9th Regional JODI Training Workshop
25-27 February 2014, Baku, Azerbaijan

JODI Oil Data Quality Assessment

Data Validation (Data Techniques),
Consistency with other Energy Statistics
Availability of Metadata

Revised and Presented by Dr. Pantelis Christodoulides (OPEC)

Prepared by Mr. Leonardo Souza
Revised by Dr. Ryo Eto (IEEJ)

Data Quality Evaluation

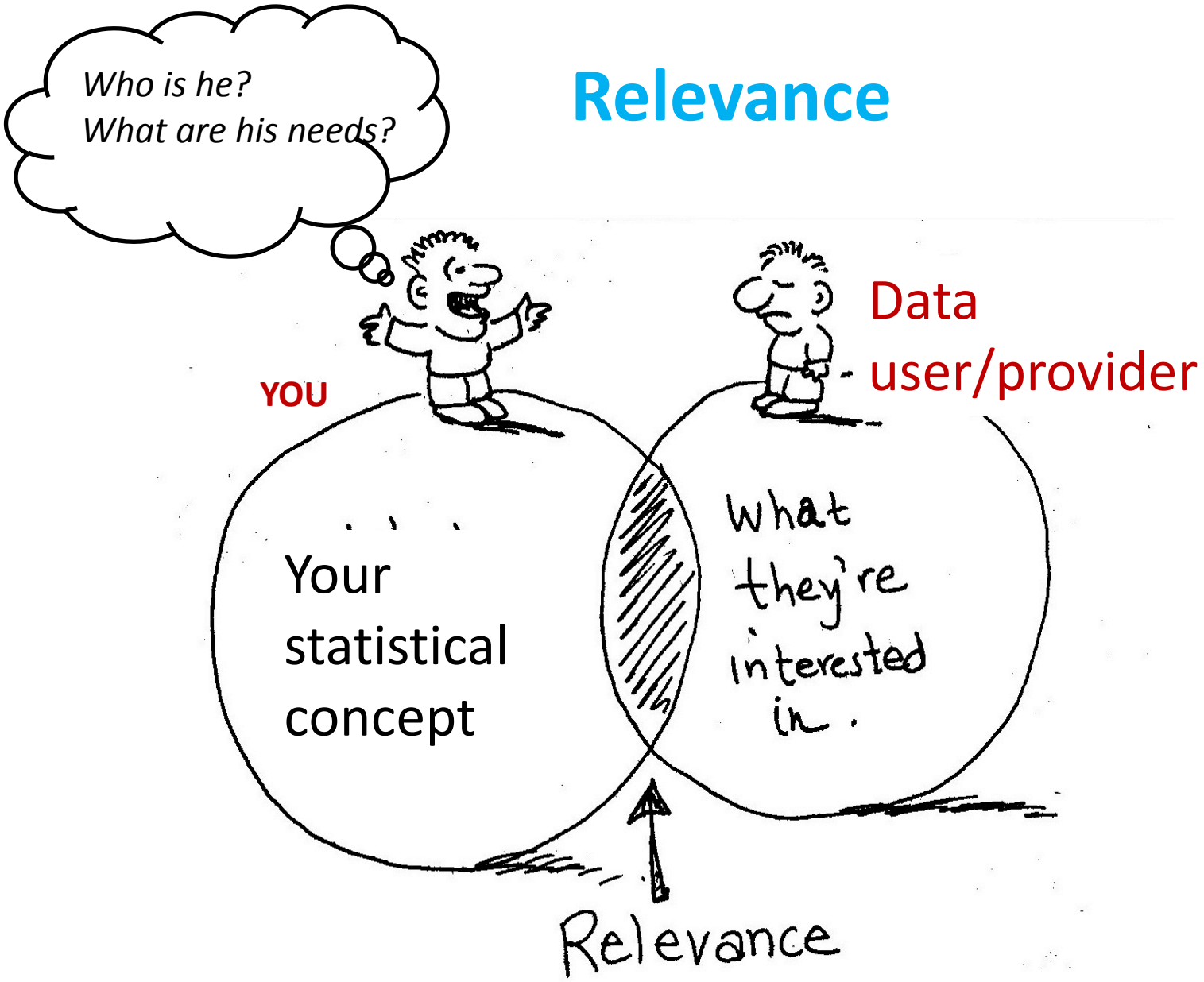
Data quality is a multi-dimensional concept

- Relevance (of statistical concepts)
- Accuracy
- Timeliness
- Accessibility and clarity of information
- Comparability of statistics
- Coherence
- Completeness/coverage
- Cost and burden
- OPEC practices
- Smiley faces
- Metadata

Before JODI, M-2 data was not collected for crude oil and petroleum product flows for such a number of countries

It requires **changes in data collection** practices and further **collaboration** between national statistical offices and country energy authorities

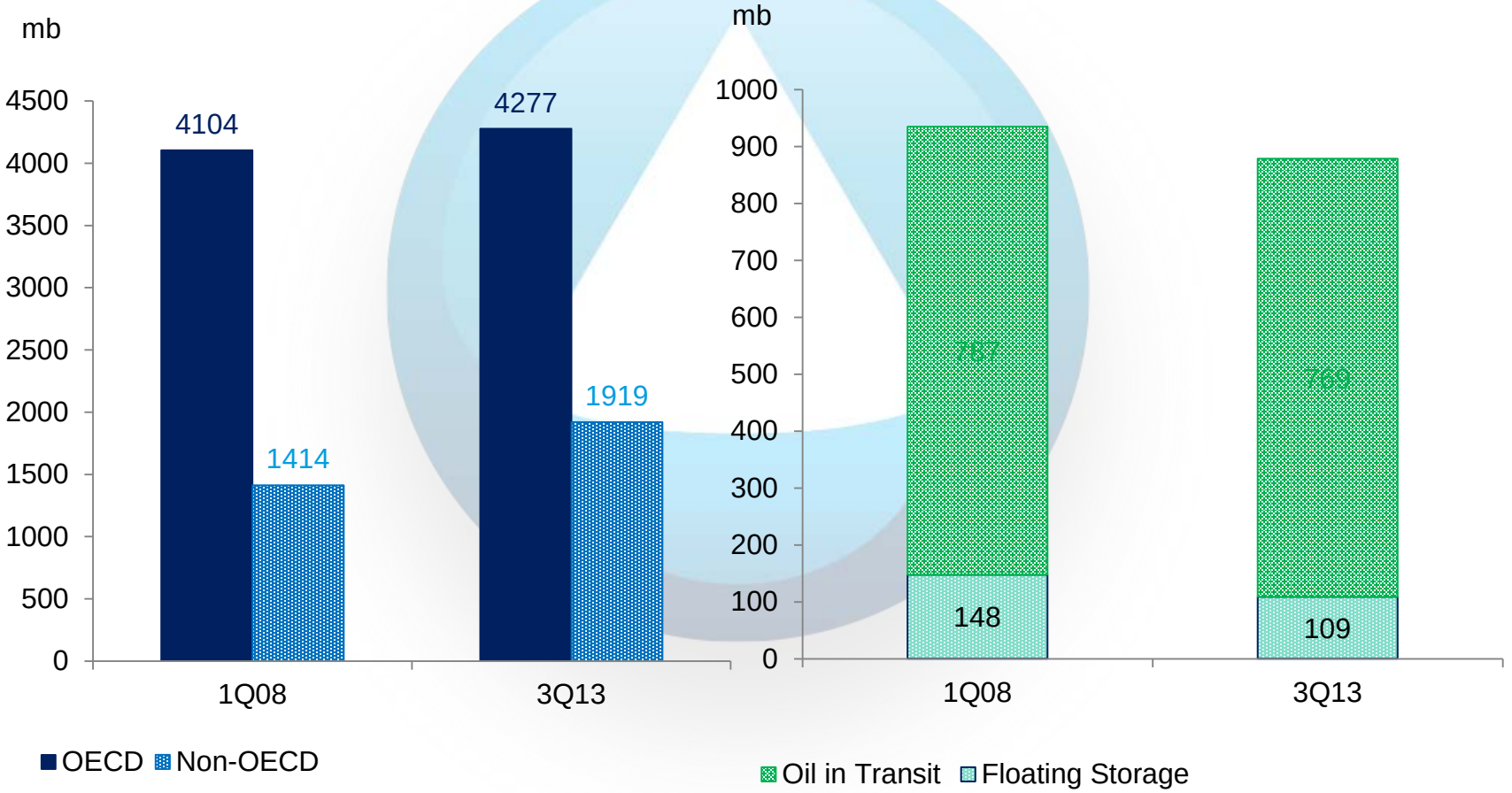
Relevance


Who is he?
What are his needs?
YOU
Data user/provider
Your statistical concept
What they're interested in.
Relevance

Relevance

- Relevance (in statistics) is assured when statistical **concepts meet** current and potential users' **needs**
- Identification of the users and their expectations is a pre-requisite
- Consult with oil companies in the country
  - How many are they? How important is each of them?
  - Listen to their expectations and needs (synthesize and prioritize)
  - Convince them to follow definitions, methodology, classifications (if not possible, at least keep a record of discrepancies)
- Consumer–Producer dialogue
  (JODI conferences: Egypt, Mexico City and Bali)

Relevance – example oil stocks



Accuracy

Accuracy

Accuracy is defined as the **closeness between the computations or estimates and the (unknown) true population value (benchmark???)**

Assessing the accuracy of an estimate involves analysing the total error associated with the estimate: Bias (+ or -?) and standard deviation (when possible)

- Sampling errors and non-sampling errors
- Sampling errors: due to problems in the design of a sample survey
- Sampling???

Accuracy

- Non-sampling errors
  - Coverage errors
  - Measurement errors
  - Processing errors
  - Non-response errors
  - Model assumption errors

- Country level: Report collected information – revisions
- International level: Revision of the time series

Accuracy

- Essential characteristic of an ideal database
- Closely related to readability & usability of database
- Usually negatively correlated to timeliness & completeness

Accuracy

- Assessment of accuracy both by international organizations & national administrations
- Data accuracy verification techniques that can be applied to JODI oil submissions
- Combination of checks/methods optimal
- Provide only indication of accuracy

Balance check

- Primary oil
- Secondary (oil products)

Country _____

Month _____ *Unit :* _____

	Crude Oil	NGL	Other	Total (1)+(2)+(3)		Petroleum Products								
						LPG	Naphtha	Gasoline	Total Kerosene	*Of which: Jet Kerosene*	Gas/ Diesel Oil	Fuel Oil	Other Products	Total Products (5)+(6)+(7) +(8)+(10) +(11)+(12)
	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)	*(9)*	(10)	(11)	(12)	(13)
+ Production					+ Refinery Output									
+ From Other sources					+ Receipts									
+ Imports					+ Imports									
- Exports					- Exports									
+ Products Transferred /Backflows					- Products Transferred									
- Direct Use					+ Interproduct Transfers									
- Stock Change					- Stock Change									
- Statistical Difference	0	0	0	0	- Statistical Difference	0	0	0	0	0	0	0	0	0
= Refinery Intake					= Demand									
Closing stocks					Closing stocks									

Balance check

Statistical Difference

		Crude Oil	NGL	Other	Total			Petroleum Products							
							LPG	Naphtha	Gasoline	Jet Kerosene	Kerosene	Gas/ Diesel Oil	Fuel Oil	Other Products	Total Products
Production						Refinery Output									
Imports						Receipts									
Exports						Imports									
Direct Use						Exports									
Transfers						Transfers									
Stocks	Closing					Stocks	Closing								
	Change						Change								
Statistical Difference*						Statistical Difference									
Refinery Intake						Demand									

Balance check – primary oil

- Calculated refinery intake ≈ reported refinery intake
- Calculated refinery intake = production + from other sources + imports –exports + products transferred/backflows – direct use – stock change
- Calculated refinery intake - reported refinery intake (statistical difference) should ideally be relatively small
- Statistical difference should be small in relative and absolute terms

Balance check – primary oil

		Crude oil (kt)
+	Production	2
+	From other sources	0
+	Imports	3681
-	Exports	0
+	Products transferred/backflows	0
-	Direct use	200
-	Stock change	-295
-	Statistical difference: calculated – reported refinery intake	228
=	Refinery intake	3550
%	Percentage statistical difference	6.4%

Balance check – secondary oil products

- Calculated demand ≈ reported demand
- Calculated demand = refinery output + receipts + imports –exports - products transferred + interproduct transfers – stock change
- Calculated demand - reported demand should ideally be relatively small
- Statistical difference should be small in relative and absolute terms

Balance check – secondary oil products

		Total products (kt)
+	Refinery output	126
+	Receipts	0
+	Imports	59
-	Exports	13
-	Products transferred	0
+	Interproduct transfers	0
-	Stock change	-2
-	Statistical difference: calculated – reported demand	-2
=	Demand	176
%	Percentage statistical difference	-1%

Balance check

- Applicable only if all data are complete and reliable
- Large deviations would require review and/or verification/correction
- Re-submission in the form of revisions during the following month
- Application on every column of the JODI oil questionnaire
- Range of 5% quite large for physical quantities (1% or even 0.5%)

Internal consistency check

- Another indicator of accuracy
- Fuel checks – total oil products should be equal to the sum of reported products (excluding jet fuel)
- Statistician should ensure that this property holds in all submissions

Country _____

Month _____

Unit : _____

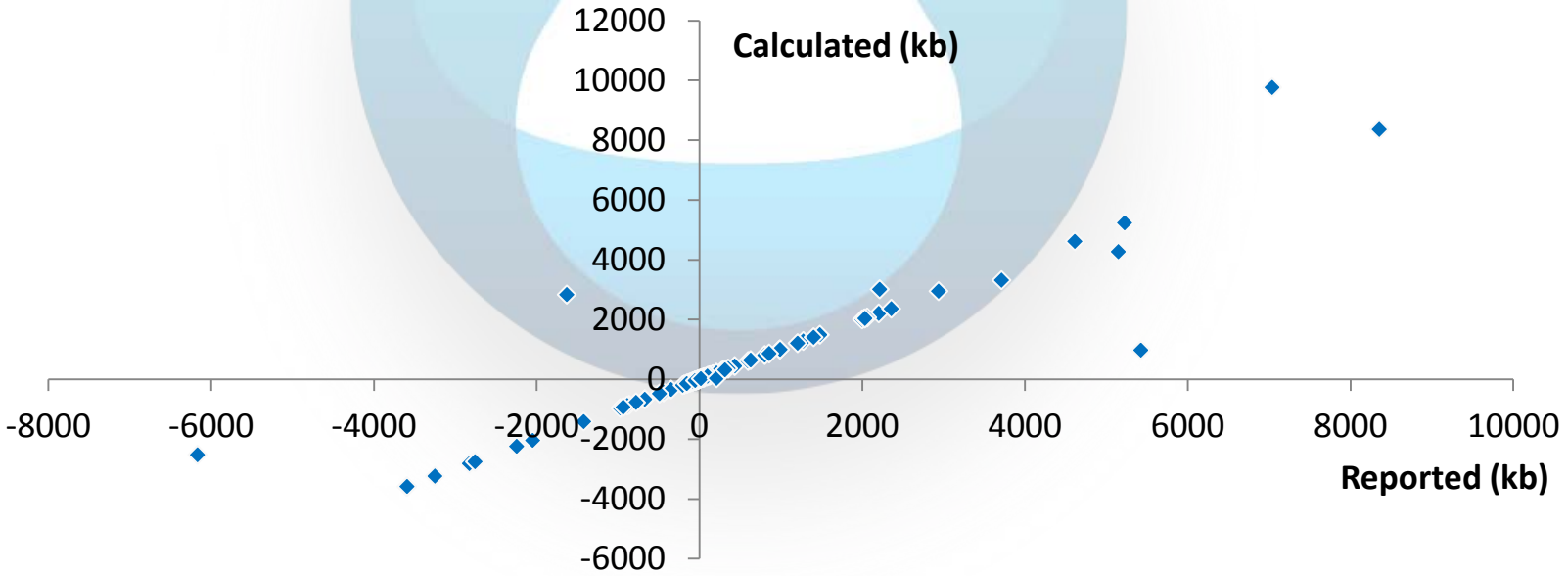
	Crude Oil	NGL	Other	Total (1)+(2)+(3)		Petroleum Products								
						LPG	Naphtha	Gasoline	Total Kerosene	*Of which: Jet Kerosene*	Gas/ Diesel Oil	Fuel Oil	Other Products	Total Products (5)+(6)+(7)+(8)+(10)+(11)+(12)
	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)	*(9)*	(10)	(11)	(12)	(13)
+ Production					+ Refinery Output									
+ From Other sources					+ Receipts									
+ Imports					+ Imports									
- Exports					- Exports									
+ Products Transferred /Backflows					- Products Transferred									
- Direct Use					+ Interproduct Transfers									
- Stock Change					- Stock Change									
- Statistical Difference	0	0	0	0	- Statistical Difference	0	0	0	0	0	0	0	0	0
= Refinery Intake					= Demand									
Closing stocks					Closing stocks									

Internal consistency check

- Automatic checks are incorporated in the questionnaire to point out inconsistencies
- Fuel checks – total oil products should be equal to the sum of reported products (excluding jet fuel)
- Statistician should ensure that this property holds in all submissions
- Indication of misreporting of data
- Example of a country with imports of LPG, fuel and other products (kt)
  - Reported total products imports (1021) < LPG imports (59) + fuel oil imports (60) + other product imports (10) = 1029

Internal consistency check

- Stock checks
- Stock change (M) = Closing stock level (M) – Closing stock level (M-1)
- Calculated stock change ≈ reported stock change
- Tolerance range of 5% (in relation to closing stocks)

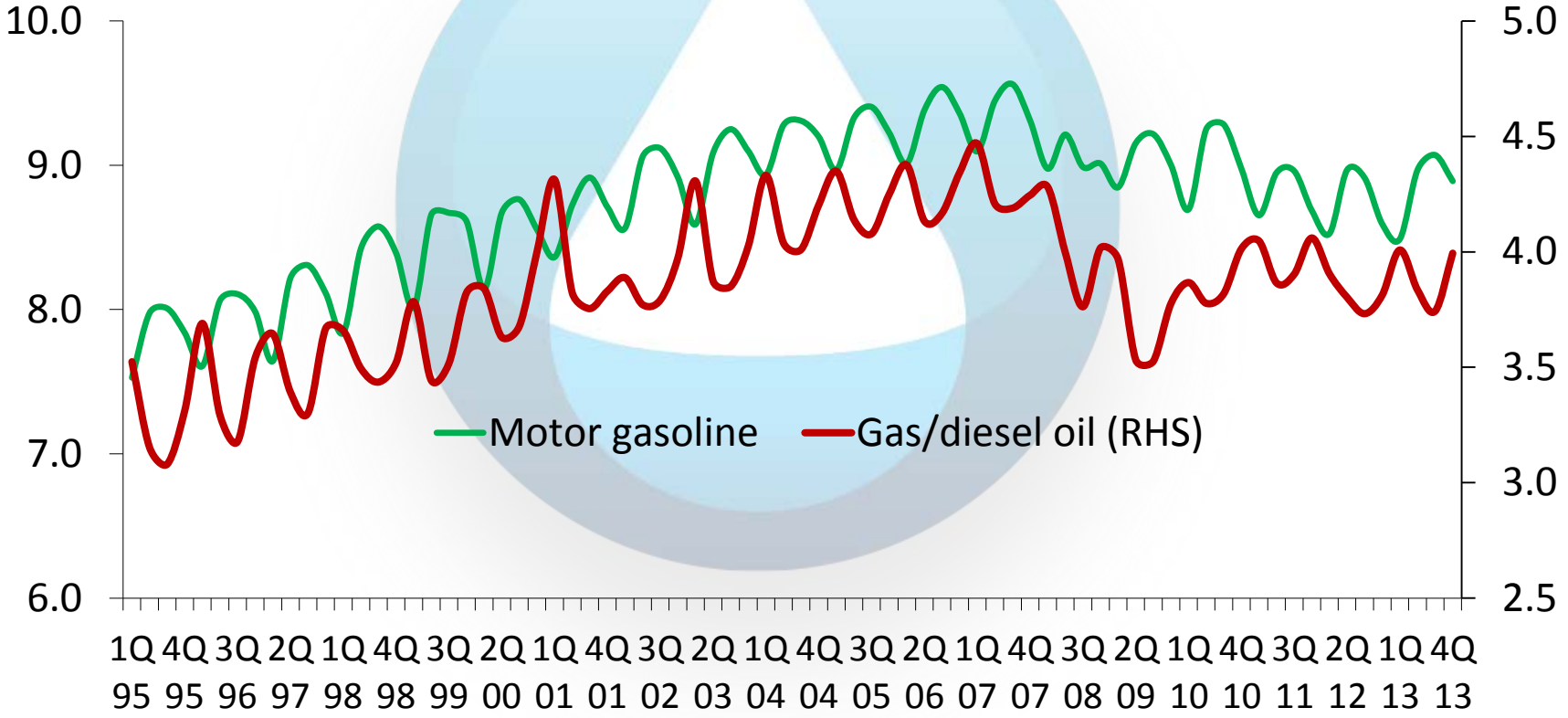


Time series check

- Similar percentage changes (m-o-m, y-o-y) indicator of "good" data quality
- "Unusually large" percentage changes may require verification of data
- Seasonality of oil data
- Trade data
- Refinery intake/output check
- Refinery yield (%) = Refinery output (total oil products)/Refinery intake (total primary products) *100

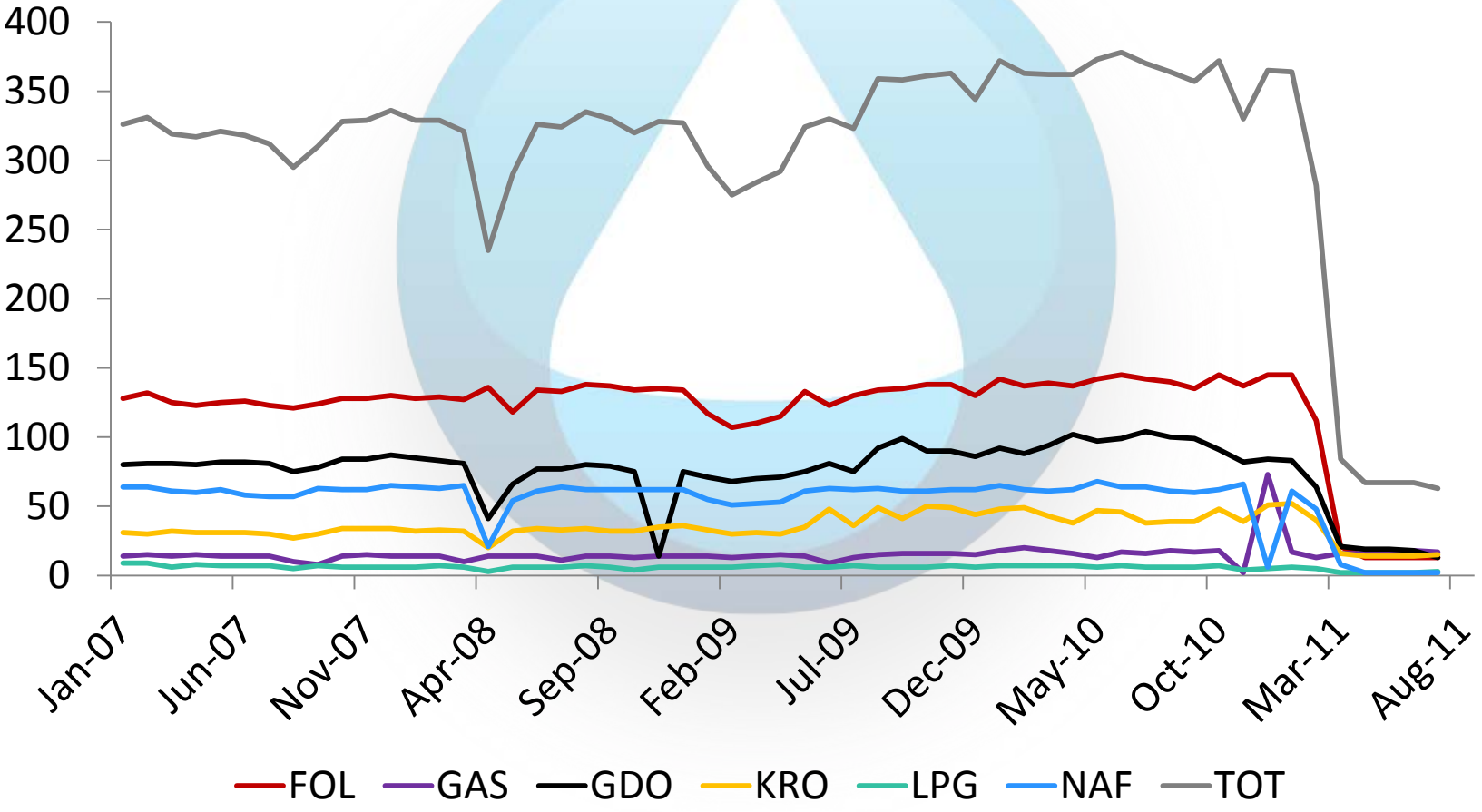
Time series check – example 1

- Seasonality in gasoline & distillate consumption (mb/d)



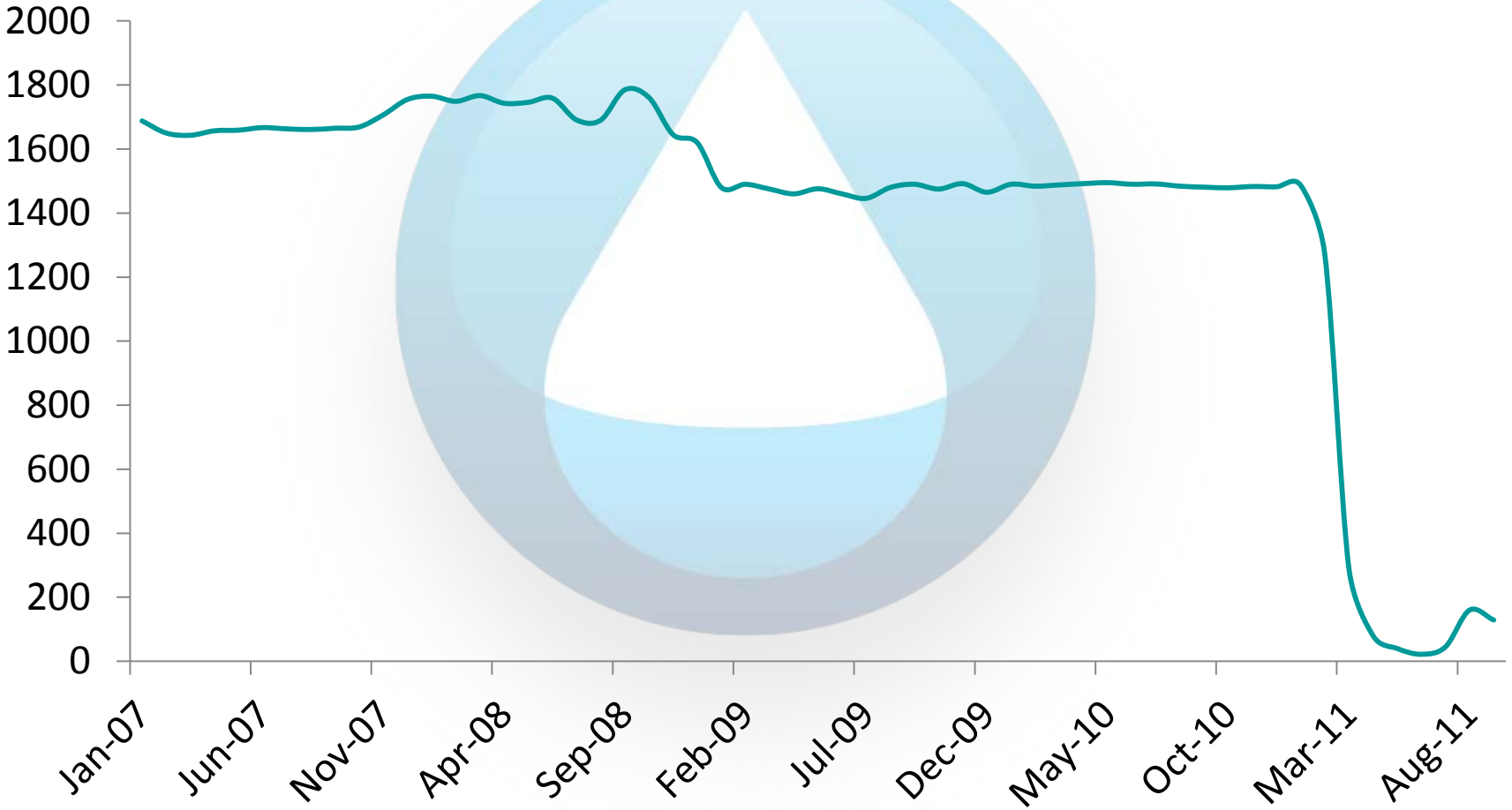
Time series check – example 2

- Refinery output (kb/d)



Visual check – example

- Crude oil production (kb/d)



Timeliness

Users want the **latest** data that are published **frequently** and **on time** at pre-established dates

Managing

- Data collection
- Editing
- Consolidation
- Dissemination

Accessibility and clarity of information

Statistical data are most valuable when they are

- Easily accessible by users
- Available in the form users desire
- There is adequately documented **metadata**

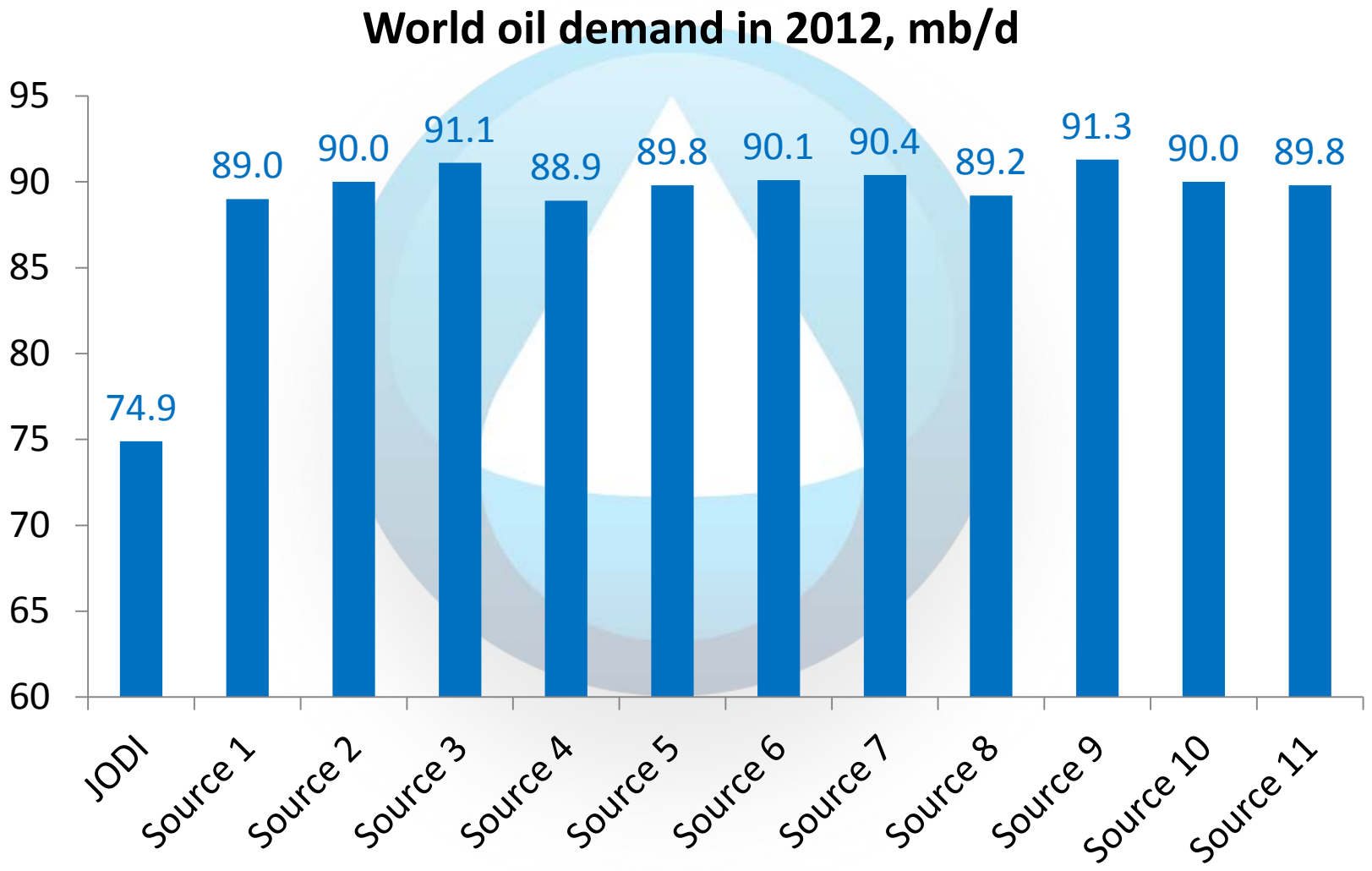
Assistance in using and interpreting the statistics should also be forthcoming from the providers

Comparability of statistics

Statistics for a given characteristic have the greatest usefulness when they enable reliable comparisons of values across space and over time

Providing comparable data makes it possible to publish regional and world totals

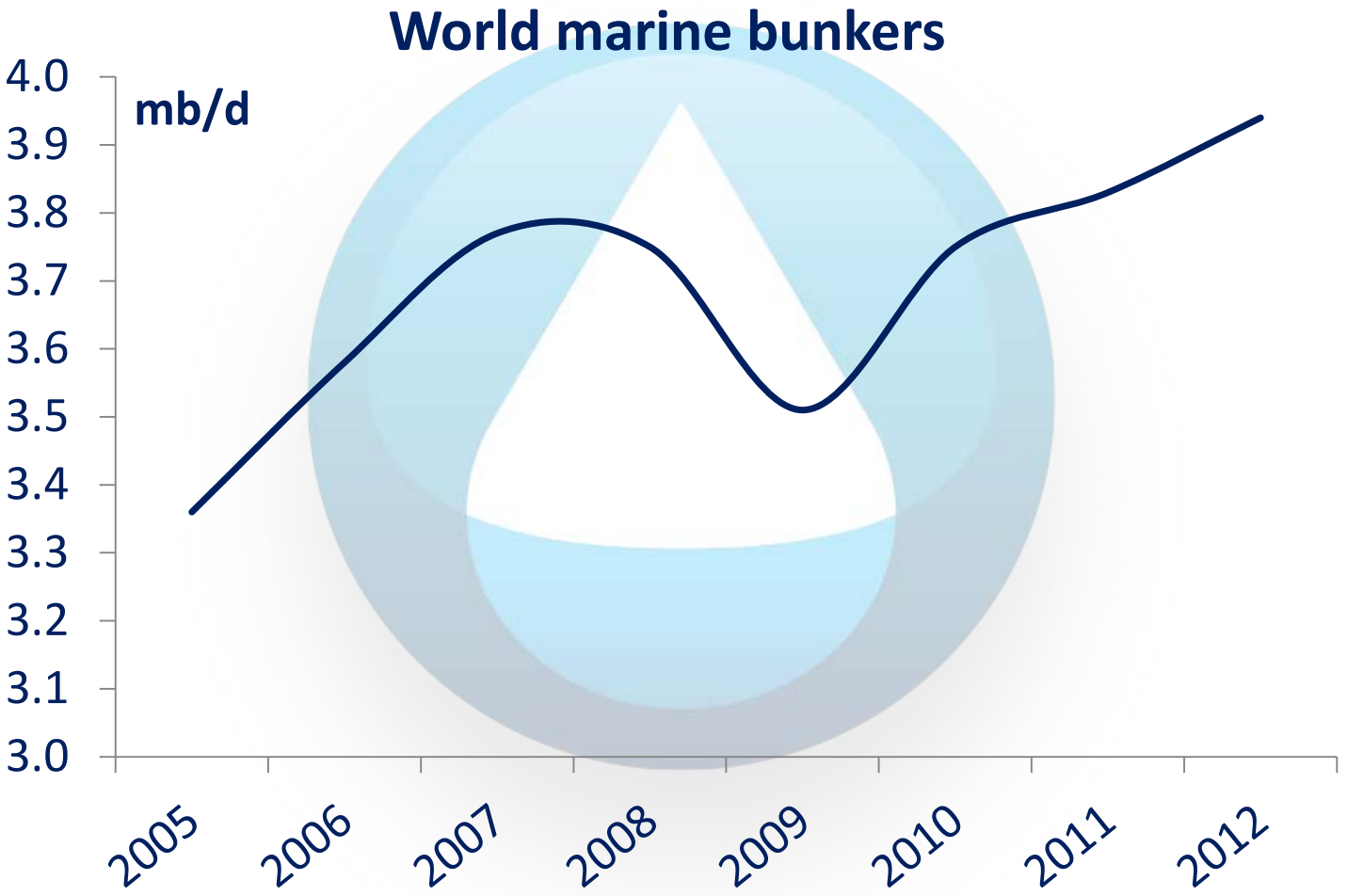
Comparability of statistics



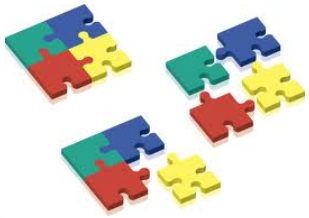
For comparability the following are needed

- Unified definitions
- Knowledge of the conversion factors at country level
- Common unit of measurement
- Unified methodology
- Timely submission of data

Comparability of statistics



Coherence



Coherence

- Coherence is the measure of the extent to which **one set of statistical characteristics agrees with another and can be used together** (with each other) **or as an alternative** (to each other)

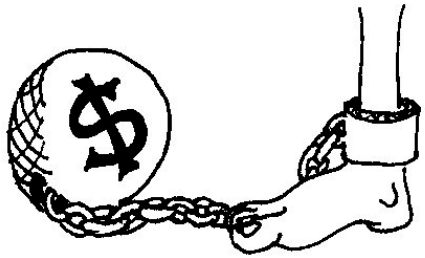
- To assess the coherence of the statistics collected, comparisons with other statistics relating to the JODI data could be made, e.g. comparisons with monthly, quarterly and yearly oil statistics of international organisations

Completeness/ coverage

The component of completeness reflects the extent to which the statistical system in place answers the users' needs and priorities by comparing all user demands with the availability of statistics

- How many participating countries
- How complete the questionnaire/ share of completed cells to the total numbers of cells in the questionnaire

Cost and burden



Cost and burden

- The quality of the data will be affected by available resources to collect, analyze and store energy statistics
- Although not measures of quality, they are positively correlated with quality
- Costs: Office space, utility bills, staff-hours involved, software, etc.
- Response burden: Simplest way to measure is the time spent by the respondent to provide information
- A compromise between quality and cost and burden must be achieved

Cost and burden

- Functions of cost/burden
  - Collection of data
  - Level of disaggregation
  - Time lags, frequencies of data
  - Applied methodologies

Data quality evaluation
Common practices at OPEC

Common practices at OPEC

- Balance check
- Internal consistency check
- Times series check
- Visual check
- Comparison with other monthly data
- Comparison with annual data
- Focus on maximizing information available in metadata

- Only flows of the short 42-points questionnaire

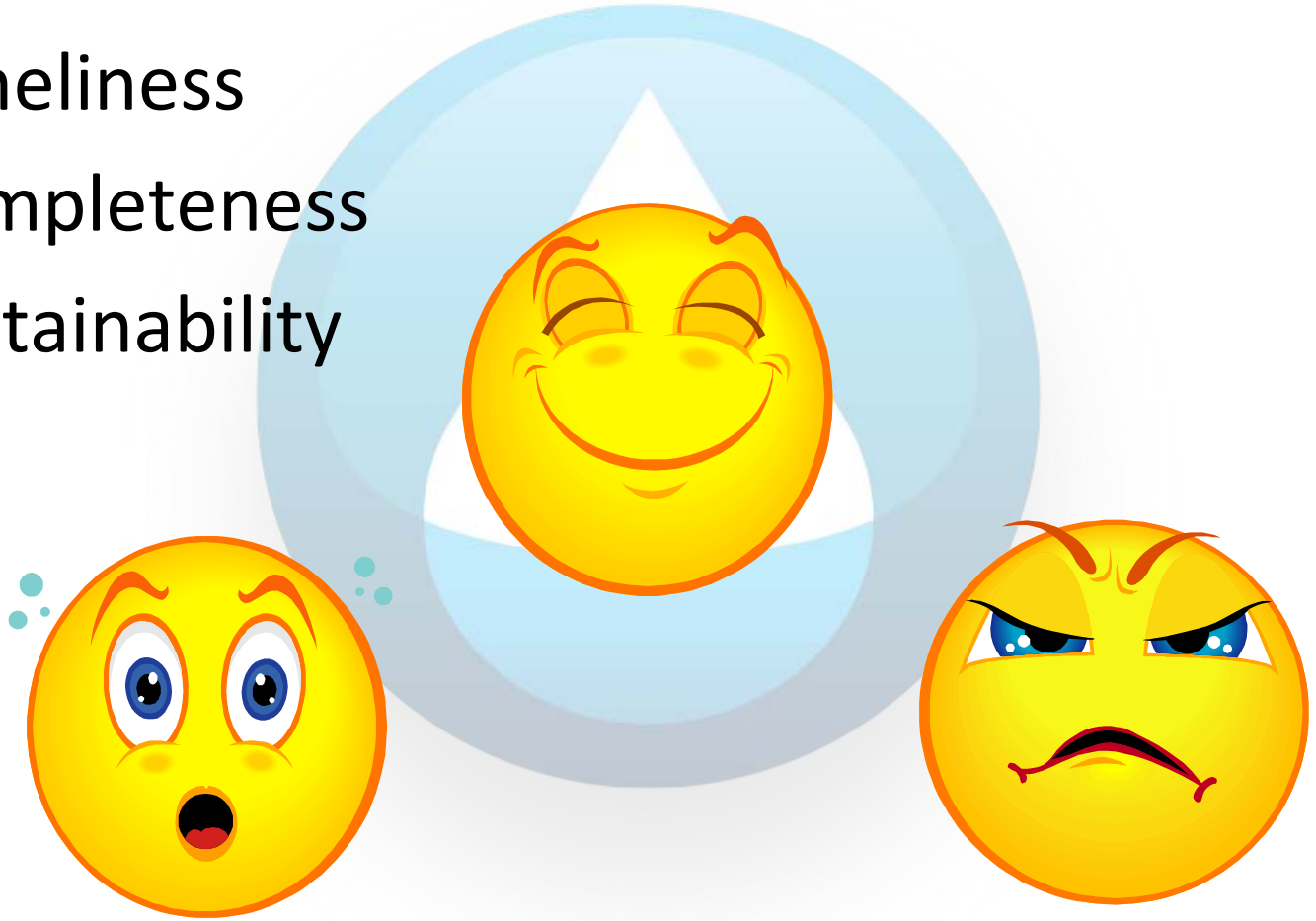
Common practices at OPEC – crude

Balance = Production + Imports– Exports - Refinery Intake – Direct Burning ± Stock Changes

Country	2011	2012	1Q13
1	-1%	7%	3%
2	3%	2%	2%
3	-1%	-2%	-2%
4	0%	2%	--
5	-5%	-4%	-3%
6	-2%	-1%	0%
7	1%	--	--
8	0%	-1%	7%
9	-1%	-1%	0%
10	0%	0%	0%
11	0%	0%	--
12	-5%	-6%	-5%

Smiley faces

- Timeliness
- Completeness
- Sustainability



Smiley faces (timeliness)

- The JODI database is expected to be updated regularly.
- The timeliness indicates whether submissions were submitted at the expected deadline

"good" when 6 submissions received within 45 days after the end of the reference month

"fair" when 4 or 5 submissions received

"less reliable" when less than 4 submissions received

Smiley faces (completeness)

Completeness measures the number of expected data points out of the maximum 42 in the JODI questionnaire which are filled in

"good" when more than 90% of the data are given for production, stock change/closing and demand

"fair" when between 60% and 90% of the data are given

"less reliable" when less than 60% of the data are given

Smiley faces (sustainability)

Sustainability is the number of the monthly JODI data (timely) submissions evaluated 2 months after the end of the six-month period

"good" if the 6 questionnaires have been submitted

"fair" if 4 or 5 questionnaires have been submitted

"less reliable" when less than 4 questionnaires have been submitted

Color code assignment

Color code assignment

- **Blue:** A blue background indicates that results of the assessment show reasonable levels of comparability
- **Yellow:** A yellow background indicates that the metadata should be consulted
- **White:** A white background indicates that data has not been assessed
- **Purple:** data under verification

Color code assignment approach by OPEC

- All checks are carried out for flows of the 42 point oil questionnaire
- Comparison with official annual data
- No assessments based on comparisons to secondary sources

Metadata

Metadata

- The simplest definition of metadata is that it is **data about data**. More specifically information (data) about a particular content (data)
- Metadata describes **how and when and by whom** a particular set of data was collected; how the data is **formatted**
- Metadata **must be updated** when there is a change in resource it describes
- It can be useful to **keep** metadata even when the resource no longer exists
- Metadata **enhances data transparency** and is essential for understanding information stored in a database

Thank you

For more information at

www.jodidata.org