Understanding and Applying the Basic Building Blocks of Data Quality

Emily Carnahan, M&E Officer
Emily Beylerian, M&E Data Analyst

PATH/Eric Becker

PATH

How would you go about assessing the quality of this data?

SiteID	Site	Report_Period	PMCTA_1stVisit_ANC	MCTA_ReVisit_AN	PMCTANCClientsT	PMCTHB7	PMCTIPT1	PMCTIPT2	PMCTANCClients4
729	Asumbi Mission Hospital	01-Jan-11	34	55	89	0	14	14	16
625	Got Kojowi Health Centre	01-Jan-11	0	0	0	0	0	0	0
710	Kenya Accorn Health Cente	01-Jan-11			0				
820	Manyatta Dispnsary	01-Jan-11	11	5	16	0	11	0	1
808	Marindi Dispensary	01-Jan-11			0				
949	Mirogi MCH	01-Jan-11	36	28	64		27	3	3
727	Ndiru Health Center	01-Jan-11			0				
804	Nyagoro HC	01-Jan-11	32	56	88	0	48	18	15
690	Ogande Dispensary	01-Jan-11	22	17	39	0	14	6	4
725	Pala Health Center	01-Jan-11			0				
686	Rangwe Sub-District Hospit	01-Jan-11			0				
640	St Paul Mission Dispensary	01-Jan-11	5	4	9	0	3	1	1
1015	Kandiege Health Center	01-Jan-11			0				
656	Kendu Adventist	01-Jan-11	60	61	121	1	24	9	17
842	Kendu Sub District Hospita	01-Jan-11			0				
868	Mawego Mission Hosp	01-Jan-11	5	7	12		5	6	1
712	Miriu Health Center	01-Jan-11	25	44	69		21	22	5
834	Oriang Dispensary	01-Jan-11	15	25	40		15	6	1
706	Atemo Maternity & Nursin	01-Jan-11	20	26	46		25	9	4
762	Godber Health Centre	01-Jan-11			0				
761	Kauma Dispensary	01-Jan-11			0				
600	Matata Nursing Home	01-Jan-11	29	49	78		50	11	18
840	Ober Health Centre	01-Jan-11			0				
931	Metaburo Mission	01-Jan-11	14	8	22		12	7	6
874	Nyamagwa Mission Hosp	01-Jan-11	7	18	25				6
890	Riokindo SDA Disp	01-Jan-11	21	29	50		24	24	5

Why is data quality important?

Outline

- Orientation to PATH
- Introduction to PATH's Data Quality Framework
- Case example: Malaria data in Zambia
- Application to your project

PATH is a leader in global health innovation

We harness our entrepreneurial insight, scientific and public health expertise, and passion for health equity…

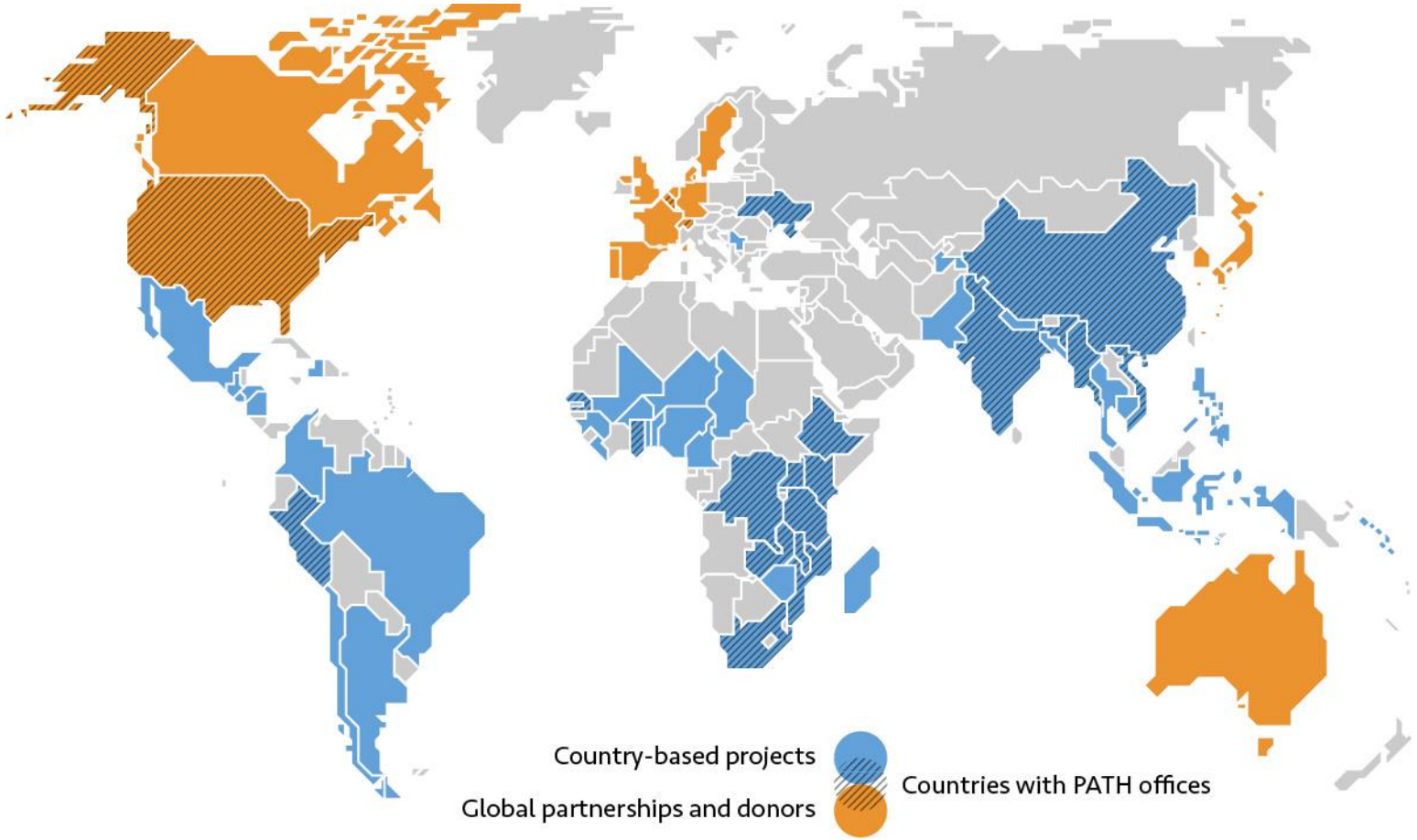
…to save the lives of women and children.

PATH/Gabe Bienczycki

PATH

Our global impact

Work in more than 70 countries
150 million people reached each year (average)



6 billion vaccine vial
monitors ensuring that vaccines are potent when given

6.3 million people
reached with rice fortified with critical micronutrients

6.2 million lives saved
with PATH-pioneered approaches to malaria control

System and service innovations

Improving care for women, children, and communities

PATH/Gabe Bienczycki

PATH

We need high quality data to inform a

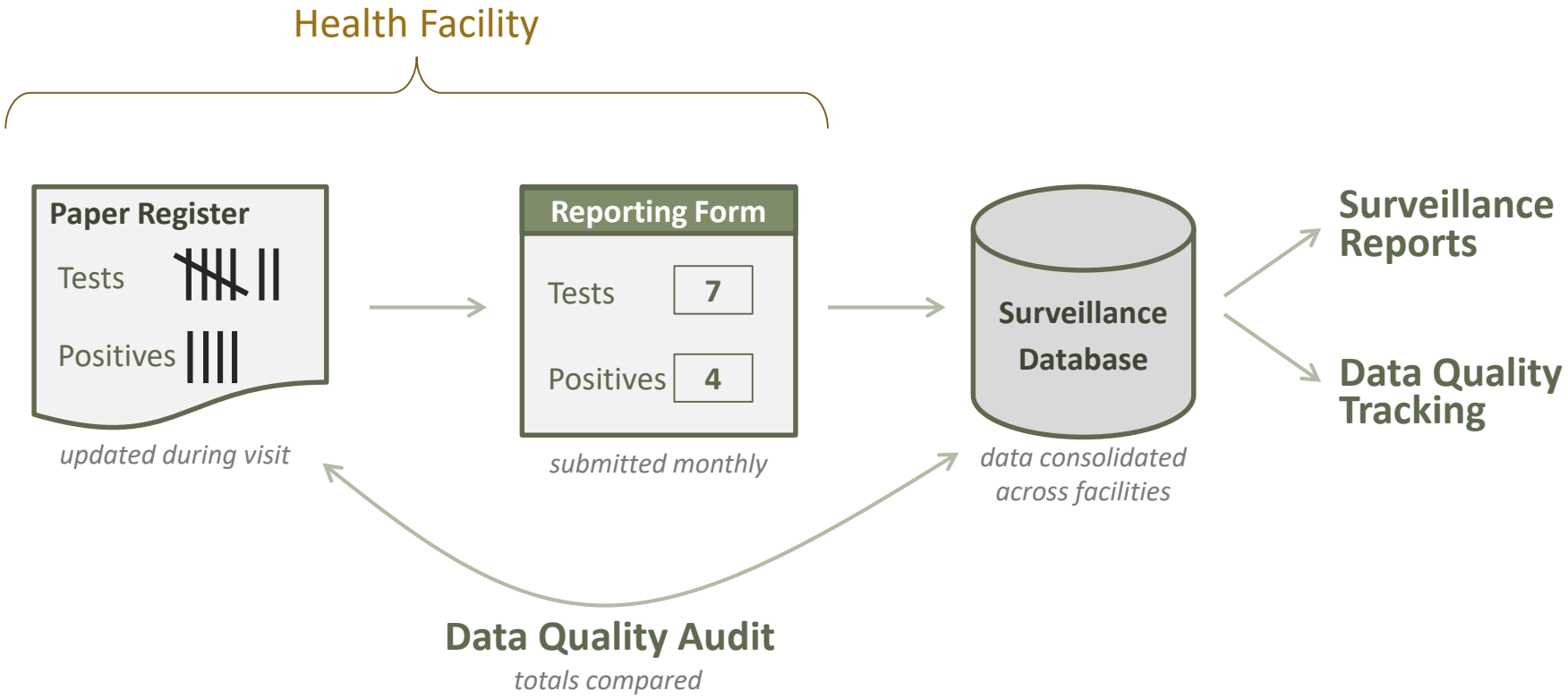
“data-driven culture”

Framework development

- Principles:
  - Marry conceptual with practical
  - Comprehensive assessment of data quality
  - Broadly applicable across projects and data sources
  - Customizable to specific project needs
  - Specific enough to identify areas of low quality

- Assumptions:
  - Have an existing dataset on hand
  - Have complete meta-data to describe that dataset
  - Not an assessment the data flow / data management process

Data Flows

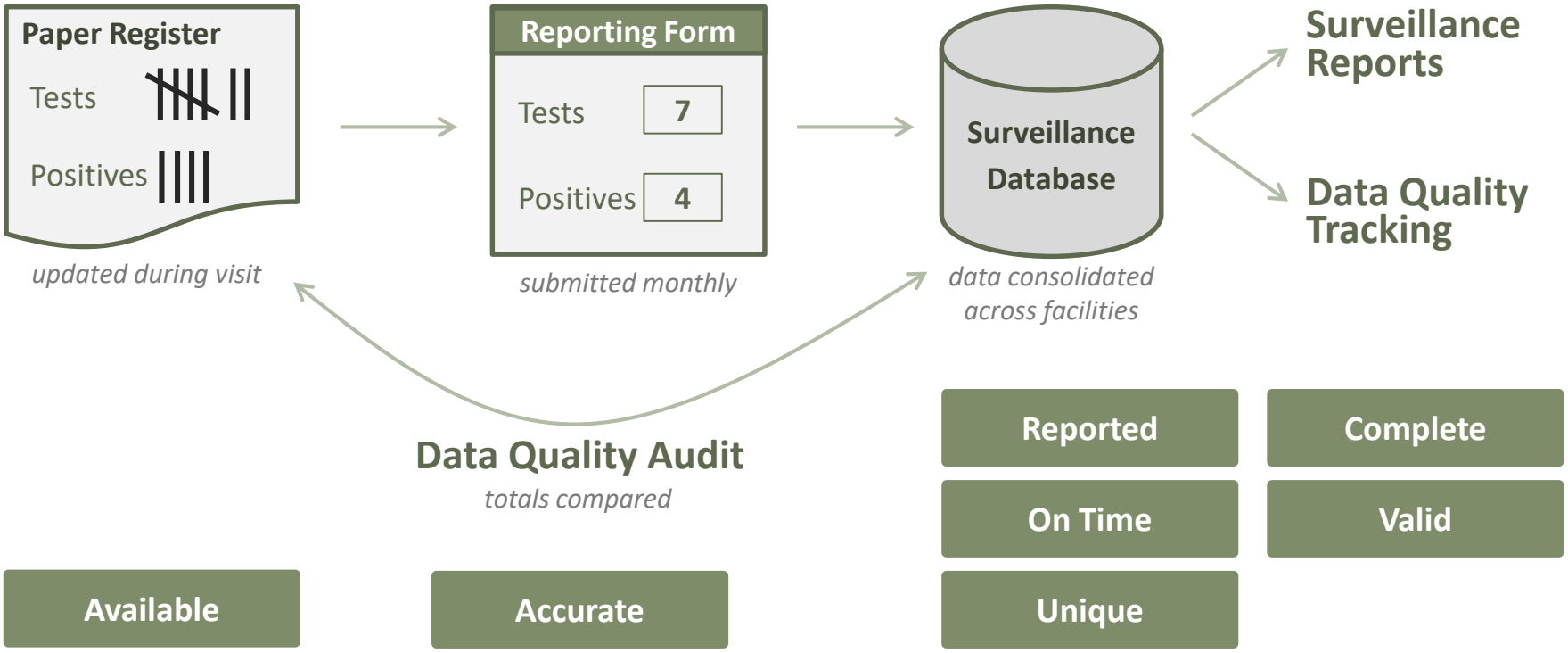

Health Facility
Paper Register
Tests
Positives
updated during visit
Reporting Form
Tests
7
Positives
4
submitted monthly
Surveillance Database
data consolidated across facilities
Surveillance Reports
Data Quality Tracking
Data Quality Audit
totals compared

Basic Data Quality Measures

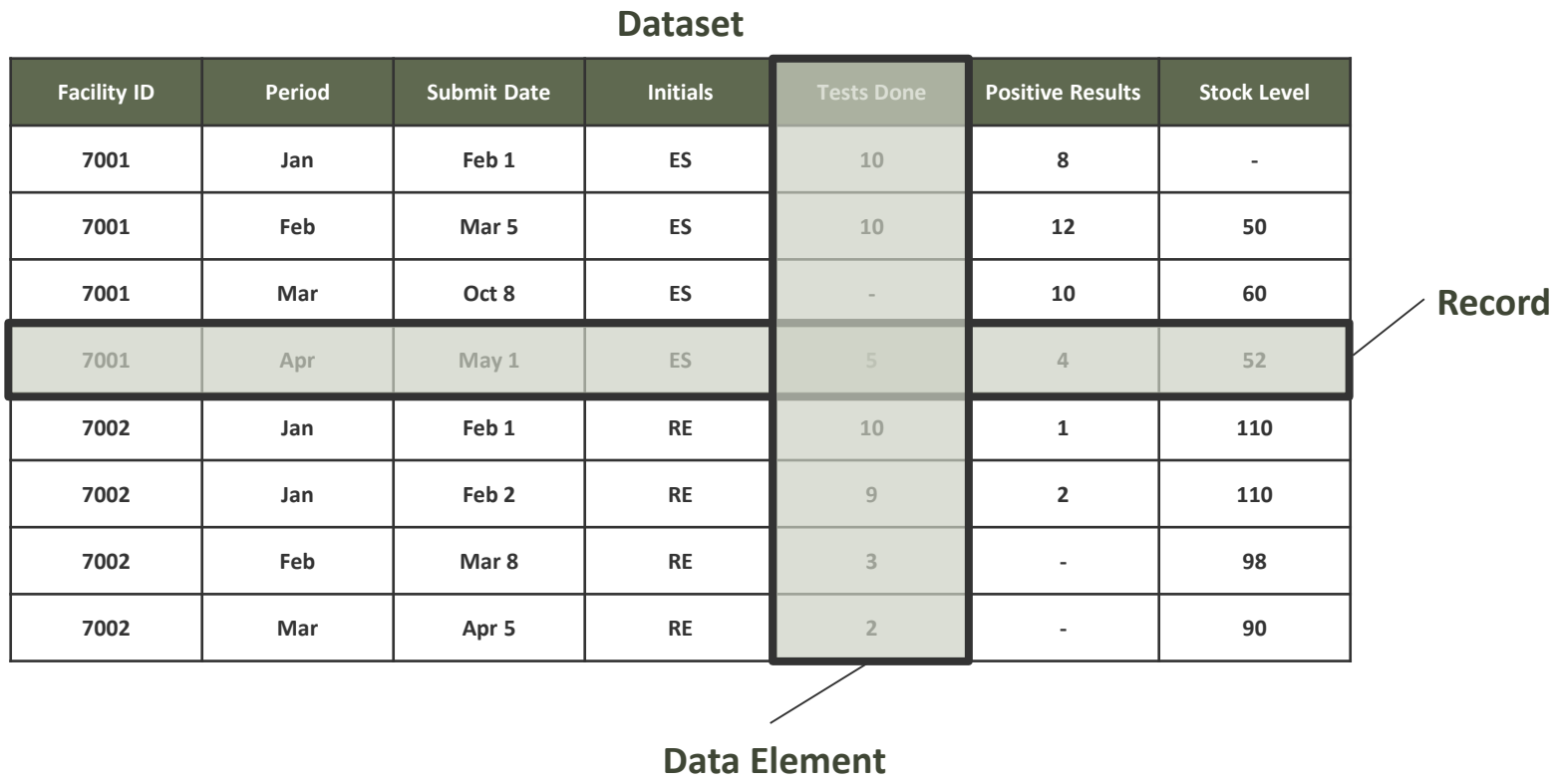
Reported	Was an expected report submitted or record entered?
On Time	Was data reported within the allotted time?
Unique	Is the data non-duplicate?
Complete	Is required data filled in?
Valid	Does the data comply with rules?
Available	Is data from a trusted source available?
Accurate	Does the data match reality or a trusted source?

*It may not be practical or necessary to measure them all. Pick the fitting ones.

Basic Data Quality Measures



Terminology



Record perspective vs. Data Element perspective

Facility ID	Period	Submit Date	Initials	Tests Done	Positive Results	Stock Level
7001	Jan	Feb 1	ES	10	8	-
7001	Feb	Mar 5	ES	10	12	50
7001	Mar	Oct 8	ES	-	10	60
7001	Apr	May 1	ES	5	4	52
7002	Jan	Feb 1	RE	10	1	110
7002	Jan	Feb 2	RE	9	2	110
7002	Feb	Mar 8	RE	3	-	98
7002	Mar	Apr 5	RE	2	-	90

Record perspective

5/8 records have all required data elements filled in

Data element perspective

7/8 values are filled in for data element Tests Done

Application of Data Quality Measures

Reported	Record perspective
On Time	Record perspective
Unique	Record perspective
Complete	Record or data element perspective
Valid	Record or data element perspective
Available	Record or data element perspective
Accurate	Record or data element perspective

Usage Examples for Each Measure

Reported

- Health facilities are expected to submit a surveillance report each week. Was there a report submitted by health facility A for week 10 of 2016?

On Time

- Health facilities are expected to submit a surveillance report within 14 days from the end of the week. Was a report received from health facility A within that allotted timeframe for the 2016 week 10 report?

Usage Examples for Each Measure

Unique

- Is a particular record in the dataset non-duplicated?

Complete

- **For a data element:** In a certain dataset, we expect Malaria Case Count to always be filled in. For a particular record, was that data element filled in?
- **For a record:** In a certain dataset, we expect 5 data elements to always be filled in. For a particular record, were all of these data elements filled in?

Valid

- **For a data element:** For a particular record, is Malaria Case Count a positive integer and less than or equal to the Malaria Test Count.
- **For a record:** For all data elements in a particular record, are the following conditions met? Values fit data element definitions; correct format, data type, and precision; validation rules are adhered to.

Usage Examples for Each Measure

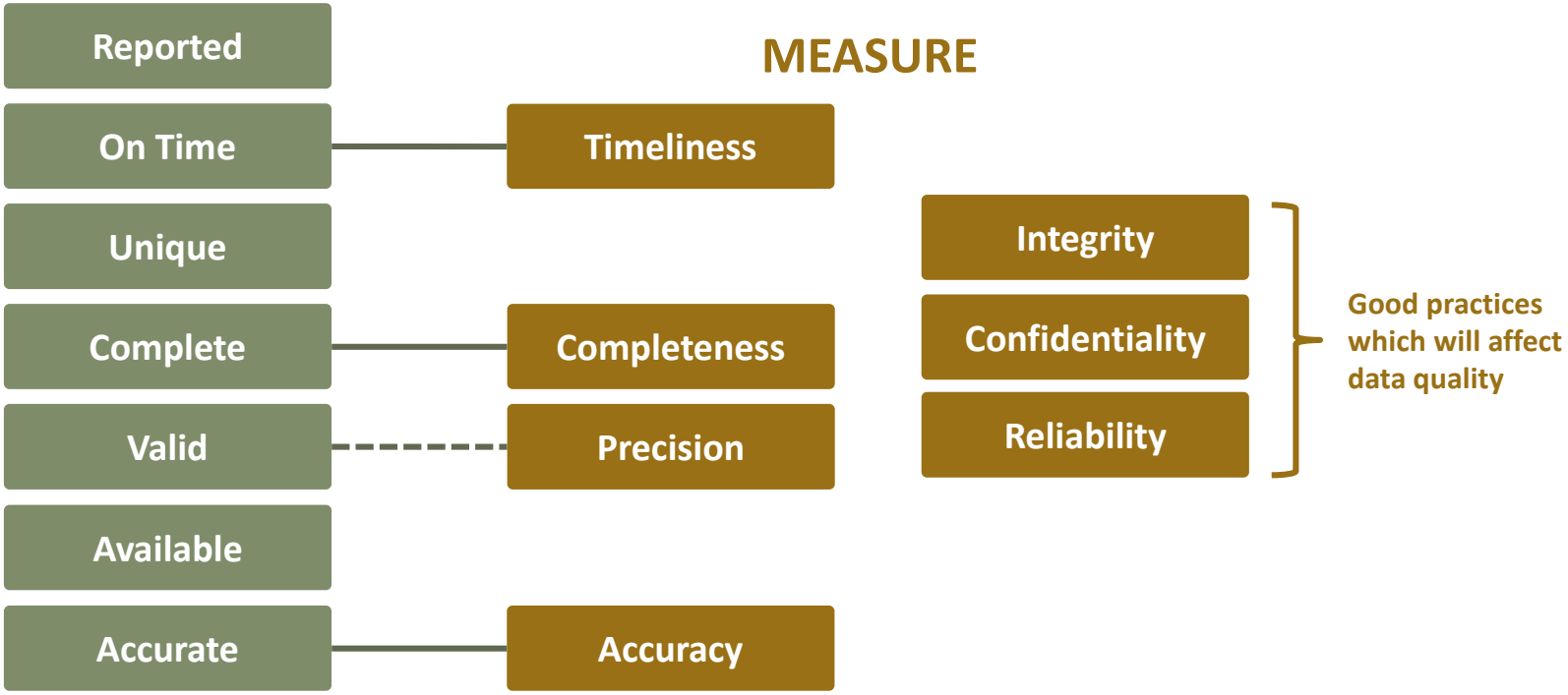
Available

- Paper registers at a health facility (considered a trusted source) are to be used to check the accuracy of reported values. Are the paper registers available to check Malaria Case Count for week 10 of 2016?

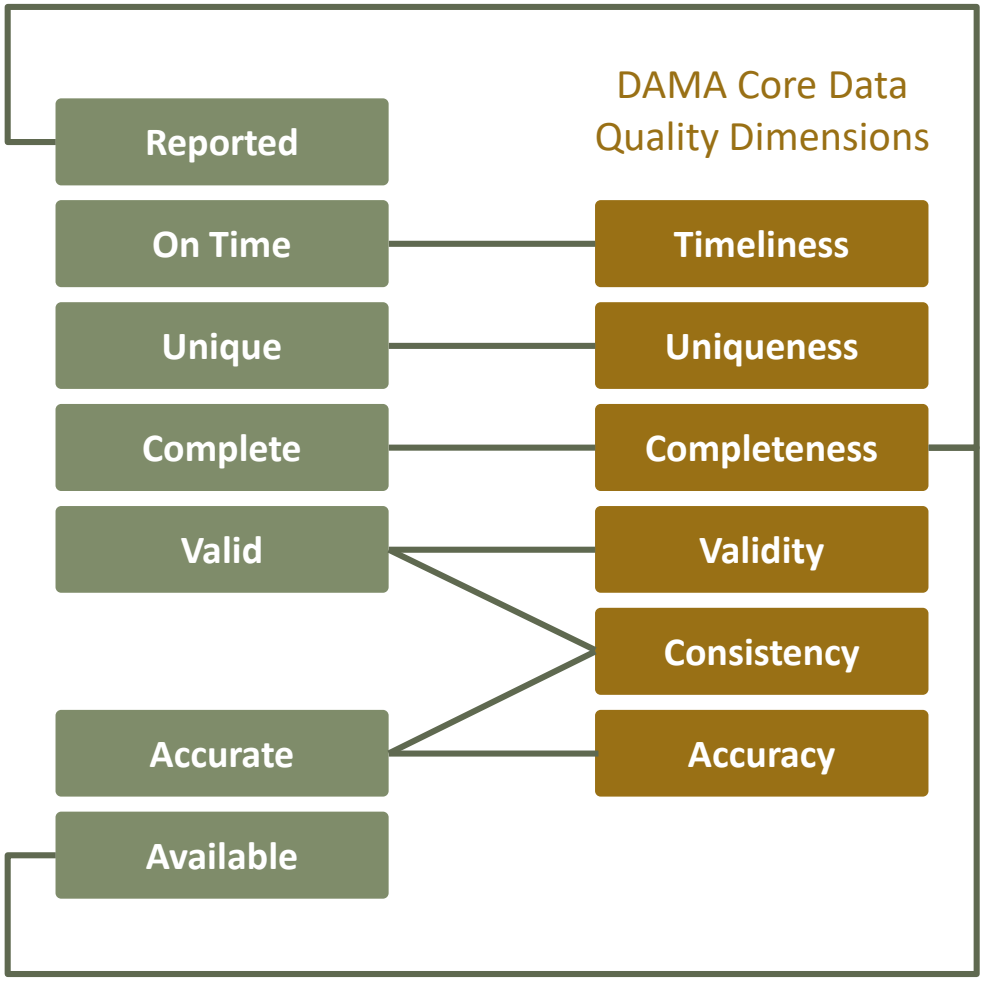
Accurate

- **For a data element:** Does the value of Malaria Case Count reported for 2016 week 10 match the paper registers (a trusted source)?
- **For a record:** Do all the reported indicators for 2016 week 10 match the paper registers?

Comparing Data Quality Frameworks - MEASURE



Comparing Data Quality Frameworks - DAMA

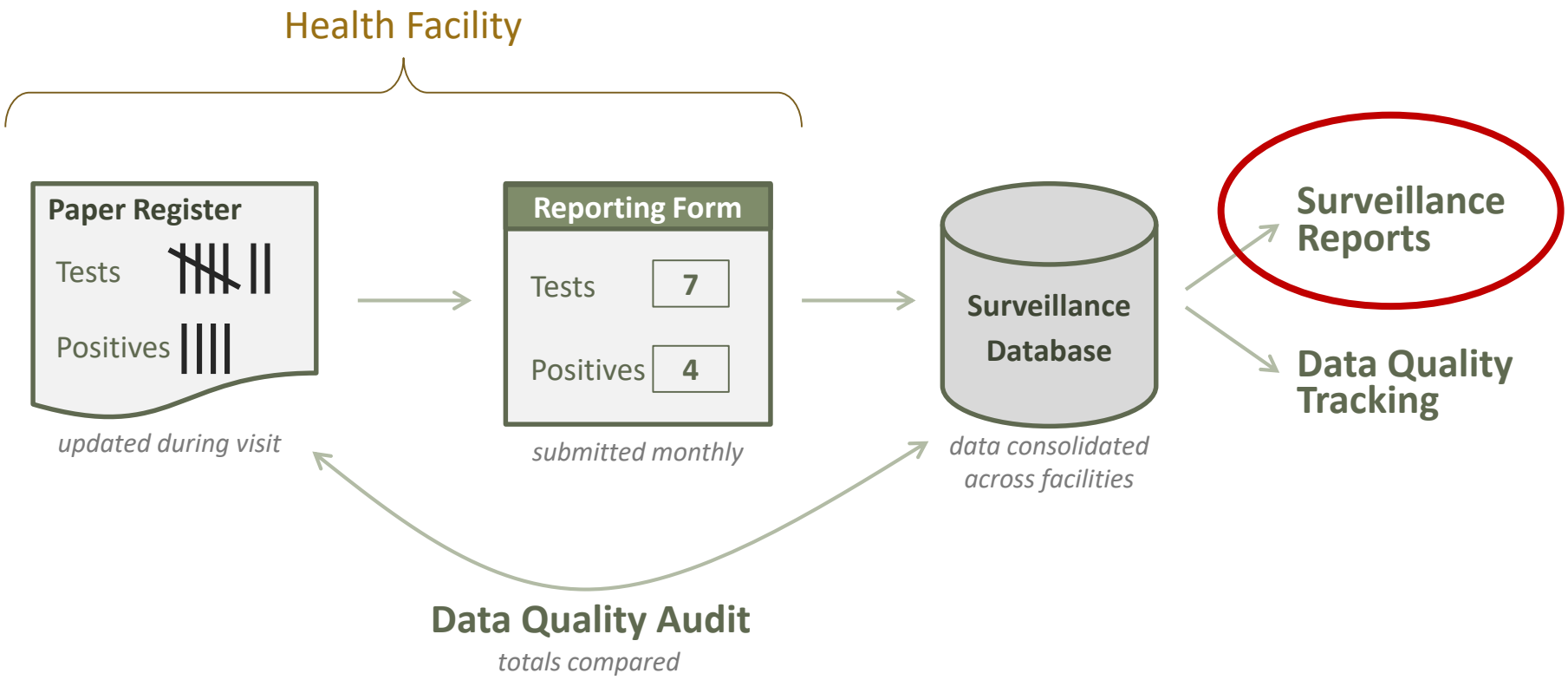


Ending Malaria in Zambia

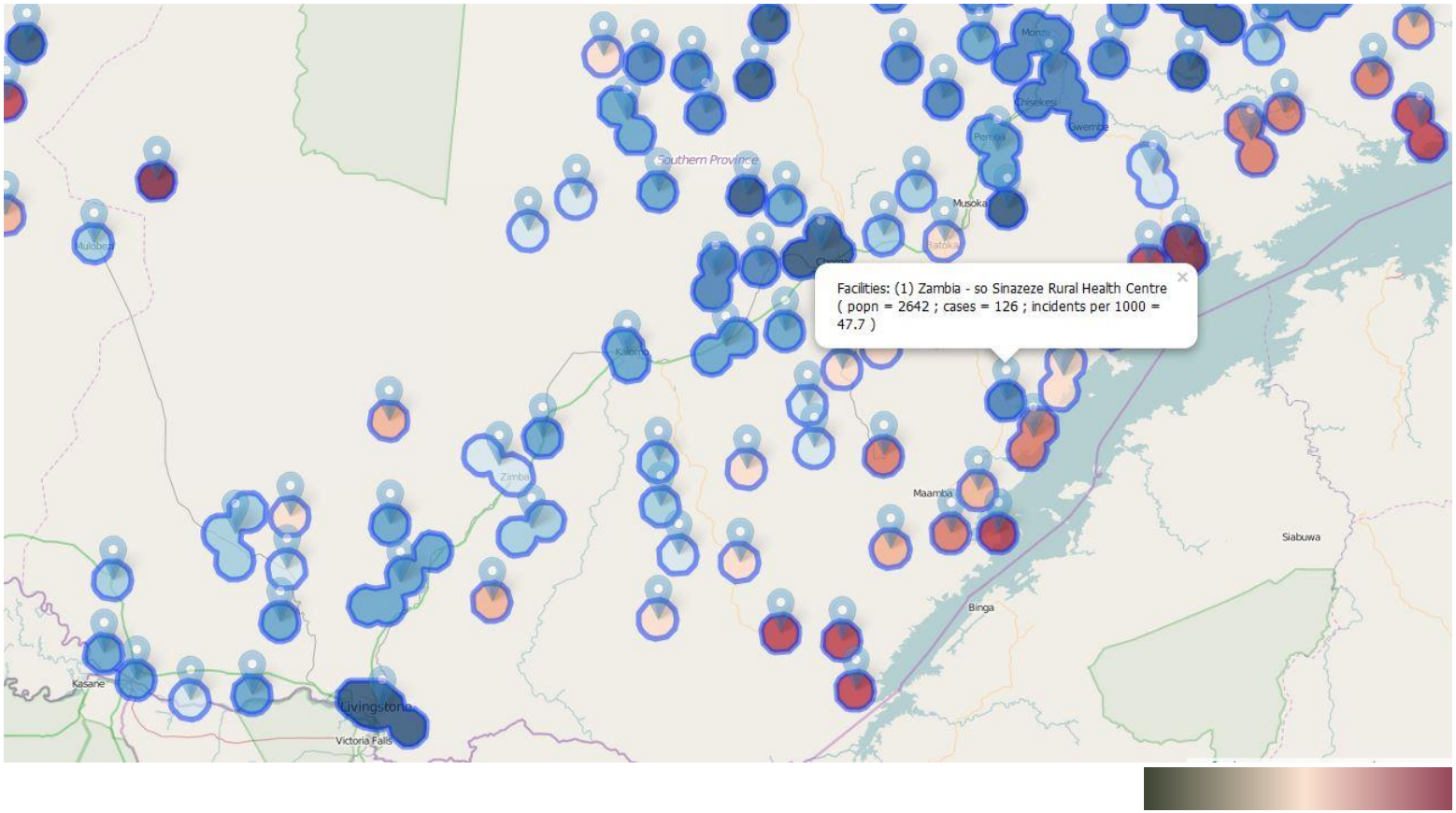
PATH/Gabe Bienczycki

PATH

Data Flows

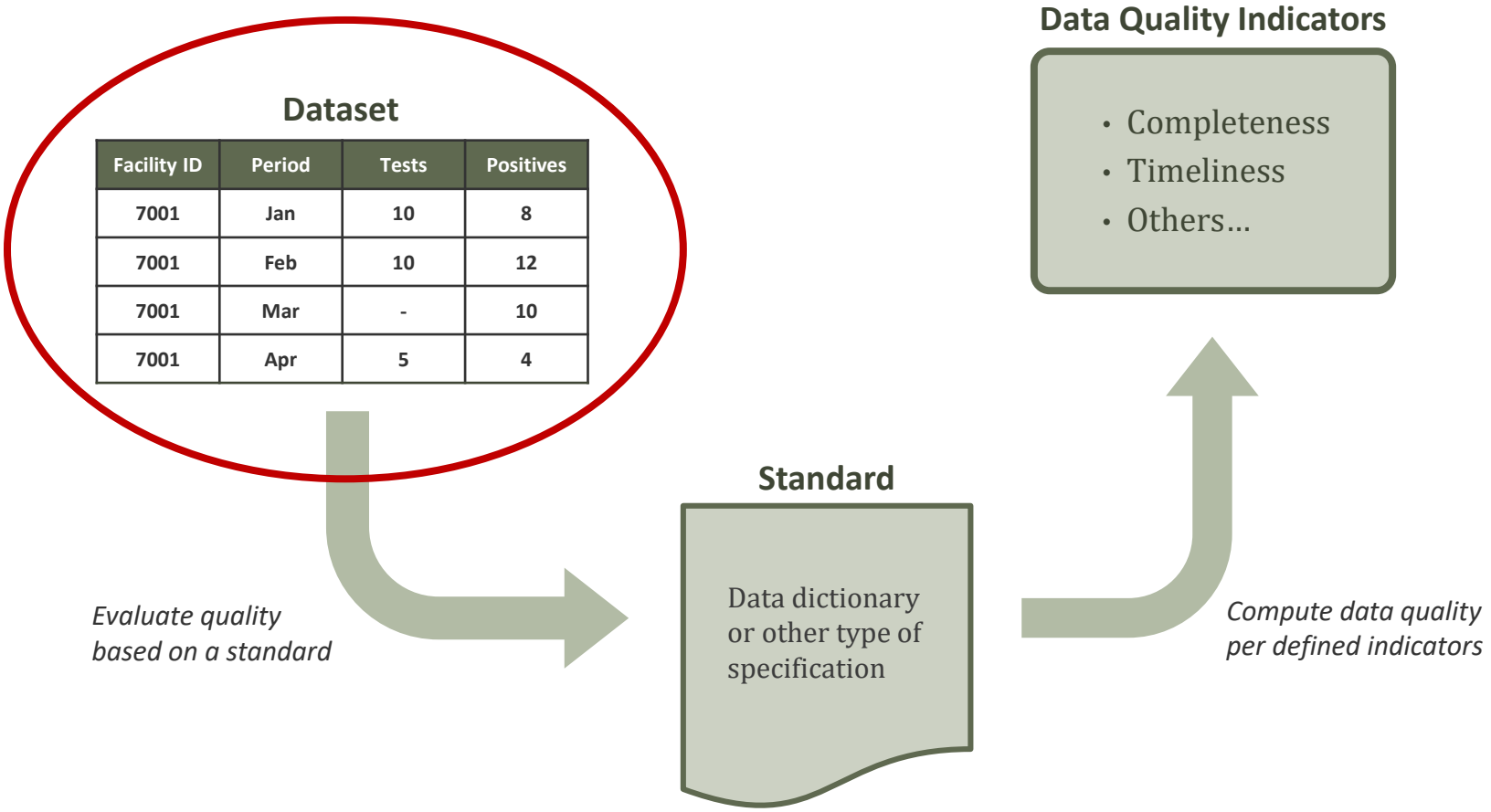

Health Facility
Paper Register
Tests
Positives
updated during visit
Reporting Form
Tests
7
Positives
4
submitted monthly
Surveillance Database
data consolidated across facilities
Surveillance Reports
Data Quality Tracking
Data Quality Audit
totals compared

Data can help us target resources



Health facilities and 5km radii, colored by incidence (cases per 1000 person-years); combined polygons

Data Quality Approach



Step 1: Define the dataset to be evaluated

Monthly Reporting of Malaria Surveillance Indicators

Facility ID	Period	Submit Date	Initials	Tests Done	Positive Results	Stock Level
7001	Jan	Feb 1	ES	10	8	-
7001	Feb	Mar 5	ES	10	12	50
7001	Mar	Oct 8	ES	-	10	60
7001	Apr	May 1	ES	5	4	52
7002	Jan	Feb 1	RE	10	1	110
7002	Jan	Feb 2	RE	9	2	110
7002	Feb	Mar 8	RE	3	-	98
7002	Mar	Apr 5	RE	2	-	90

Step 2: Determine the scope of data to include

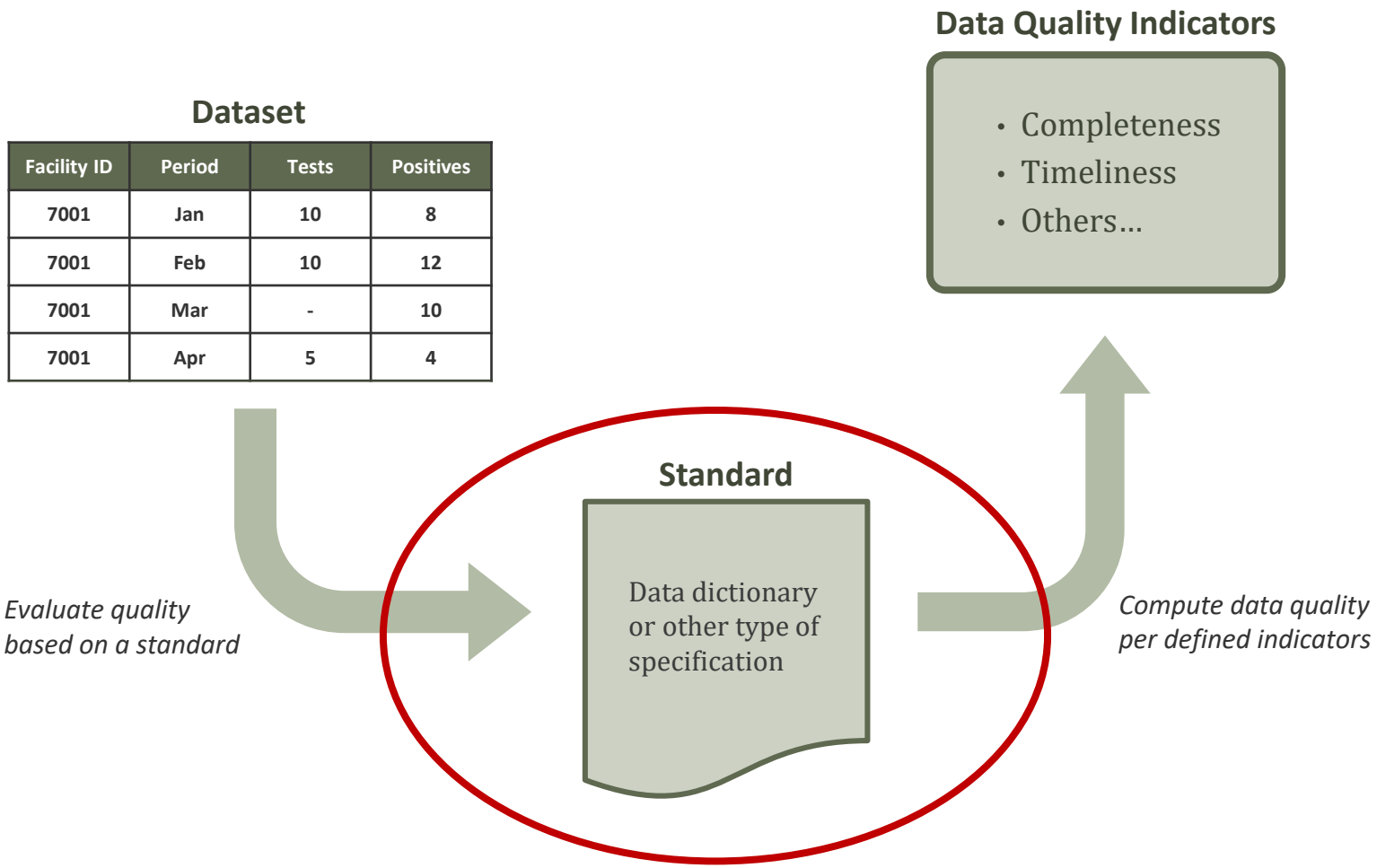
not of interest

concerned only with Jan-Mar

Facility ID	Period	Submit Date	Initials	Tests Done	Positive Results	Stock Level
7001	Jan	Feb 1	ES	10	8	-
7001	Feb	Mar 5	ES	10	12	50
7001	Mar	Oct 8	ES	-	10	60
7001	Apr	May 1	ES	5	4	52
7002	Jan	Feb 1	RE	10	1	110
7002	Jan	Feb 2	RE	9	2	110
7002	Feb	Mar 8	RE	3	-	98
7002	Mar	Apr 5	RE	2	-	90

- Data elements: Facility ID, Period, Submit Date, Tests Done, Positive Results
- Time period: January – March 2016
- Facilities: 7001 – 7003

Data Quality Approach



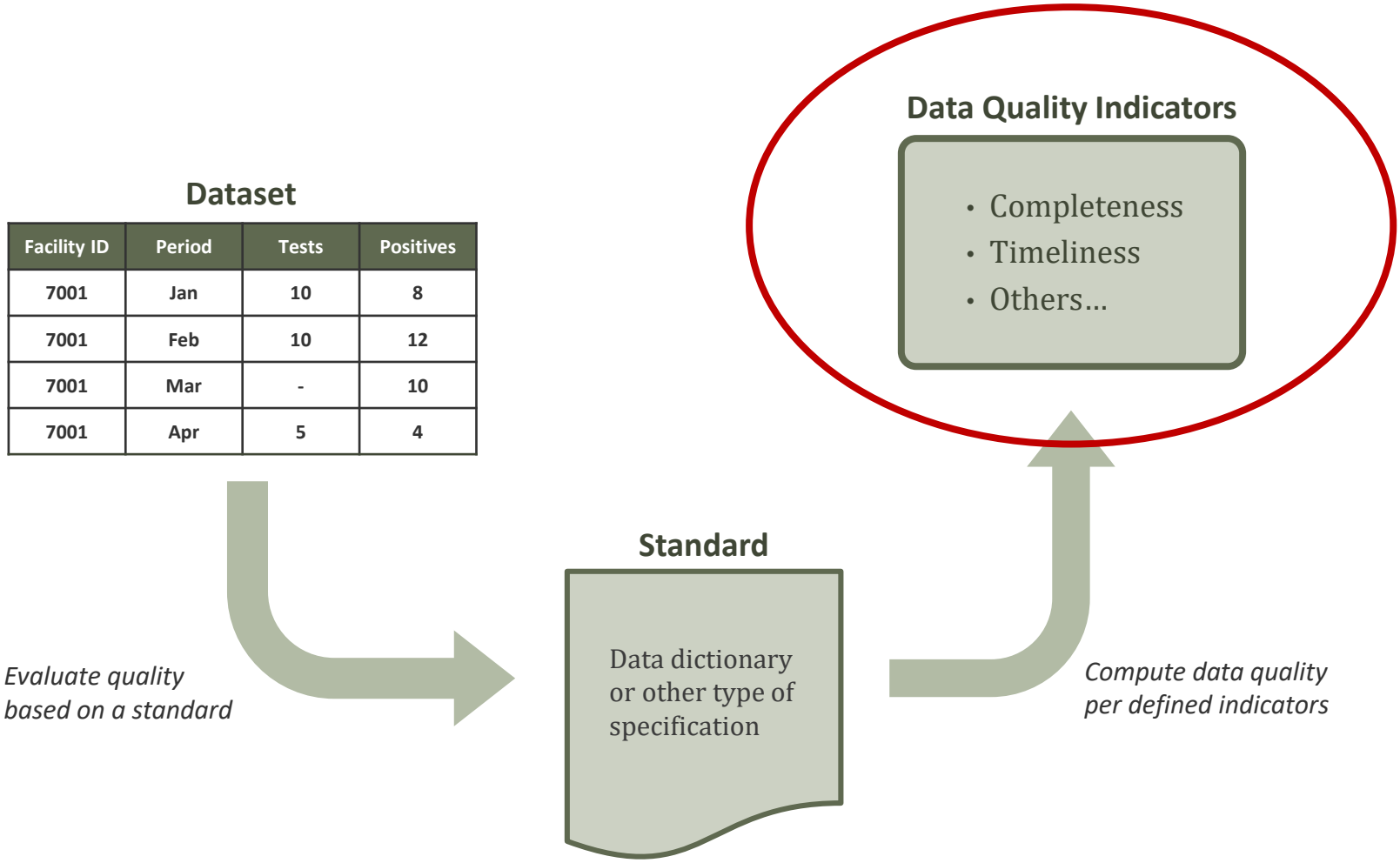
Step 3: Define standards for this context

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level
7001	Jan	Feb 1	10	8	-
7001	Feb	Mar 5	10	12	50
7001	Mar	Oct 8	-	10	60
7002	Jan	Feb 1	10	1	110
7002	Jan	Feb 2	9	2	110
7002	Feb	Mar 8	3	-	98
7002	Mar	Apr 5	2	-	90
7003	*Jan*				
7003	*Feb*				
7003	*Mar*				
Required Integer	Required Date	Required Date	Required Integer	Required Integer	Optional Integer

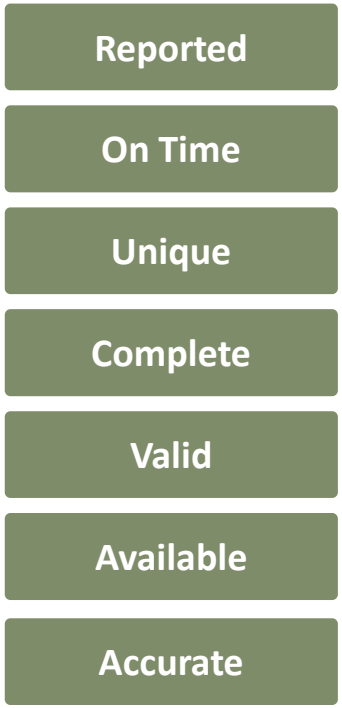
Unique Key

Validation Rule: Positive Results <= Tests Done

Data Quality Approach



Step 4: Describe each data quality measure for this context



- Which data quality measures are you most interested in for this context?
- Which are most important?
- Which may not be relevant?

Step 4: Describe each data quality measure for this context

Reported	The facility submitted a report for the month.
On Time	The report Submitted Date is within 10 days from the end of the month.
Unique	A report is not a duplicate, based on the unique key of Facility ID and Period.
Complete	All required data elements are filled in for the report.
Valid	Positive cases is less than or equal to tests done.
Available	*Not applicable*
Accurate	*Not applicable*

Step 4: Describe each data quality measure for this context

Reported	The facility submitted a report for the month.
On Time	The report Submitted Date is within 10 days from the end of the month.
Unique	A report is not a duplicate, based on the unique key of Facility ID and Period.
Complete	All required data elements are filled in for the report.
Valid	Positive cases is less than or equal to tests done.
Available	*Not applicable*
Accurate	*Not applicable*

Step 5: Measure data quality for each record

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●				
7001	Feb	Mar 5	10	12	50	●				
7001	Mar	Oct 8	-	10	60	●				
7002	Jan	Feb 1	10	1	110	●				
7002	Jan	Feb 2	9	2	110	●				
7002	Feb	Mar 8	3	-	98	●				
7002	Mar	Apr 5	2	-	90	●				
7003	*Jan*					○				
7003	*Feb*					○				
7003	*Mar*					○				
Data Quality Indicators										

Step 6: Compute data quality indicators

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●				
7001	Feb	Mar 5	10	12	50	●				
7001	Mar	Oct 8	-	10	60	●				
7002	Jan	Feb 1	10	1	110	●				
7002	Jan	Feb 2	9	2	110	●				
7002	Feb	Mar 8	3	-	98	●				
7002	Mar	Apr 5	2	-	90	●				
7003	*Jan*					○				
7003	*Feb*					○				
7003	*Mar*					○				
					Data Quality Indicators	7/10				

Step 4: Describe each data quality measure for this context

Reported	The facility submitted a report for the month.
On Time	The report Submitted Date is within 10 days from the end of the month.
Unique	A report is not a duplicate, based on the unique key of Facility ID and Period.
Complete	All required data elements are filled in for the report.
Valid	Positive cases is less than or equal to tests done.
Available	*Not applicable*
Accurate	*Not applicable*

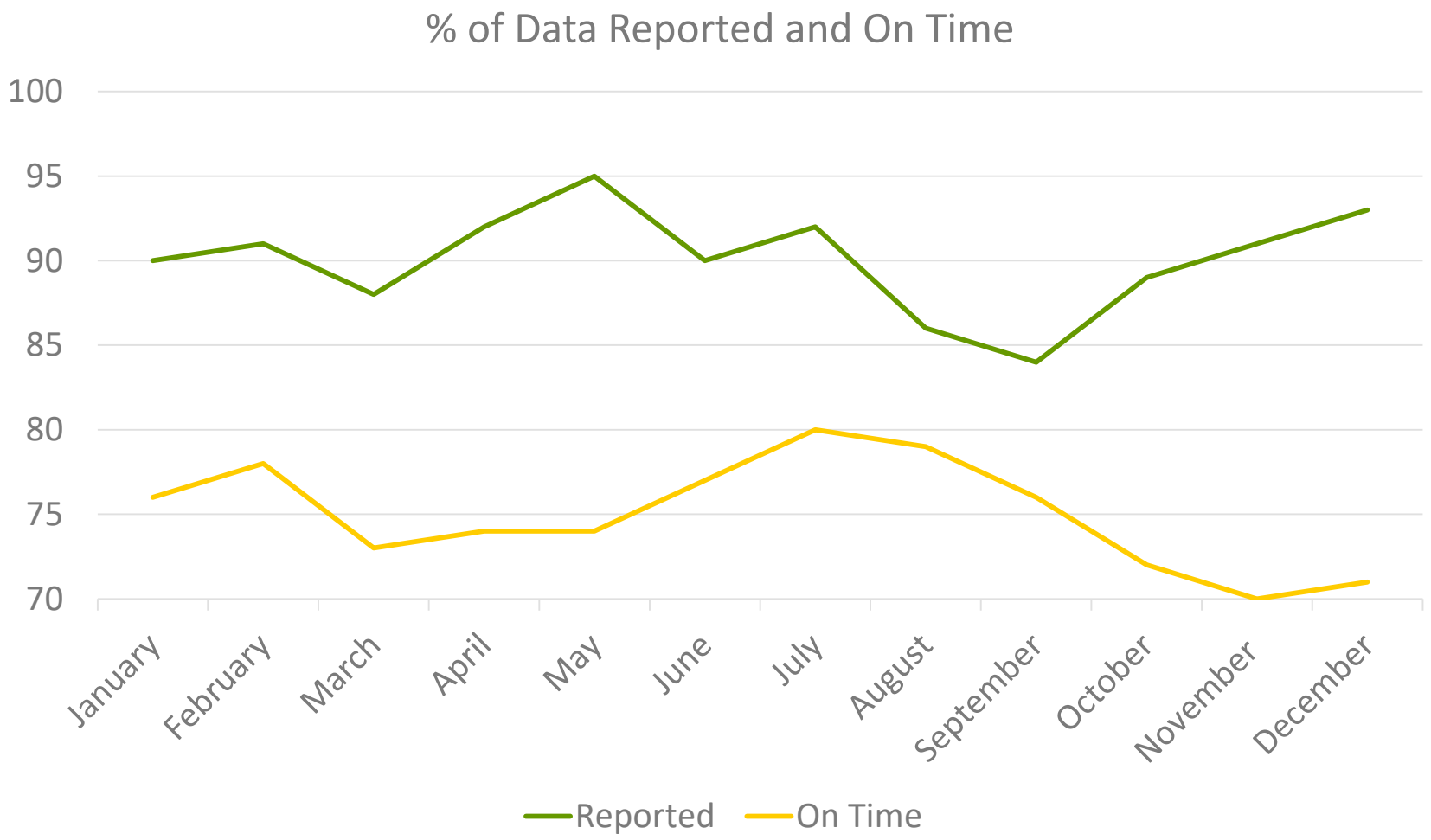
Step 5: Measure data quality for each record

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●	●			
7001	Feb	Mar 5	10	12	50	●	●			
7001	Mar	Oct 8	-	10	60	●	○			
7002	Jan	Feb 1	10	1	110	●	●			
7002	Jan	Feb 2	9	2	110	●	●			
7002	Feb	Mar 8	3	-	98	●	●			
7002	Mar	Apr 5	2	-	90	●	●			
7003	*Jan*					○	○			
7003	*Feb*					○	○			
7003	*Mar*					○	○			
					Data Quality Indicators	**7/10**				

Step 6: Compute data quality indicators

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●	●			
7001	Feb	Mar 5	10	12	50	●	●			
7001	Mar	Oct 8	-	10	60	●	○			
7002	Jan	Feb 1	10	1	110	●	●			
7002	Jan	Feb 2	9	2	110	●	●			
7002	Feb	Mar 8	3	-	98	●	●			
7002	Mar	Apr 5	2	-	90	●	●			
7003	*Jan*					○	○			
7003	*Feb*					○	○			
7003	*Mar*					○	○			
Data Quality Indicators						**7/10**	**6/10**			

Visualizing data quality



Step 4: Describe each data quality measure for this context

Reported	The facility submitted a report for the month.
On Time	The report Submitted Date is within 10 days from the end of the month.
Unique	A report is not a duplicate, based on the unique key of Facility ID and Period.
Complete	All required data elements are filled in for the report.
Valid	Positive cases is less than or equal to tests done.
Available	*Not applicable*
Accurate	*Not applicable*

Step 5: Measure data quality for each record

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●	●	●		
7001	Feb	Mar 5	10	12	50	●	●	●		
7001	Mar	Oct 8	-	10	60	●	○	○		
7002	Jan	Feb 1	10	1	110	●	●	●		
7002	Jan	Feb 2	9	2	110	●	●	●		
7002	Feb	Mar 8	3	-	98	●	●	○		
7002	Mar	Apr 5	2	-	90	●	●	○		
7003	*Jan*					○	○	-		
7003	*Feb*					○	○	-		
7003	*Mar*					○	○	-		
					Data Quality Indicators	7/10	6/10			

PATH

Step 6: Compute data quality indicators

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●	●	●		
7001	Feb	Mar 5	10	12	50	●	●	●		
7001	Mar	Oct 8	-	10	60	●	○	○		
7002	Jan	Feb 1	10	1	110	●	●	●		
7002	Jan	Feb 2	9	2	110	●	●	●		
7002	Feb	Mar 8	3	-	98	●	●	○		
7002	Mar	Apr 5	2	-	90	●	●	○		
7003	*Jan*					○	○	-		
7003	*Feb*					○	○	-		
7003	*Mar*					○	○	-		
					Data Quality Indicators	7/10	6/10	4/7		

Step 4: Describe each data quality measure for this context

Reported	The facility submitted a report for the month.
On Time	The report Submitted Date is within 10 days from the end of the month.
Unique	A report is not a duplicate, based on the unique key of Facility ID and Period.
Complete	All required data elements are filled in for the report.
Valid	Positive cases is less than or equal to tests done.
Available	*Not applicable*
Accurate	*Not applicable*

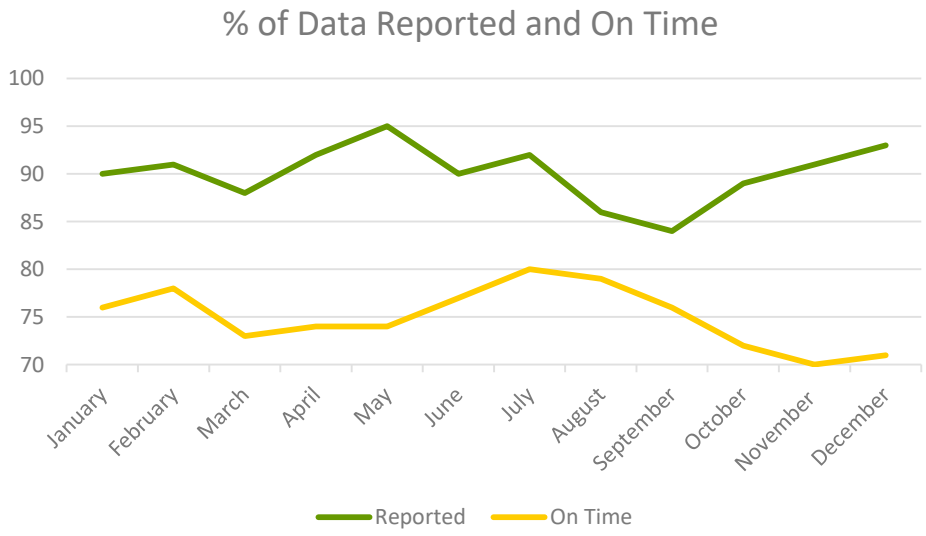
Step 5: Measure data quality for each record

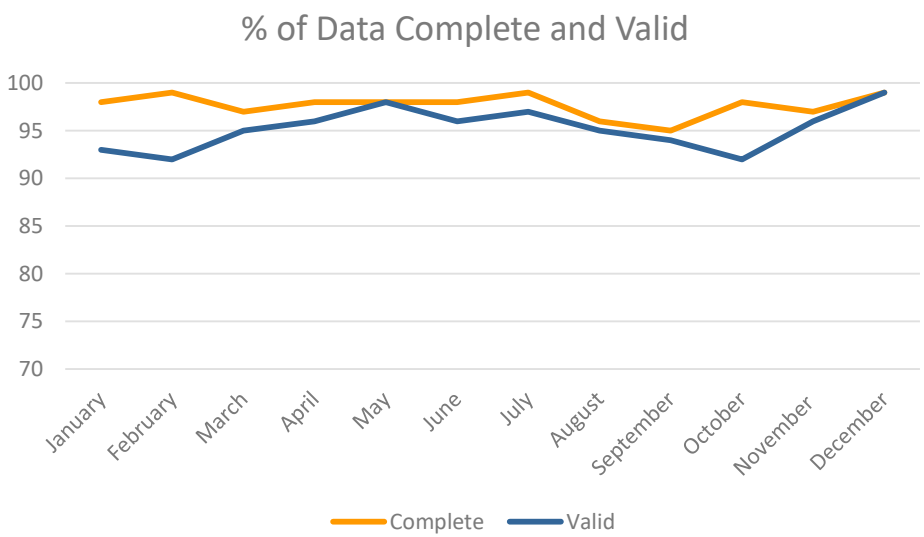
Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●	●	●	●	
7001	Feb	Mar 5	10	12	50	●	●	●	○	
7001	Mar	Oct 8	-	10	60	●	○	○	●	
7002	Jan	Feb 1	10	1	110	●	●	●	●	
7002	Jan	Feb 2	9	2	110	●	●	●	●	
7002	Feb	Mar 8	3	-	98	●	●	○	●	
7002	Mar	Apr 5	2	-	90	●	●	○	●	
7003	*Jan*					○	○	-	-	
7003	*Feb*					○	○	-	-	
7003	*Mar*					○	○	-	-	
					Data Quality Indicators	7/10	6/10	4/7		

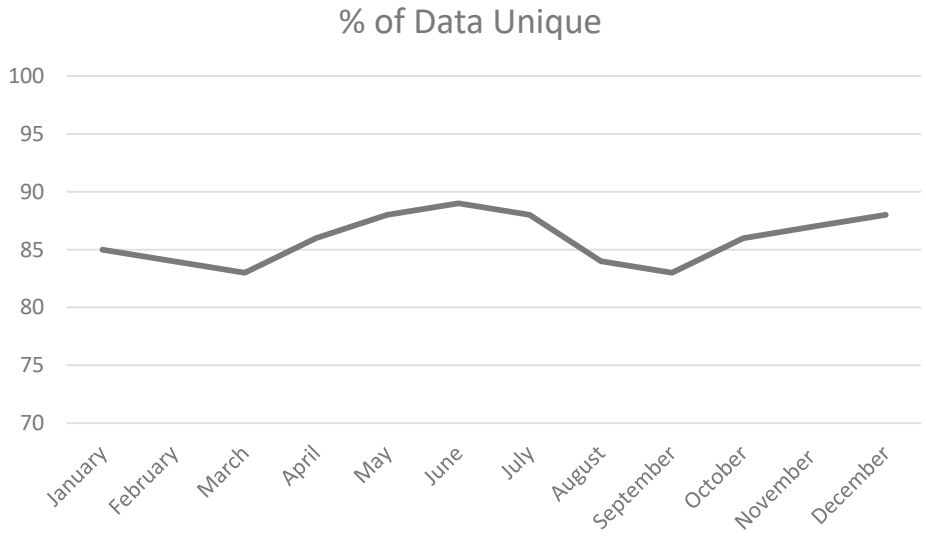
Step 6: Compute data quality indicators

Facility ID	Period	Submit Date	Tests Done	Positive Results	Stock Level	Reported	On Time	Complete	Valid	Unique
7001	Jan	Feb 1	10	8	-	●	●	●	●	
7001	Feb	Mar 5	10	12	50	●	●	●	○	
7001	Mar	Oct 8	-	10	60	●	○	○	●	
7002	Jan	Feb 1	10	1	110	●	●	●	●	
7002	Jan	Feb 2	9	2	110	●	●	●	●	
7002	Feb	Mar 8	3	-	98	●	●	○	●	
7002	Mar	Apr 5	2	-	90	●	●	○	●	
7003	*Jan*					○	○	-	-	
7003	*Feb*					○	○	-	-	
7003	*Mar*					○	○	-	-	
					Data Quality Indicators	7/10	6/10	4/7	6/7	

Visualizing Indicators for Data Quality Over Time







*In this example, the following rules hold:

- Reported and On Time are measured from the full expected sample
- Complete and Valid are measured from the full reported sample
- Unique is measured from the full reported, complete, and valid sample.

PATH

so Kazungula District Data Quality Metrics

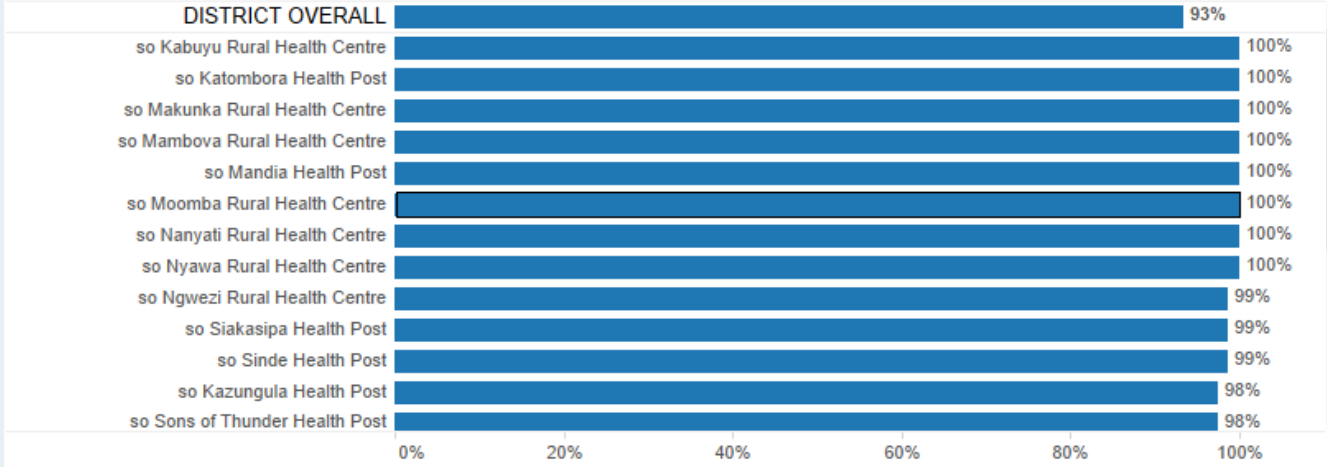
Province: so Southern Province
District: so Kazungula District
Date Range: 1/23/2015 – 8/9/2016
Primary Metric: % Reported
Facilities to Include: Has reported

Reported	Reported On Time
93%	64%

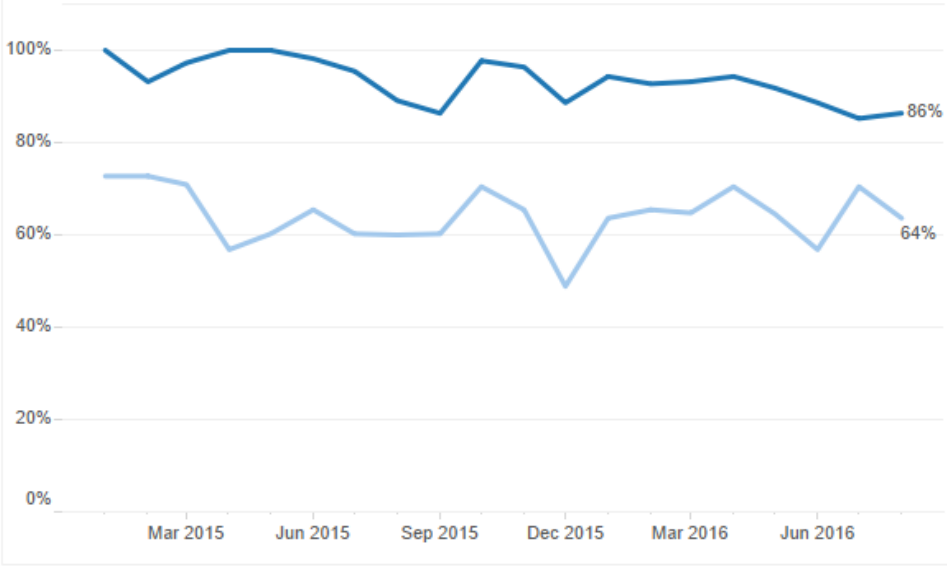
Complete	Valid
99%	100%

% Reported & On Time based on # of expected reports.
% Complete & Valid based on # of submitted reports.

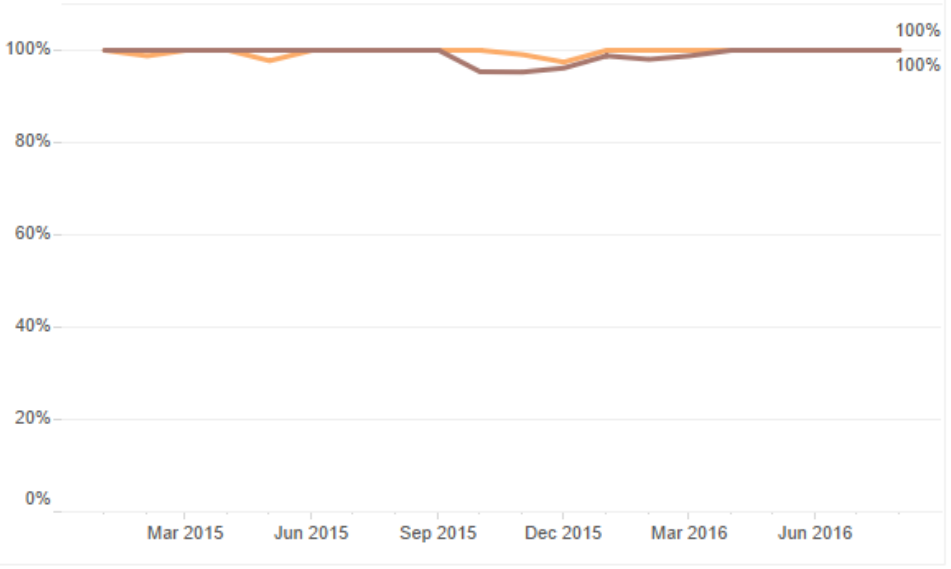
% Reported



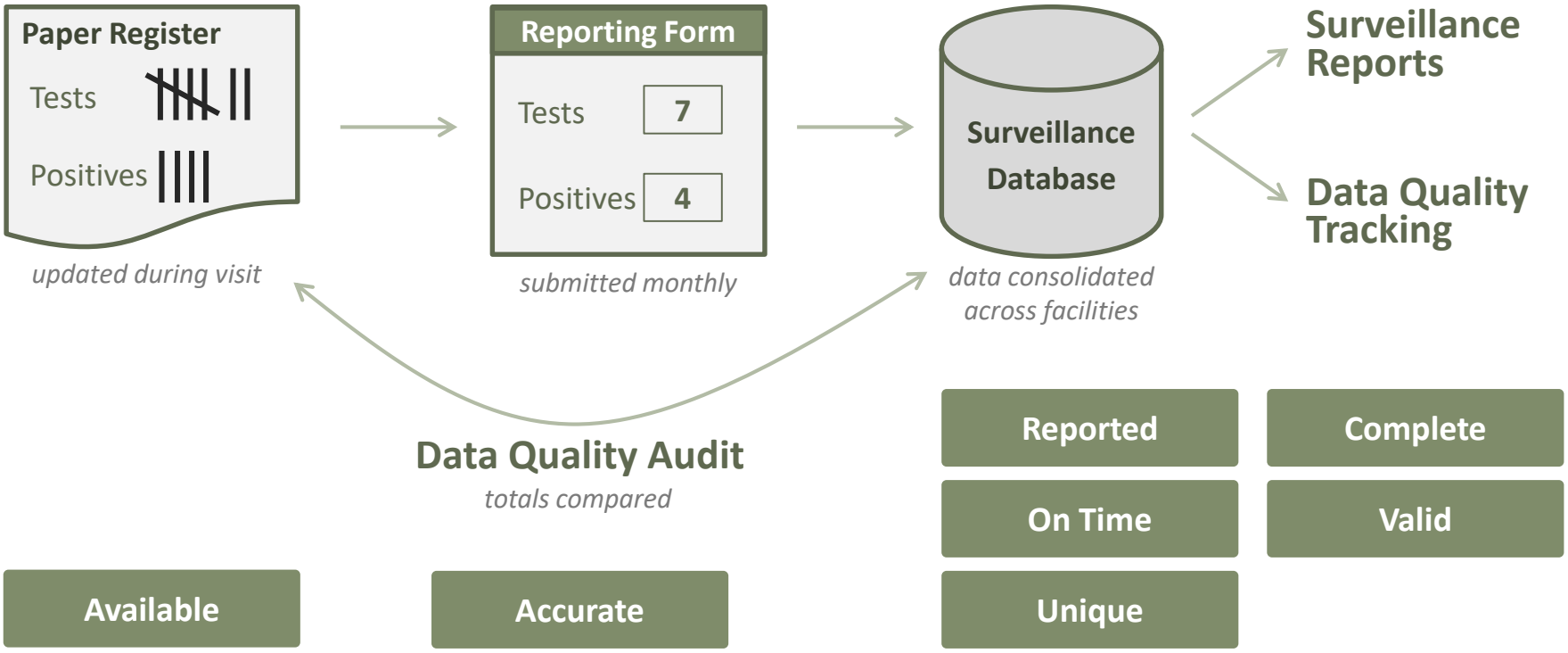
% Reported and % Reported On Time



% Complete and % Valid of Submitted Reports



Basic Data Quality Measures



Trusted Value vs. Reported Value

Facility ID	Period	Data Element	Trusted Value	Reported Value	✓
7004	Jan	Tests Done	4	4	✓
7004	Jan	Positive Results	3	2	
7004	Feb	Tests Done	15	-	
7004	Feb	Positive Results	-	14	
7004	Mar	Tests Done	5	5	✓
7004	Mar	Positive Results	7	7	✓

Measuring Accuracy - Option 1: Percent Error

Percent Error = (| reported value – trusted value | / trusted value) x 100

Trusted Value	Reported Value	Difference	Denominator	Percent Error
4	4	0	4	0%
4	3	1	4	25%
1	4	3	1	300%
0 **1** adjusted	4 **5** adjusted	4	1	400%

Measuring Accuracy - Option 2: Accuracy Score

Accuracy Score = 100 x Min/Max of trusted and reported values

Trusted Value	Reported Value	Min	Max	Accuracy Score
4	4	4	4	100
4	3	3	4	75
1	4	1	4	25
0	4	1 adjusted	5 adjusted	20

Accuracy score ranges from 0 to 100.

Percent Error vs. Accuracy Score

Percent Error

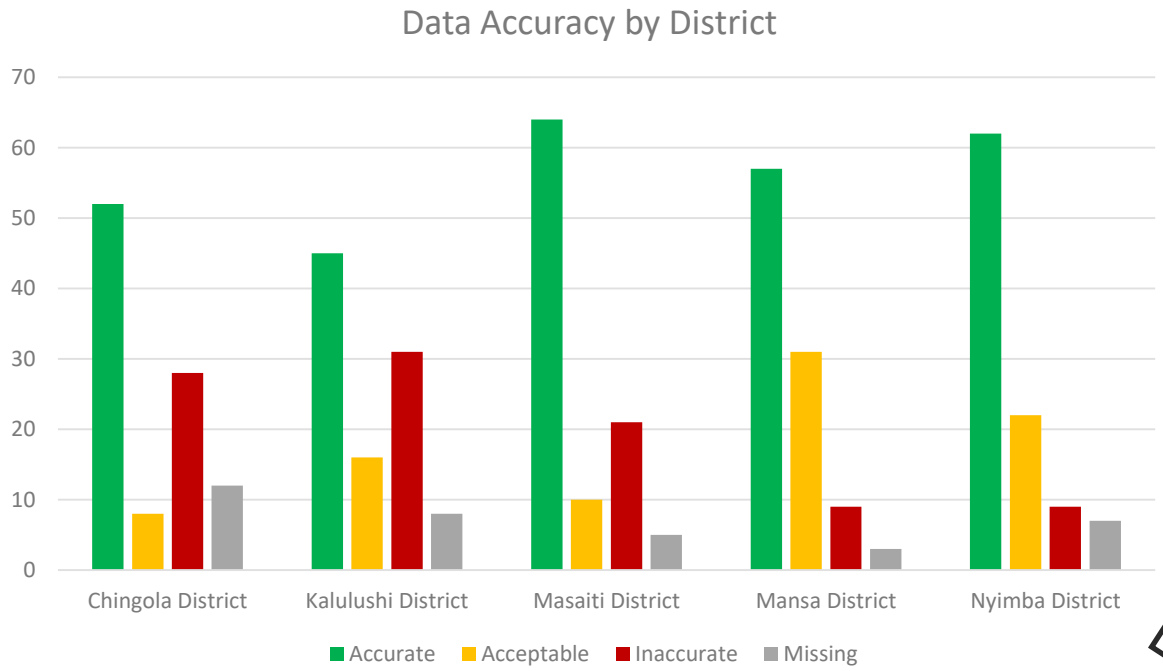
How far off is the reported value from the trusted value?

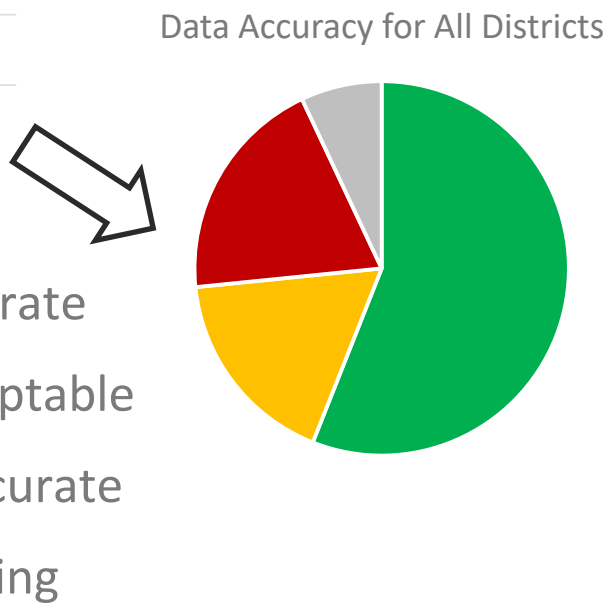
Accuracy Score

How close are the trusted and reported values?

Trusted Value	Reported Value	Percent Error	Accuracy Score
4	4	0%	100
4	3	25%	75
1	4	300%	25
4	1	75%	25
20	1	95%	5
1	20	1900%	5

Visualizing Indicators for Data Quality Audits





Application

- How can these measures apply to your project data?