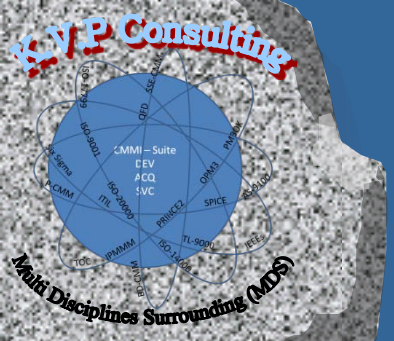
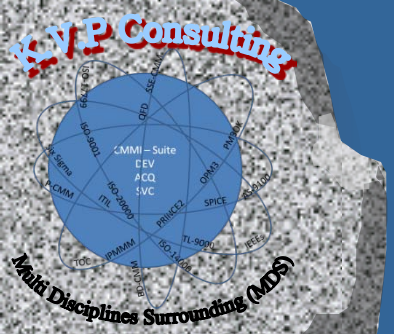


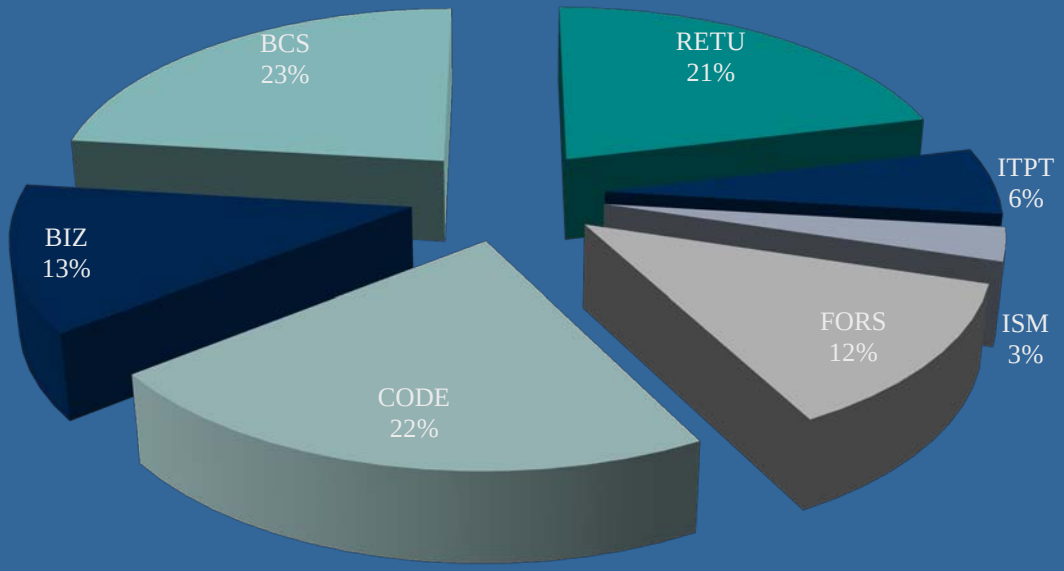
Data Quality and Integrity



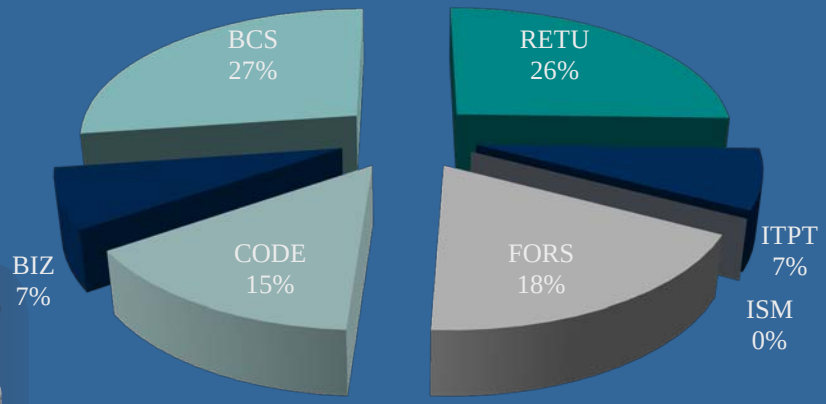
The Challenge

When an organization is building / planning its measurements capabilities and target it to support the business and the decision makers, one of the most critical element in the process is data quality and integrity. If the organization is compromising it all what will come after will be damaged and misleading, therefore will cause more damage then improvements

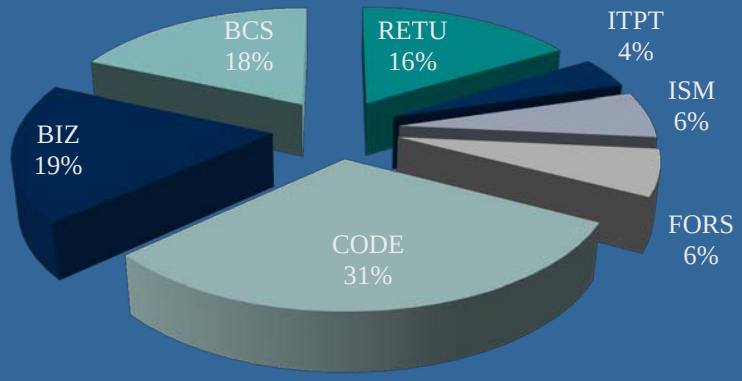
All Areas

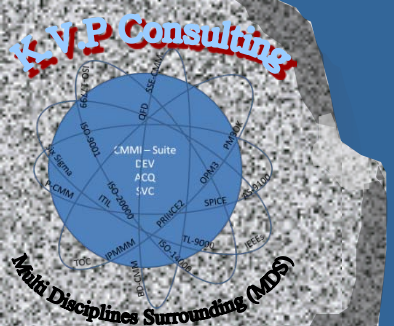


All Areas @ L2

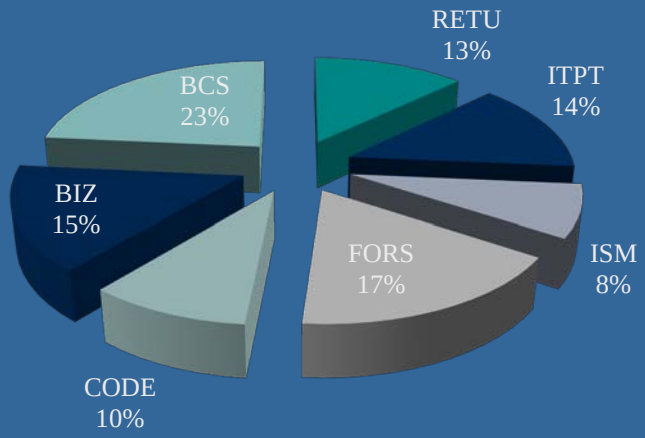


All Areas @ L3

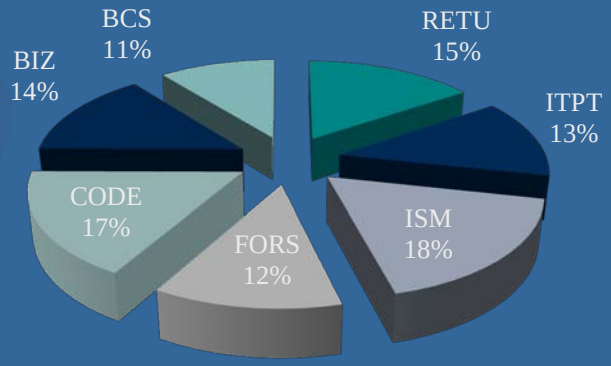




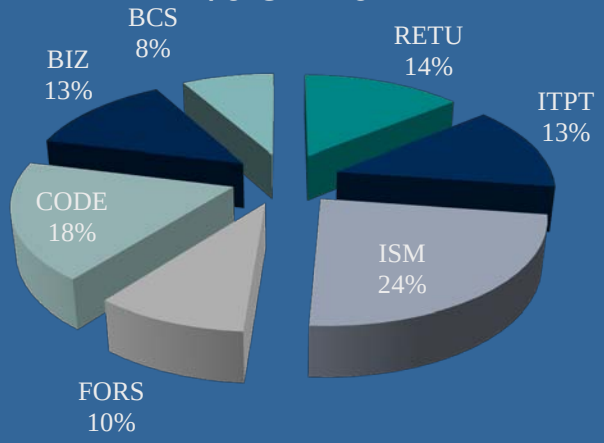
% of ≥ 75%



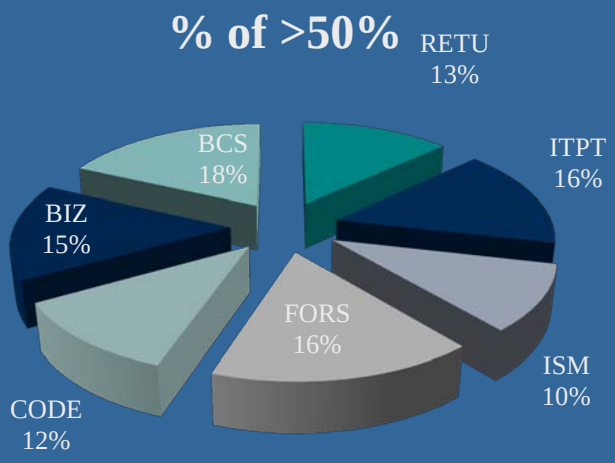
% of <50%



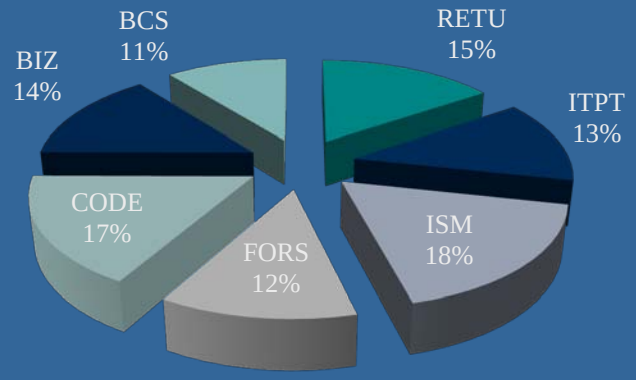
% of = 0



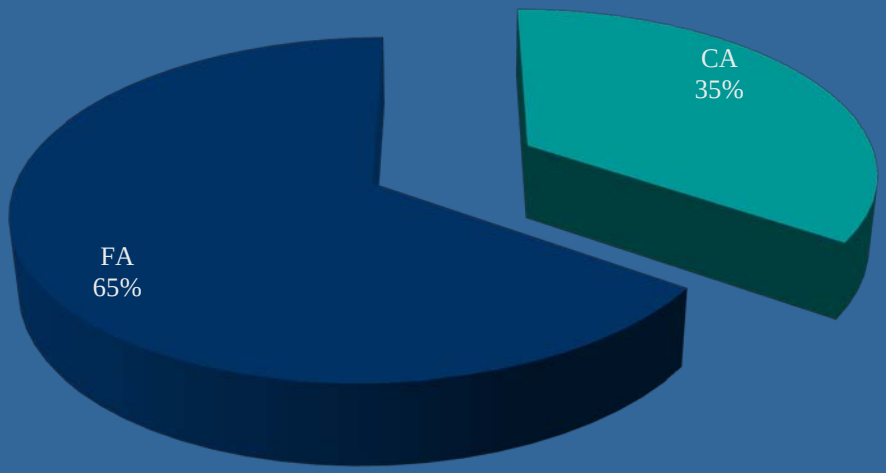
% of >50%



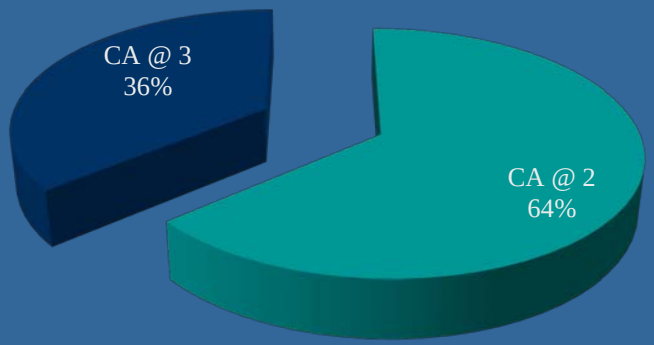
% of <50%



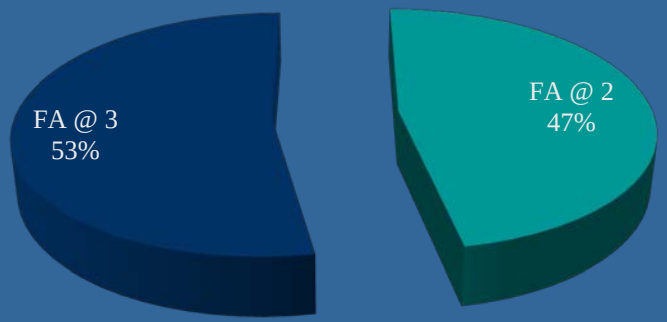
FA vs. CA



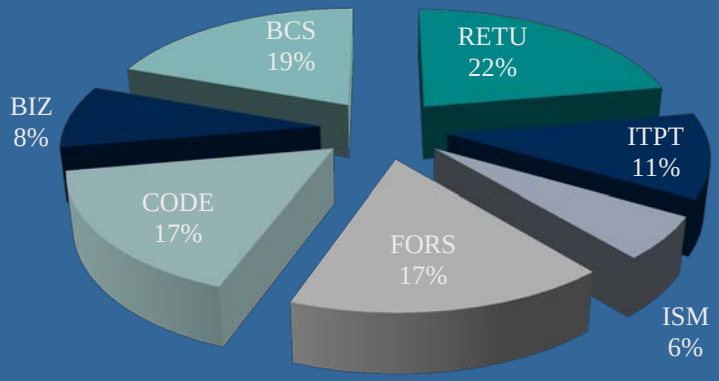
CA vs. Levels



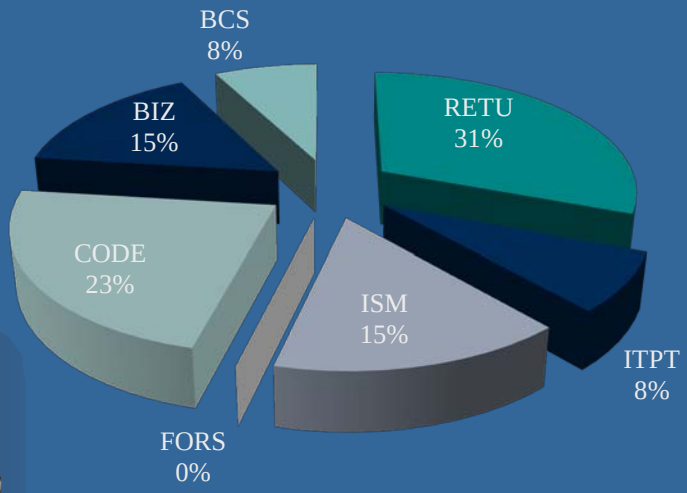
FA vs. Levels



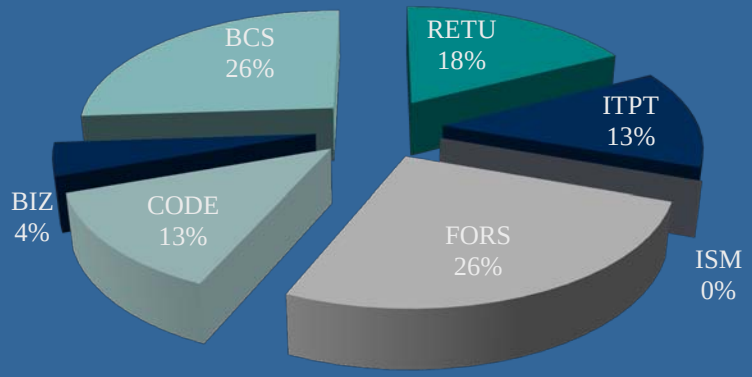
All Areas @ CA



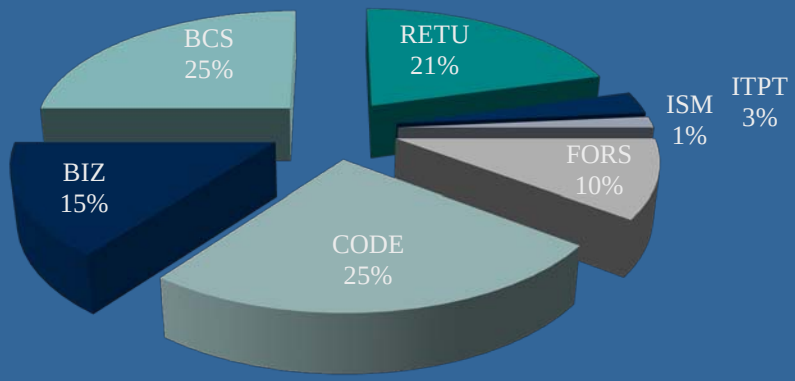
All Areas @ CA ML3



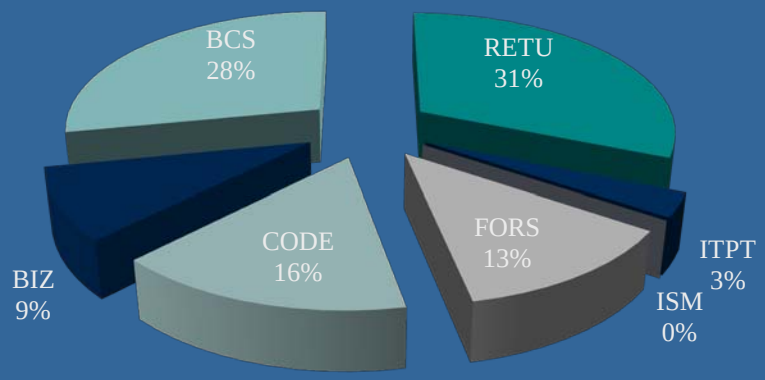
All Areas @ CA ML2



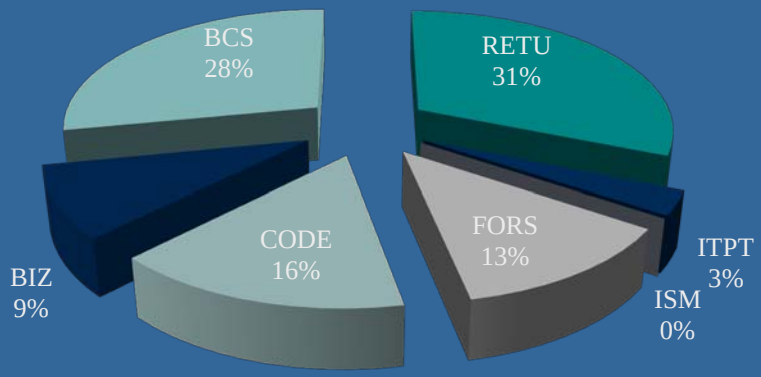
All Areas @ FA

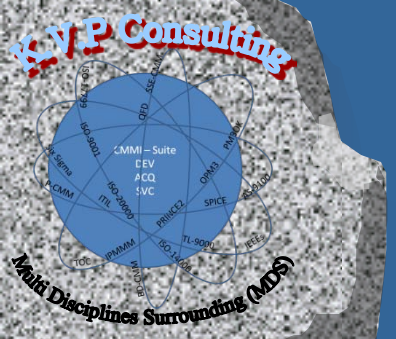


All Areas @ FA ML2



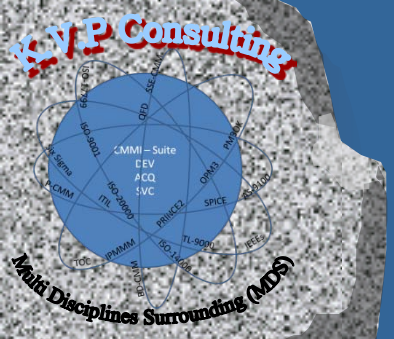
All Areas @ FA ML3





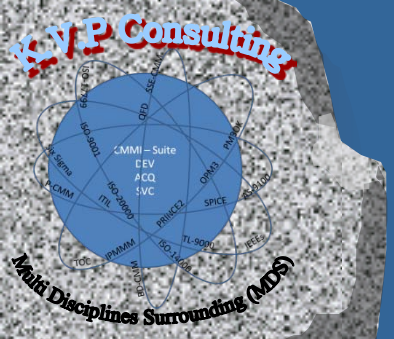
Some Definitions

- "the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs"
- "fitness for purpose"
  - measure of the degree to which the data meets the needs of the particular application
- "performance against specification"
  - how closely does the data fit to the specified requirements for the job



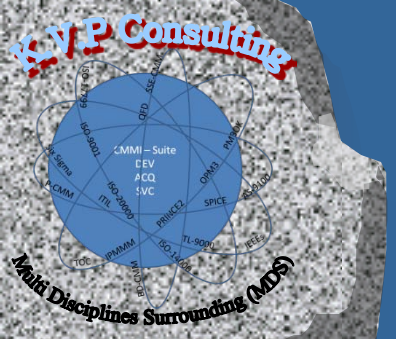
The Insight

- Quality is a relative term not an absolute; it is the relationship between the properties of the data, the purpose for which it is being used and the degree to which the requirements (whether explicitly stated or implied) are being met.
- Statements such as,
  - “this is quality data”, or,
  - “my data is 100%” are meaningless.
- Users need to define exactly what their quality criteria are and state how they are to be evaluated.



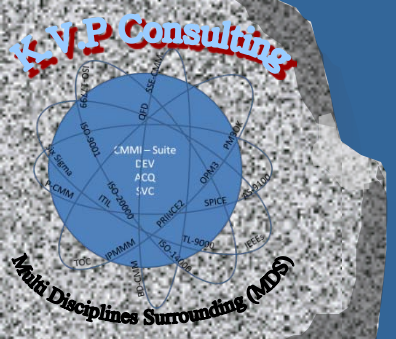
Characterization of Performance-Based Data

- Can be both descriptively and quantitatively
- Descriptively means
  - Purpose,
  - Usage
  - Lineage.
- These are non quantitative and tell potential users of the data:
  - Why the data was captured,
  - How it was created and subsequently modified or maintained
  - How it has been used
- This is enable users to have give a useful indication of the suitability of a dataset for a particular purpose



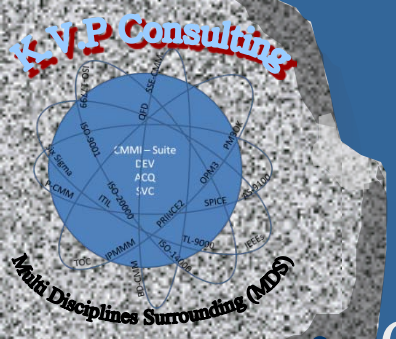
Characterization of Performance-Based Data

- Quantitatively means
- The capability of measurement and can yield quantitative results:
  - Positional accuracy – this can be absolute accuracy - closeness of values to values accepted as being true or relative accuracy - closeness of the relative positions of features in a dataset to the relative positions accepted as being true;
  - Temporal accuracy - accuracy of time measurement. This can include temporal consistency - correctness of ordered events or sequences and temporal validity - the validity of the date assigned to a data item;



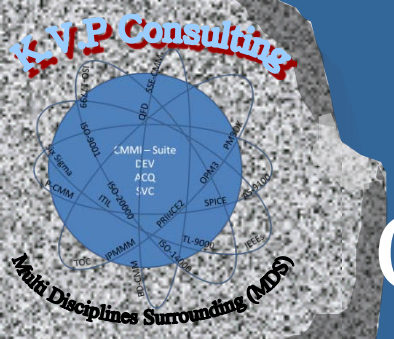
Characterization of Performance-Based Data

- Quantitatively means
- The capability of measurement and can yield quantitative results:
  - Thematic accuracy – accuracy of the attribution of the data. This can include classification correctness – comparison of the classes or attributes assigned to data items to the real world or other sources and non-quantitative and quantitative attribute correctness;
  - Completeness – this is either excess or missing data i.e. commission or omission when compared to the data source at the time of capture;



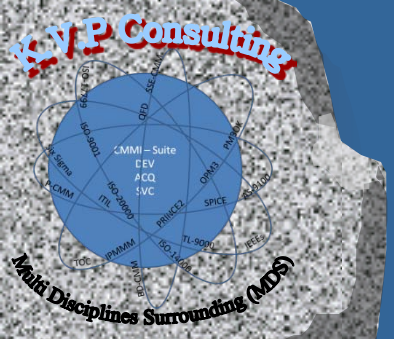
Characterization of Performance-Based Data

- Quantitatively means
- The capability of measurement and can yield quantitative results:
  - Logical consistency – this can include conceptual consistency – conformance to the data model or schema, domain consistency - adherence of values to the value domains, format consistency - degree to which data accords with the physical structure of the dataset and topological consistency – degree to which the geometry is correctly structured topologically.

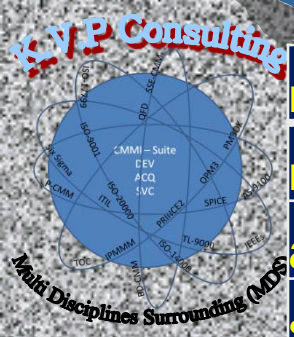


Characterization Data Integrity

- Data accuracy,
- Completeness
- Validity
- Preservation during storage and transfer



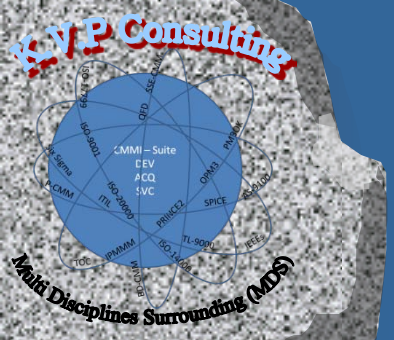
Definitions



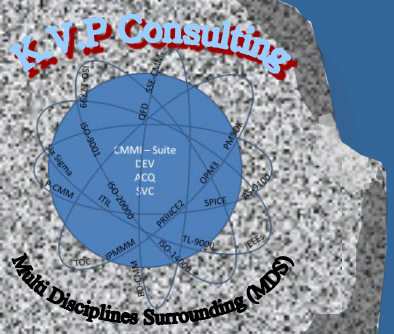
min	Returns the smallest number in a set of values
max	Returns the largest value in a set of values
ave	Returns the average (arithmetic mean) of the arguments
samp	counts the number of cells that contain numbers
>4	Returns the number of cells with value larger then 4
% of >4	Returns the percentage of cells contain numbers that are larger then 4
<4	Returns the number of cells with value smaller then 4
% of <4	Returns the percentage of cells contain numbers that are smaller then 4
is 4	Returns the number of cells with value equal to 4
% of is 4	Returns the percentage of cells contain numbers that are equal to 4
>6	Returns the number of cells with value larger then 6
% of ≥6	Returns the percentage of cells contain numbers that are larger then 6
mean	Returns the geometric mean of an array or range of positive data
median	Returns the median of the given numbers. The median is the number in the middle of a set of numbers
mode	Returns the most frequently occurring, or repetitive, value in an array or range of data
VAR	Estimates variance based on a sample

Understanding Variance

																				Is Consistent	% of Consistency	Variance
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	1	5.00%	35.00
1	1	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	5	25.00%	27.37
1	1	1	1	1	1	1	1	1	1	7	8	9	10	11	12	13	14	15	16	10	50.00%	33.36
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12	13	14	15	16	15	75.00%	33.88
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	100.00%	0.00

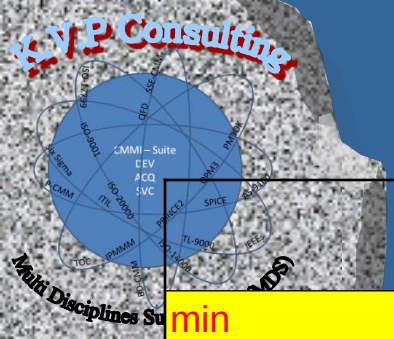


Example walkthrough



Unit perspective analysis

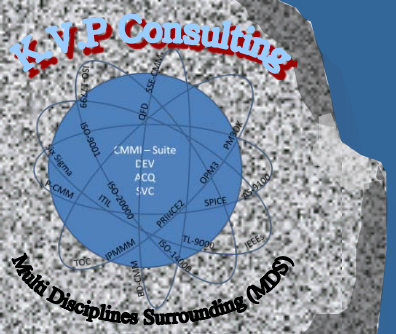
- Center
- Areas
- Focus projects



Center	
min	0%
max	100%
ave	50%
sample Projects	104
% From ORG	100.00%
Sample Practices	19629
% From Sample	100.00%
is 0	2649
% of is 0	13.50%
>4	9147
% of >4	46.60%
<4	7828
% of <4	39.88%
is 4	2654
% of is 4	13.52%
>6	4818
% of >6	24.55%
mean	#NUM!
median	4
mode	8
VAR	7.279

Areas

	RETU	ITPT	ISM	FORS	CODE	BIZ	BCS
min	0%	0%	0%	0%	0%	0%	0%
max	100%	100%	100%	100%	100%	100%	100%
ave	50%	50%	37.5%	62.5%	50%	50%	75%
sample Projects	22	6	3	13	23	13	24
% From ORG	21.15%	5.77%	2.88%	12.50%	22.12%	12.50%	23.08%
Sample Practices	3733	957	647	2069	4961	2914	4348
% From Sample	19.02%	4.88%	3.30%	10.54%	25.27%	14.85%	22.15%
is 0	526	127	154	195	914	378	355
% of is 0	14.09%	13.27%	23.80%	9.42%	18.42%	12.97%	8.16%
>4	1575	476	213	1092	1850	1413	2528
% of >4	42.19%	49.74%	32.92%	52.78%	37.29%	48.49%	58.14%
<4	1626	347	322	705	2358	1165	1305
% of <4	43.56%	36.26%	49.77%	34.07%	47.53%	39.98%	30.01%
is 4	532	134	112	272	753	336	515
% of is 4	14.25%	14.00%	17.31%	13.15%	15.18%	11.53%	11.84%
>6	779	211	82	579	775	733	1659
% of >6	20.87%	22.05%	12.67%	27.98%	15.62%	25.15%	38.16%
mean	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!
median	4	4	4	5	4	4	6
mode	2	6	0	6	0	6	8
VAR	7.058	6.898	6.750	6.853	6.654	7.142	7.265



Analysis Disclaimer 1

Sample Size

CODE	23
TREL	6
CEBL	4
ASMA	1
FORL	3
SECL	4
CUSL	3
ASFI	2

BIZ	13
CUAC	5
HRID	2
PASY	4
PRSY	2

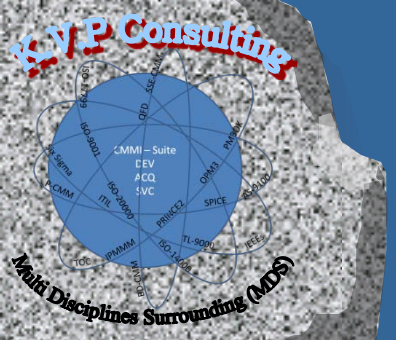
BCS	24
EDW	8
KRED	7
FPAS	3
RMS	1
BIS	4
LOAN	1

RETU	22
BKL	2
REBL	6
BRS	1
REF	7
CRMS	4
PP	2

ITPT	6
DEDA	3
FREM	3

ISM	3
KNOW	2
TEMA	1

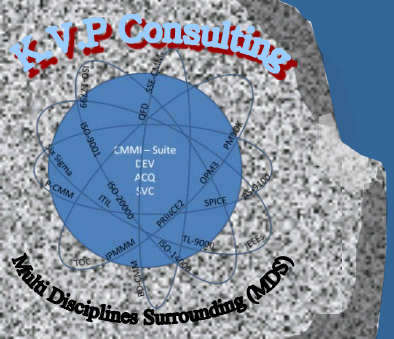
FORS	13
LL	5
BPL	4
LPL	4



Analysis Disclaimer 2

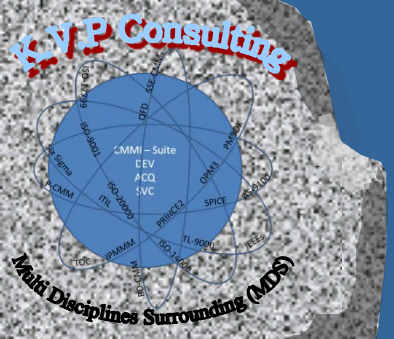
Sample Size

- The area level sample size is vary from 3 projects for an area up to 24
- The department level sample is vary from 1 to 8 sample projects
- Therefore the result decision was not to deep dive in analysis for all areas \ departments
- We have selected the largest in sample size areas for demonstrating the analysis and the expected inputs
- We will be able to provide the same analysis for all; however
- If we will do it on sample smaller than 5 different projects Results are neither accurate nor reflecting insights
- Thus will be done only upon request from an area \ department manager



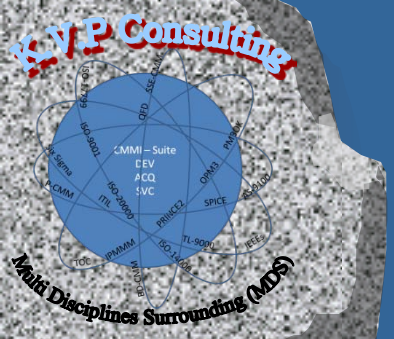
Focus projects

The Selected on Focus Projects; are Only These That We Have the Mid Year and End Year Results For Them



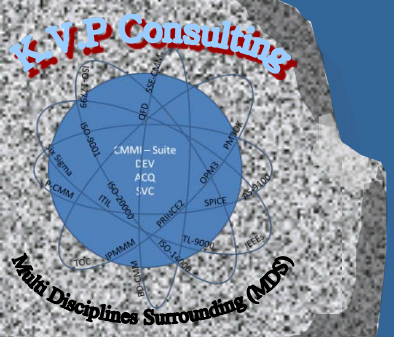
DAPROVOM

	AC @ L2	FA @ L3
min	0%	0%
max	100%	100%
ave	62.5%	62.5%
samp	123	289
>4	84	158
% of >4	68.29%	54.67%
<4	32	106
% of <4	26.02%	36.68%
is 4	7	25
% of is 4	5.69%	8.65%
>6	38	89
% of ≥6	30.89%	30.80%
mean	#NUM!	#NUM!
median	6	5
mode	8	6
VAR	5.456	7.130



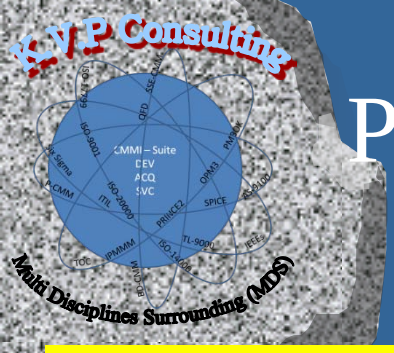
DPAL2008

	CA @ L2	FA @ L3
min	0%	0%
max	100%	100%
ave	50%	50%
samp	122	296
>4	59	141
% of >4	48.36%	47.64%
<4	42	136
% of <4	34.43%	45.95%
is 4	21	19
% of is 4	17.21%	6.42%
>6	18	74
% of ≥6	14.75%	25.00%
mean	#NUM!	#NUM!
median	4	4
mode	6	6
VAR	6.017	7.423



DINF2008

	CA @ L2	FA @ L3
min	0%	0%
max	100%	100%
ave	62.5%	50%
samp	108	288
>4	62	137
% of >4	57.41%	47.57%
<4	28	103
% of <4	25.93%	35.76%
is 4	18	48
% of is 4	16.67%	16.67%
>6	32	63
% of ≥6	29.63%	21.88%
mean	#NUM!	#NUM!
median	6	4
mode	6	6
VAR	5.262	6.853



Process and Product Quality Assurance
SP 1.1 Objectively Evaluate Processes

min	max	ave	samp	>4	% of >4	<4	% of <4	is 4	% of is 4	>6	% of >6	mean	median	mode	VAR
0%	100%	62.5%	104	64	62%	24	23%	16	15%	22	21%	#NUM!	5	6	4.13

- Knowing the development model (the latest version I have reviewed was at Nov.2008) and the missing elements during the end of year assessments with the fact that in the performed QA plans you don't have process evaluation activities (other than the OPF ones) I will challenge the higher than 50% results (under the assumption that we have used it in the past to reflect missing elements in the development model)
- However the result might reflect a need for more in-depth understanding of the practice meaning and context

Requirements Management
SP 1.5 Identify Inconsistencies Between Project Work and Requirements

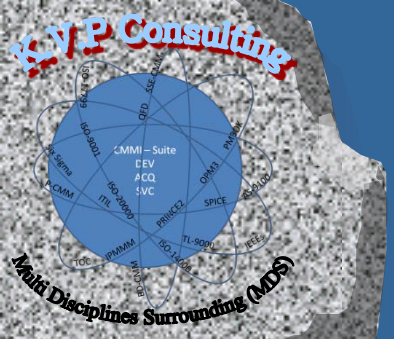
max	ave	samp	>4	% of >4	<4	% of <4	is 4	% of is 4	>6	% of >6	mean	median	mode	VAR
0%	100%	62.5%	104	65	63%	24	23%	15	14%	37	36%	#NUM!	6	6	5.19

- Knowing the development model (the latest version I have reviewed was at Nov.2008) and the missing elements during the end of year assessments with the fact that you are missing Inconsistencies Between Project Work and Requirements I will challenge the higher than 50% results (under the assumption that we have used it in the past to reflect missing elements in the development model)
- However the result might reflect a need for more in-depth understanding of the practice meaning and context

Verification
SP 3.2 Analyze Verification Results

min	max	ave	samp	>4	% of >4	<4	% of <4	is 4	% of is 4	>6	% of >6	mean	median	mode	VAR
0%	100%	62.5%	39	20	51%	12	31%	7	18%	12	31%	#NUM!	5	7	5.13

- Knowing the development model (the latest version I have reviewed was at Nov.2008) and the missing elements and references during the end of year assessments I will challenge the density of results (under the assumption that we have used it in the past to reflect missing elements in the development model)
- However the result might reflect a need for more in-depth understanding of the practice meaning and context
- This practice must go hand in hand with Measurements & Analysis



Specific Practices

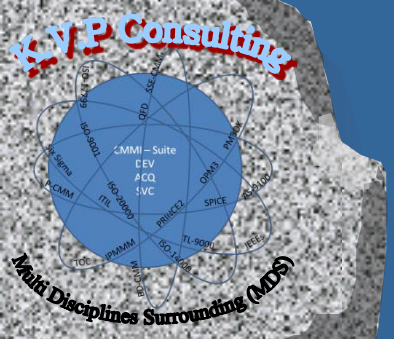
min	0%
max	100%
ave	50%
samp	8454
>4	4303
% of >4	50.90%
<4	2944
% of <4	34.82%
is 4	1207
% of is 4	14.28%
>6	2129
% of ≥6	25.18%
mean	#NUM!
median	5
mode	6
VAR	6.663

GP 2.8 Monitor and Control the Process

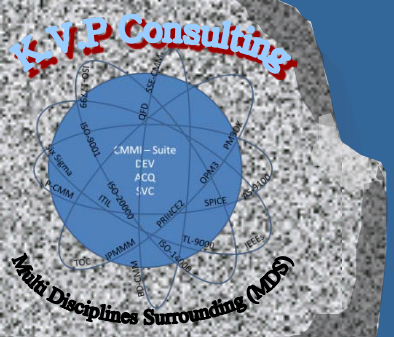
	min	max	ave	samp	>4	% of >4	<4	% of <4	is 4	% of is 4	>6	% of >6	mean	median	mode	VAR
M&A	0%	100%	12.5%	104	10	10%	84	81%	10	10%	6	6%	#NUM!	0	0	4.85
REQM	0%	100%	37.5%	104	18	17%	70	67%	16	15%	6	6%	#NUM!	2	2	4.60
CM	0%	87.5%	12.5%	104	2	2%	95	91%	7	7%	1	1%	#NUM!	0	0	2.03
IPM	0%	75%	12.5%	46	2	4%	40	87%	4	9%	0	0%	#NUM!	0	0	2.53
VAL	0%	100%	37.5%	39	12	31%	20	51%	7	18%	4	10%	#NUM!	3	4	5.03
VER	0%	100%	50%	39	13	33%	23	59%	3	8%	9	23%	#NUM!	3	3	6.68

GP 2.9 Objectively Evaluate Adherence

	min	max	ave	samp	>4	% of >4	<4	% of <4	is 4	% of is 4	>6	% of ≥6	mean	median	mode	VAR
M&A	0%	100%	25%	104	12	12%	74	71%	18	17%	6	6%	#NUM!	2	0	4.83
REQM	0%	100%	62.5%	104	57	55%	26	25%	21	20%	28	27%	#NUM!	5	6	5.29
CM	0%	100%	12.5%	104	7	7%	89	86%	8	8%	2	2%	#NUM!	1	0	3.03
IPM	0%	100%	12.5%	46	1	2%	41	89%	4	9%	0	0%	#NUM!	0	0	2.17
VAL	0%	100%	37.5%	39	12	31%	17	44%	10	26%	4	10%	#NUM!	4	4	5.92
VER	0%	100%	50%	39	14	36%	14	36%	11	28%	2	5%	#NUM!	4	4	5.41

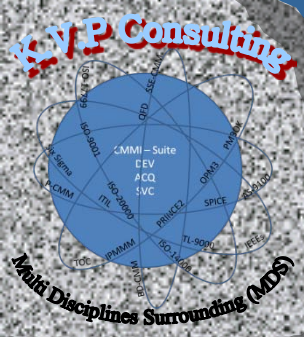


Data Quality and Integrity as 'Satellite' Project



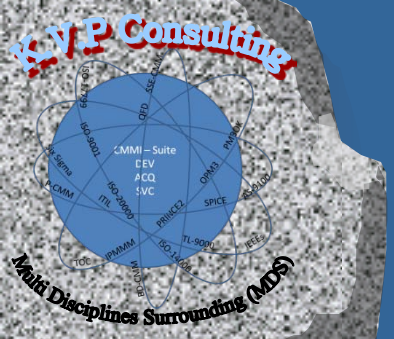
Background

- Addresses data quality issues in cooperative scenarios
- Contributions
  - A model for representing data and quality data
  - A methodology
  - A software architecture for data quality diffusion and improvement



Cooperative Information System

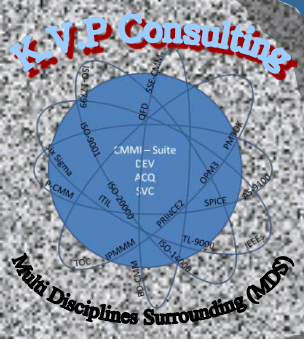
- Distributed system composed by a set of cooperating organizations
  - organizations are heterogeneous and independent
- Service-based cooperation
- Common communication infrastructure
- Organizations exports data and quality data
- Organizations can self-evaluate the quality of their own data



Quality Improvement

- Replication of a same data within the system is exploited for quality improvement with comparison and reconciliation algorithms

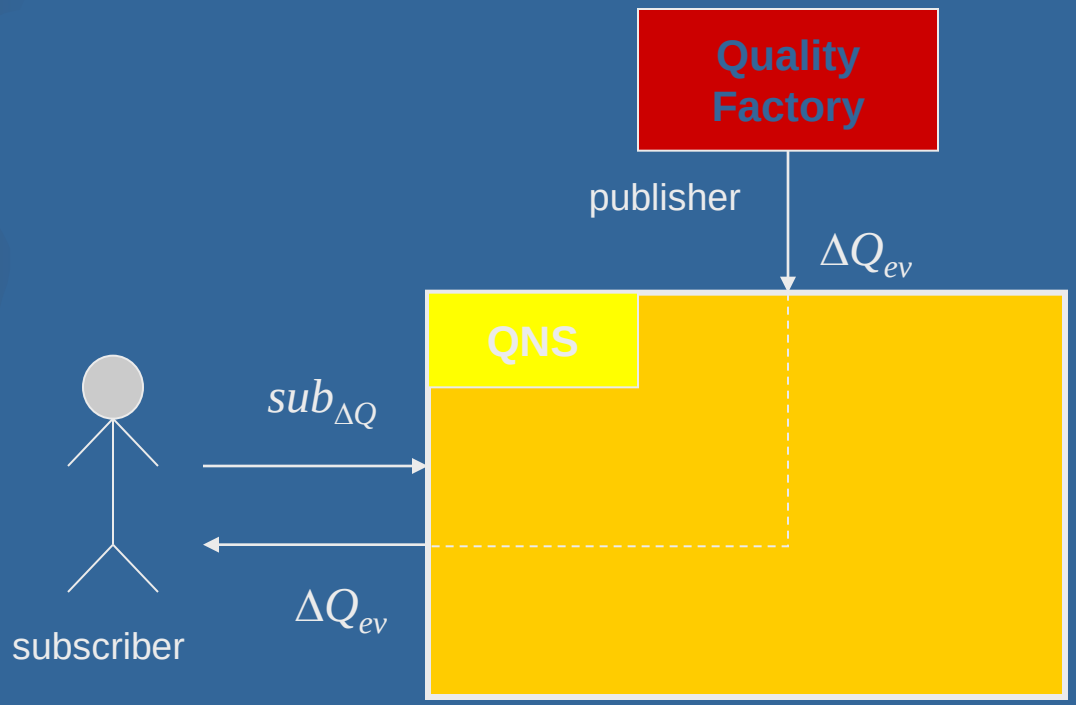
- Quality Improver: *Off-line improvement*
  - periodically matches records over different databases and tries to reconcile non-exact matches
- Data Quality agent: *On-line improvement*
  - performs queries based on quality constraints. Chooses the best copies and gives a feedback
- Quality Notification Service: *Quality maintenance*
  - Notifies quality changes to monitor overall quality

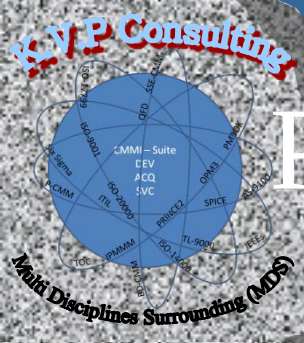


The Quality Notification Service (QNS)

- Notifies users for changes in quality of data
- Follows the publish/subscribe *(pub/sub)* paradigm
  - users subscribe to QNS using a specific subscription
  - when a change in quality happens, the QNS fires a corresponding event
  - the event is notified to all interested subscribers
- Can be used to:
  - keep track of quality changes to prevent degradation
  - automatically activate other architectural services
  - maintain overall quality at an acceptable level

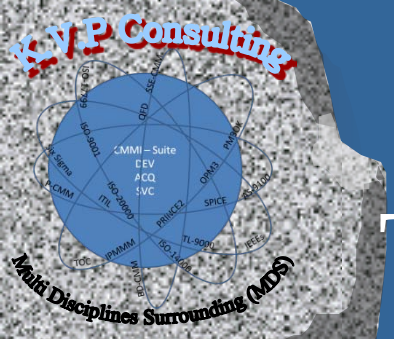
The Quality Notification Service





Publish/Subscribe Middleware variants

- Existing pub/sub middleware can be classified according to the level of expressiveness of their subscription language
- **Topic-based systems**
  - events are grouped in topics
  - subscribers simply declare their interest for a specific topic
  - each topic corresponds to a logical event channel
    - examples: TIB/RV, CORBA Event Service, JMS
- **Content-based systems**
  - subscriptions are filters on the event content
  - a subscription is a set of constraints on the attributes of an event
  - constraints include comparison operators and can be composed in AND/OR
    - examples:SIENA, IBM Gryphon, Elvin
- Hybrid approaches
  - allow to express filters over a channel
    - examples: MS COM+ Event Service, CORBA Notification Service

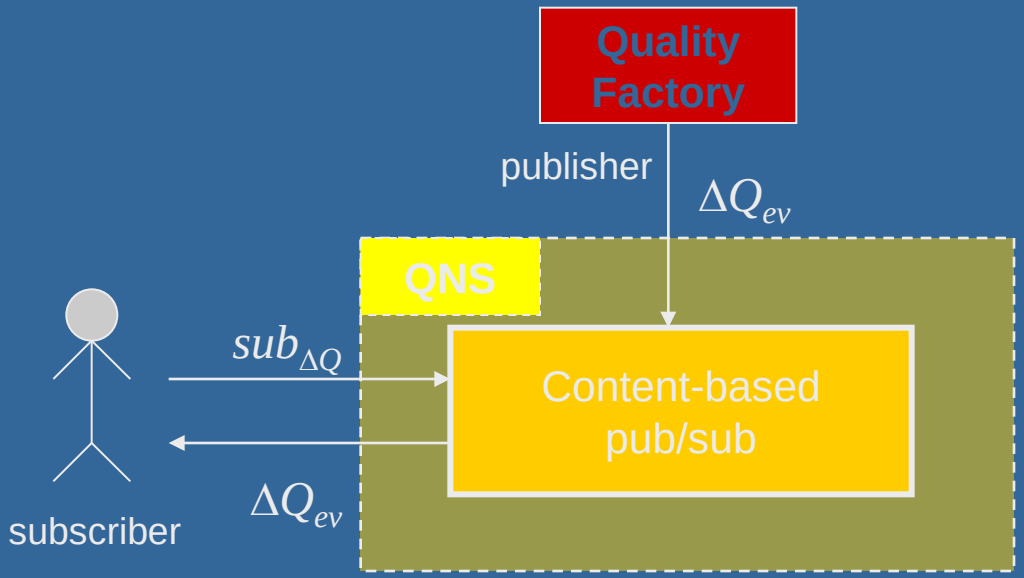


Topic-based vs. Content-based

- Trade-off between expressiveness and scalability
- Topic-based limits expressiveness but it is more efficient
  - Subscriber set for a publication is known a-priori
  - Can exploit multicast
  - Many efficient implementations are available
- Content-based is more expressive but hardly scale
  - Have to calculate receivers for each event (“matching”)
  - Events must be efficiently propagated (“routing”)

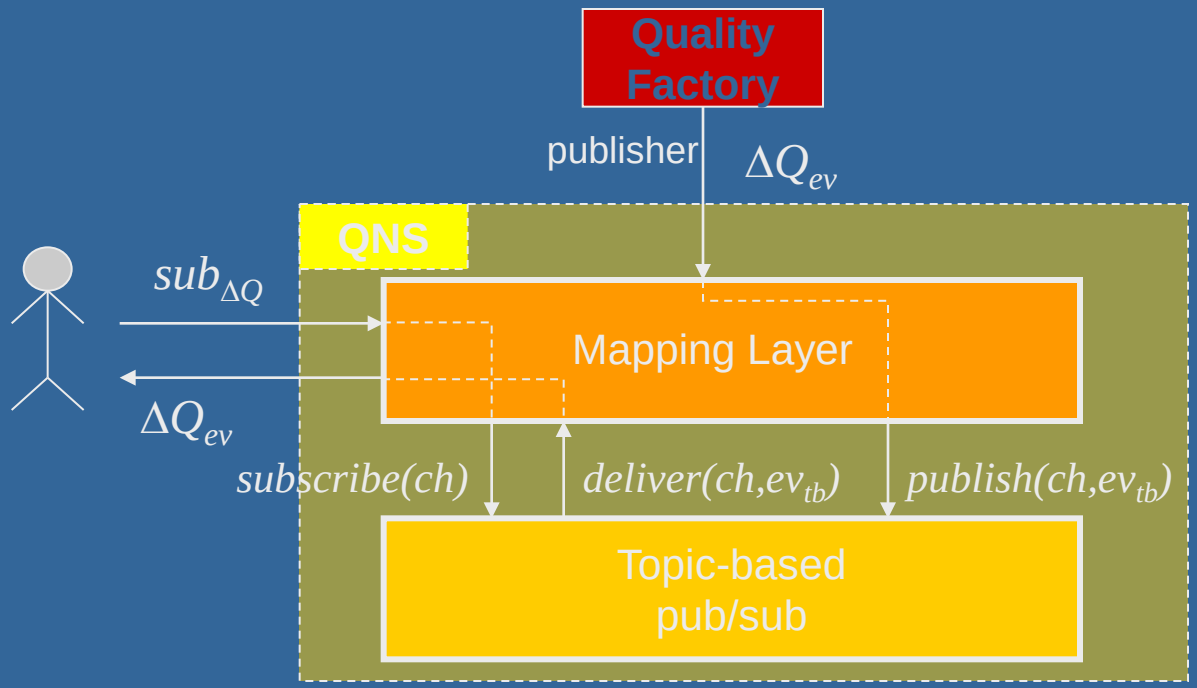
QNS: Content-based implementation

- Straightforward mapping of QNS language to tool-specific language



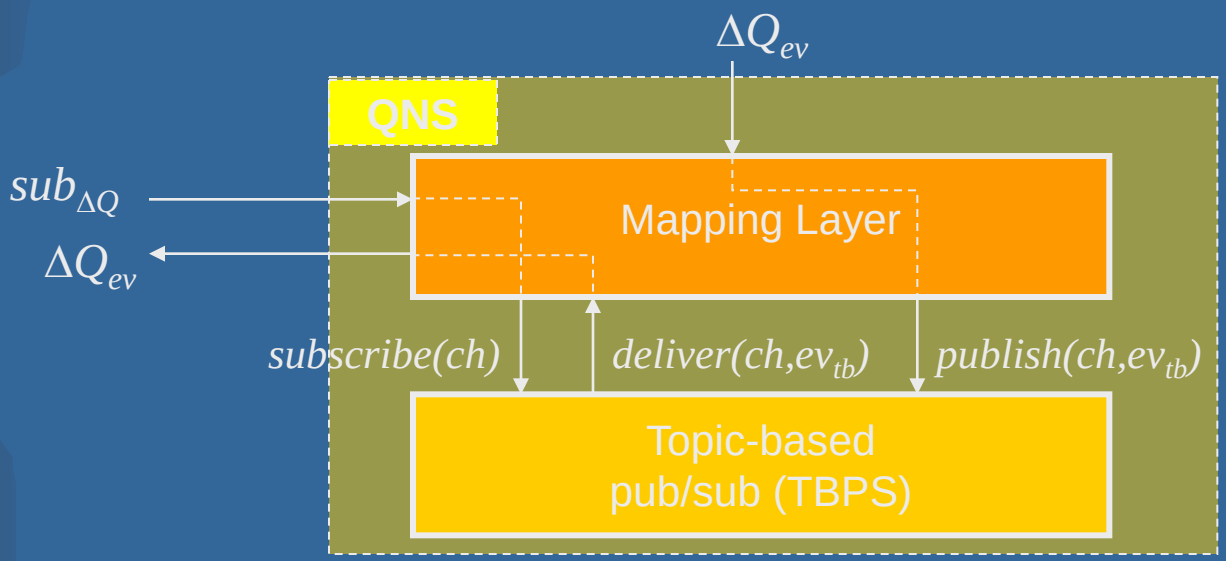
QNS: topic-based implementation

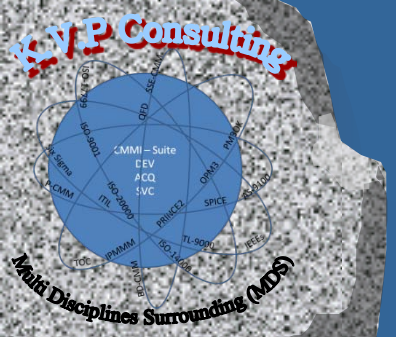
- Requires additional processing to emulate content-based behaviour
- Implemented by a Mapping Layer inside QNS



QNS: Mapping Layer

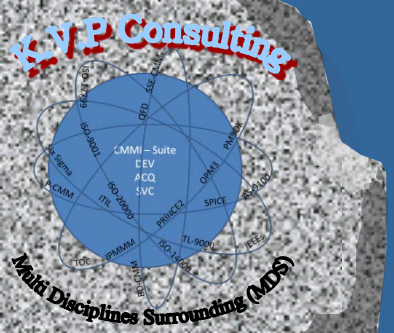
- maps QNS subscriptions $sub_{\Delta Q}$ into TBPS channels *ch*
- decides on which TBPS channel *ch* each QNS event ΔQ_{ev} should be published
- delivers events ev_{tb} from TBPS to interested QNS subscribers in the form of QNS events ΔQ_{ev}
  - implements comparison contraints





Mapping policies

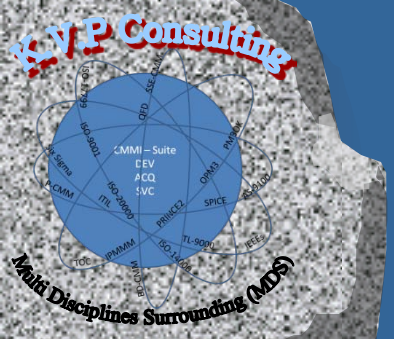
- General problem of emulating a content-based system with a topic-based one
- Cost metrics
  - *number of channels*: too many channels cloaks the TBPS level
  - *non-precision*: too few channels generate unnecessary network traffic
- Example policy
  - *channel-per-entity*: each channel corresponds to a different entity
- 3 policies are presented and evaluated in the paper
  - can be combined
- No evident one-size-fit-all solutions
  - experimental evaluations needed



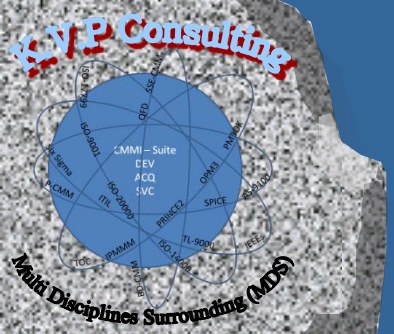
Conclusions and Future Work

- We presented a set of different solutions for implementing a Quality Notification Service upon a pub/sub middleware system
- No solution is better that the others
- Evaluation must be done for the specific case

- Future work
  - Experimental evaluation on real-world data of the different proposals
  - choose one solution and implement the service



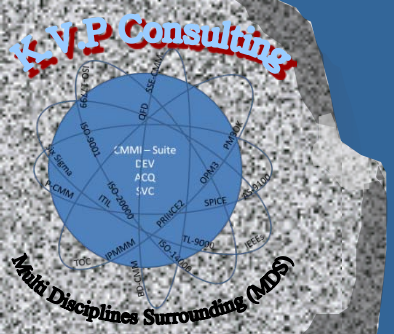
Data Quality and Integrity as 'What If' Scenarios



Benchmark Requirements

- **Relevancy**
  - Relevant for the domain of interest
- **Portability**
  - Portable to different systems
- **Scalability**
  - Applicable to small and large systems
- **Simplicity**
  - Easy to understand and implement

(Jim Gray: The Benchmark Handbook for Database and Transaction Systems, Morgan Kaufmann,1993)



The OID-Benchmark

A **Benchmark for Object Identification** is a triple (***D***, ***Q***, ***S***),

- ***D*** is a **benchmark database**,
- ***Q*** is a set of **quality criteria**,
- ***S*** is a **test specification**.

Benchmarking Example

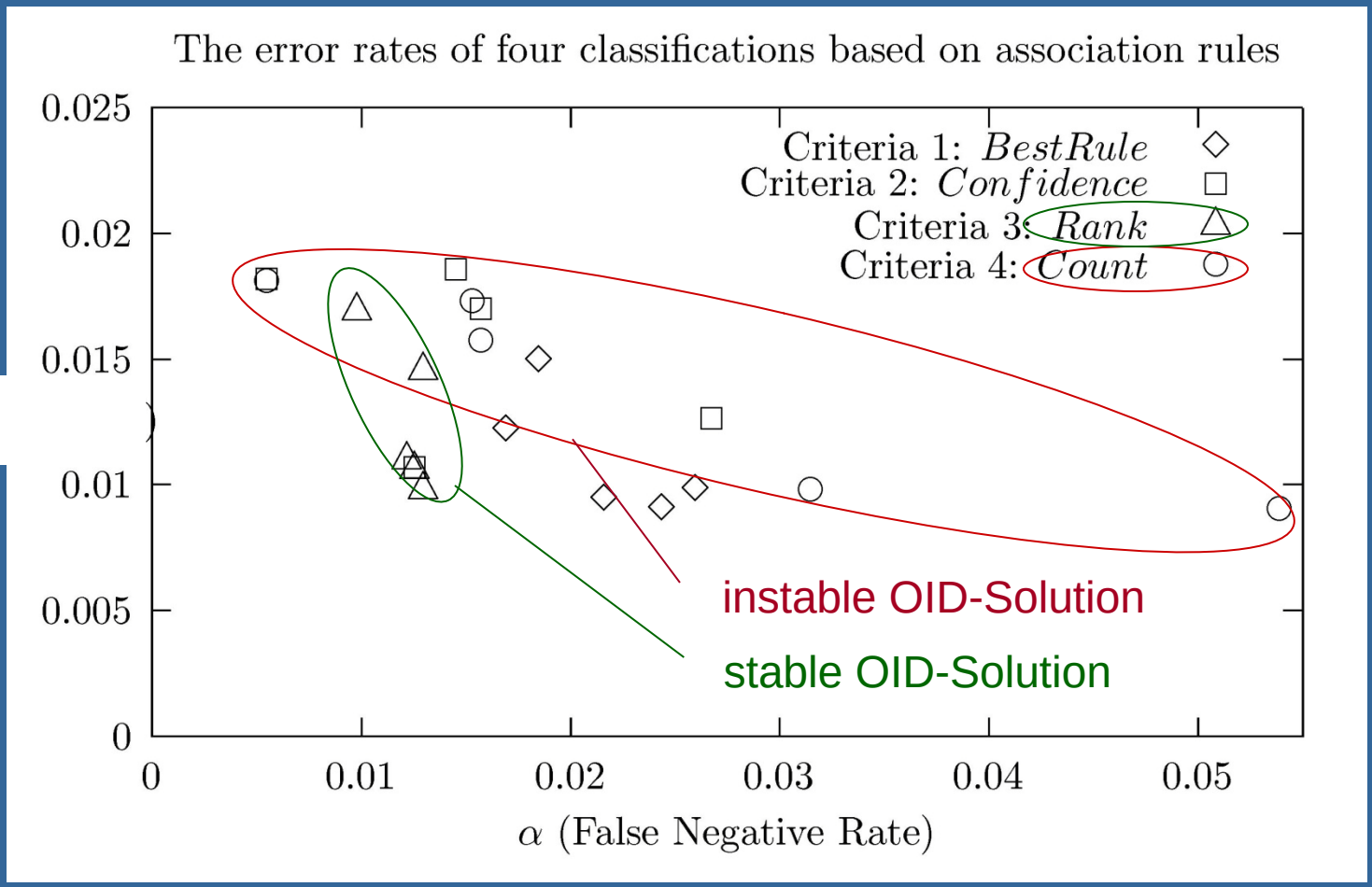
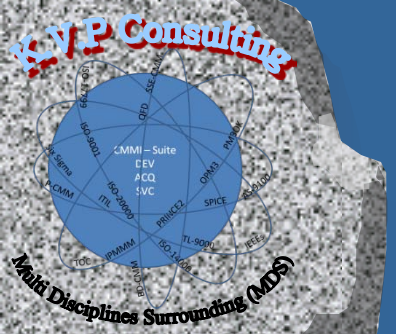


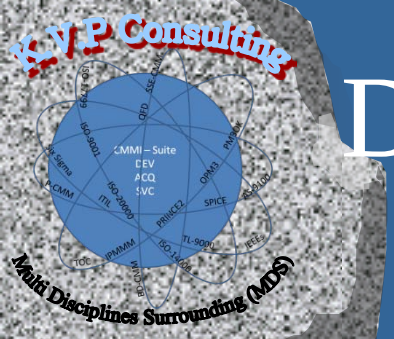
Fig. Correctness measured for samples from the database classified by aggregated Association Rules



Summary & Outlook

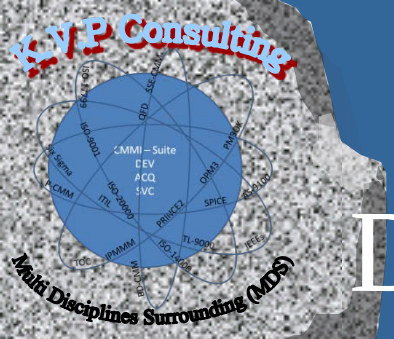
- Object Identification Quality is divided into two,
  - The quality of data, described by data characteristics,
  - The quality of object identification solutions, e.g. correctness.
- The Test Framework enables the comparisons,
  - Moreover, Benchmarks for Identification analogous to the ORG - Benchmark can be established.


Mirror, mirror upon the wall,
Who is the fairest fair of all?
O Lady Queen, though fair ye be,
Snow-White is fairer far to see.
Over the hills and far away,
She dwell with seven dwarfs to-day!



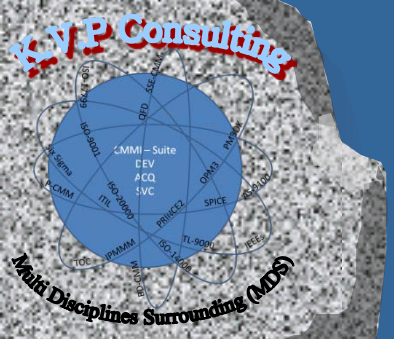
Discussion Some of the Leading Challenges and Issues

- Quality of statement
- Inconsistencies in reporting data
- Lack of quality evaluation
- Lack of data specification or feature
- No statement of requirements
- “Poor fit” across different datasets
- Inaccurate, inconsistent, incomplete and misleading information
- Lack of referential integrity in cross-referencing of business and objectives
- Problems with data sharing and interoperability because of a lack of
- Inefficiencies in operations because of missing, inaccurate or out-of-date data
- Costs resulting from invalid or incorrect results

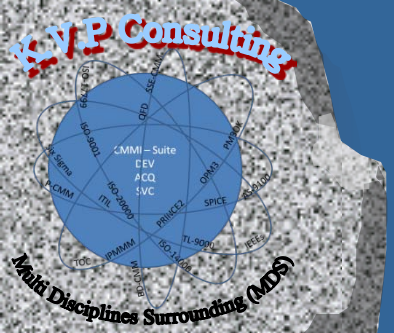


Discussion on Potential Impacts

- Inconsistencies in reporting data
- No statement of requirements
- “Poor fit” across different datasets
- Inefficiencies in operations because of missing, inaccurate or out-of-date data


J.P. Consulting
CMMI – Suite
DEV
ACQ
SVC
Multi Disciplines Surrounding (MDS)

Questions ?



Contact

Kobi Vider

K.V.P Consulting

Kobi.Vider@hotmail.com

KobiVP@aol.com

Phone: +972522946676