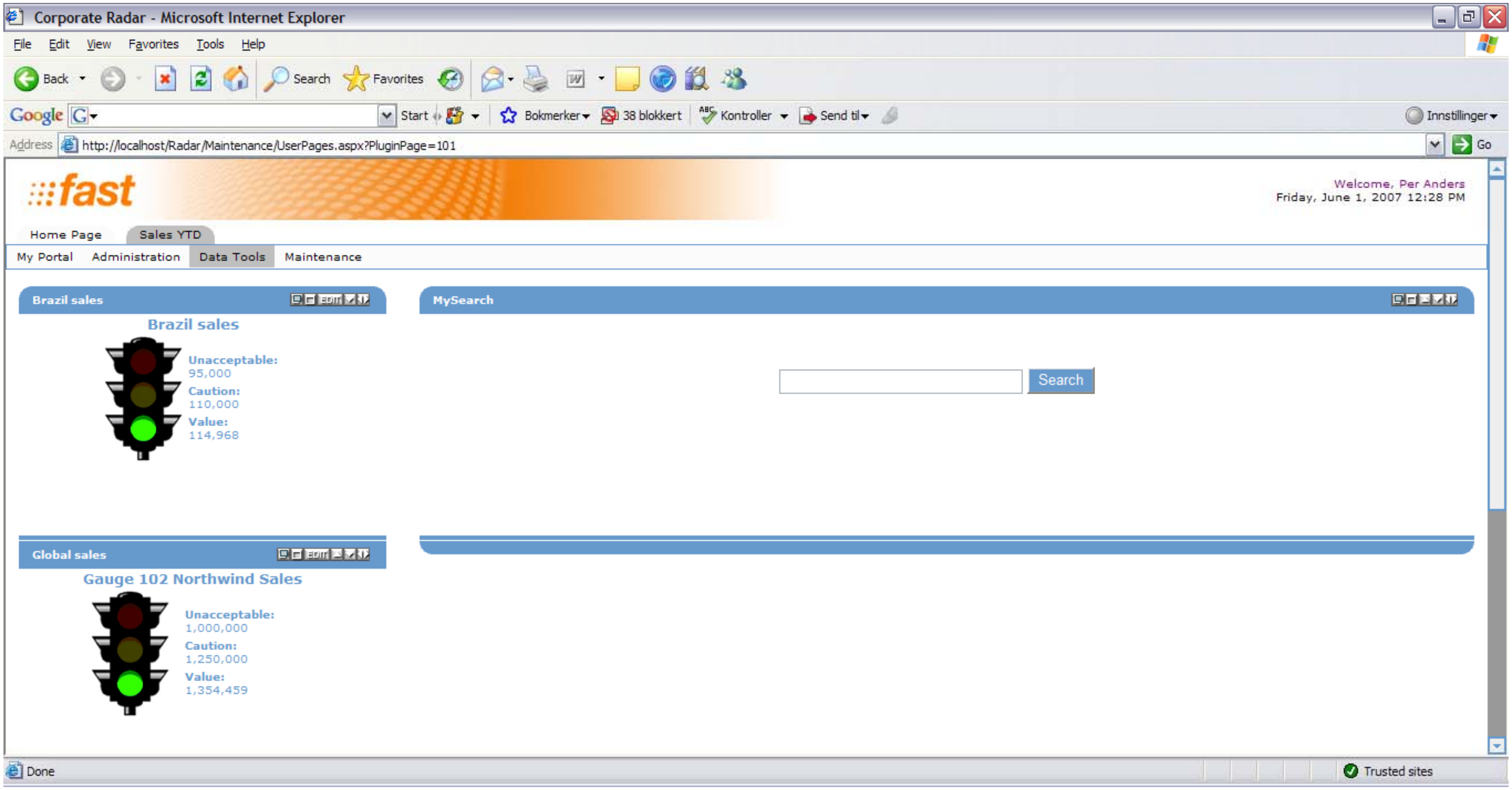


- BI
  - BI_CleanseCities
  - BI_CleanseCountries
    - BI_CleanseCountries
      - CUSTOMERS(North
      - SelectColumns
      - RunCleansing
      - ParseResults
      - UpdateSource
      - FilterChanges
      - CUSTOMERS(North
  - BI_DumpCities
  - BI_DumpCountries

Repository:
Microsoft_SQL_Server.localhost.bodirepos.sa

Datastore	Usage cc
LP_1(Live...	2
LP_1_1(Li...	2
LP_1_2(Li...	3
LP_1_3(Li...	1
LP_1_4(Li...	5
LP_1_4CP...	1
LP_1_5(Li...	1
LP_1_6(Li...	3
LP_1_7(Li...	0
LP_2(Live...	2

Now look at the result



Corporate Radar - Microsoft Internet Explorer

Address http://localhost/Radar/Maintenance/UserPages.aspx?PluginPage=101

Welcome, Per Anders
Friday, June 1, 2007 12:28 PM

Brazil sales

Brazil sales

Unacceptable:
95,000
Caution:
110,000
Value:
114,968

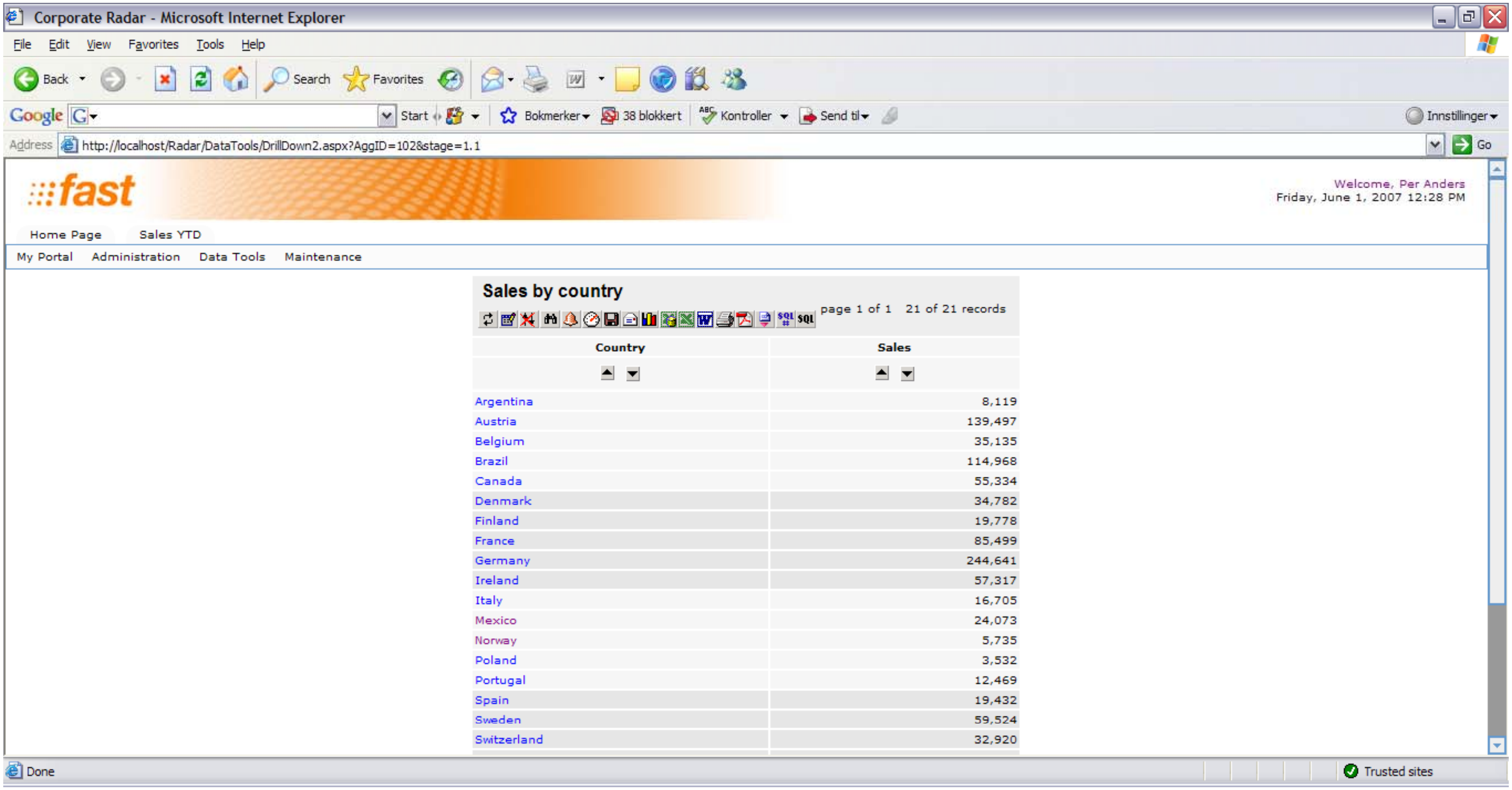
Global sales

Gauge 102 Northwind Sales

Unacceptable:
1,000,000
Caution:
1,250,000
Value:
1,354,459

:::fast

Details


Corporate Radar - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Google Start Bokmerker 38 blokkert Kontroller Send til Innstillinger
Address
Go
:::fast
Home Page Sales YTD
My Portal Administration Data Tools Maintenance
Done
Trusted sites

http://localhost/Radar/DataTools/DrillDown2.aspx?AggID=102&stage=1.1

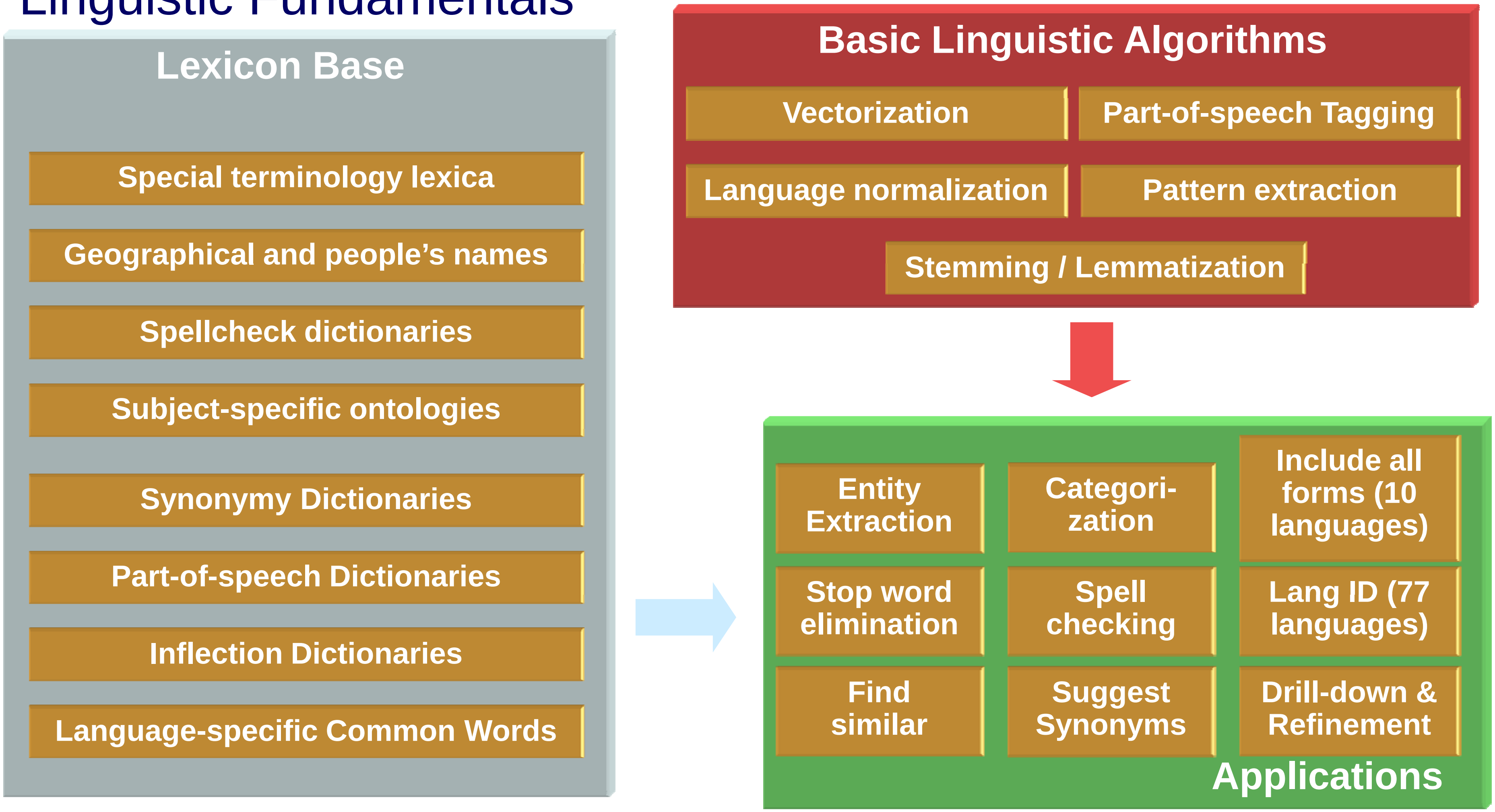
Welcome, Per Anders
Friday, June 1, 2007 12:28 PM

Sales by country

page 1 of 1 21 of 21 records

Country	Sales
Argentina	8,119
Austria	139,497
Belgium	35,135
Brazil	114,968
Canada	55,334
Denmark	34,782
Finland	19,778
France	85,499
Germany	244,641
Ireland	57,317
Italy	16,705
Mexico	24,073
Norway	5,735
Poland	3,532
Portugal	12,469
Spain	19,432
Sweden	59,524
Switzerland	32,920

Linguistic Fundamentals



Lexicon Base

- Special terminology lexica
- Geographical and people’s names
- Spellcheck dictionaries
- Subject-specific ontologies
- Synonymy Dictionaries
- Part-of-speech Dictionaries
- Inflection Dictionaries
- Language-specific Common Words

Basic Linguistic Algorithms

- Vectorization
- Part-of-speech Tagging
- Language normalization
- Pattern extraction
- Stemming / Lemmatization

Applications

- Entity Extraction
- Categori-zation
- Include all forms (10 languages)
- Stop word elimination
- Spell checking
- Lang ID (77 languages)
- Find similar
- Suggest Synonyms
- Drill-down & Refinement

Q&A

Thank you!

fast

find the real value of search

Jeff Fried
VP Advanced Solutions
Jeff.Fried@fastsearch.com