

Data Quality Indexes

Theo van Hintum

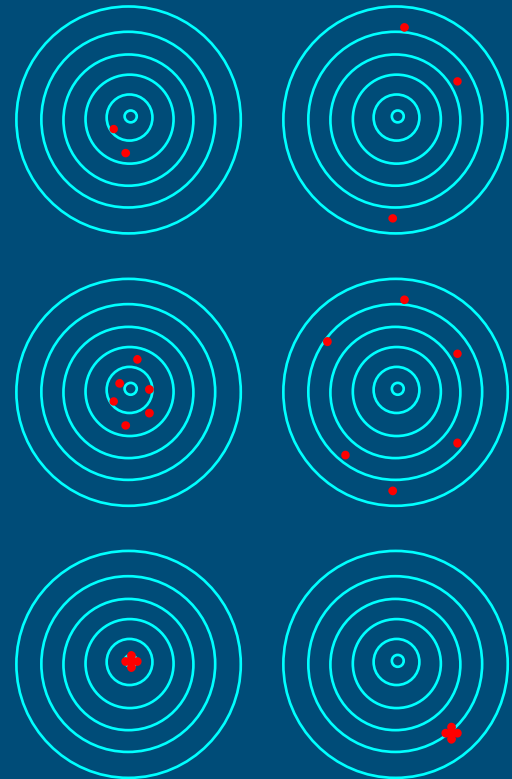
GCP Workshop on Passport Data Quality

July 4th, 2007

Data Quality Indexes

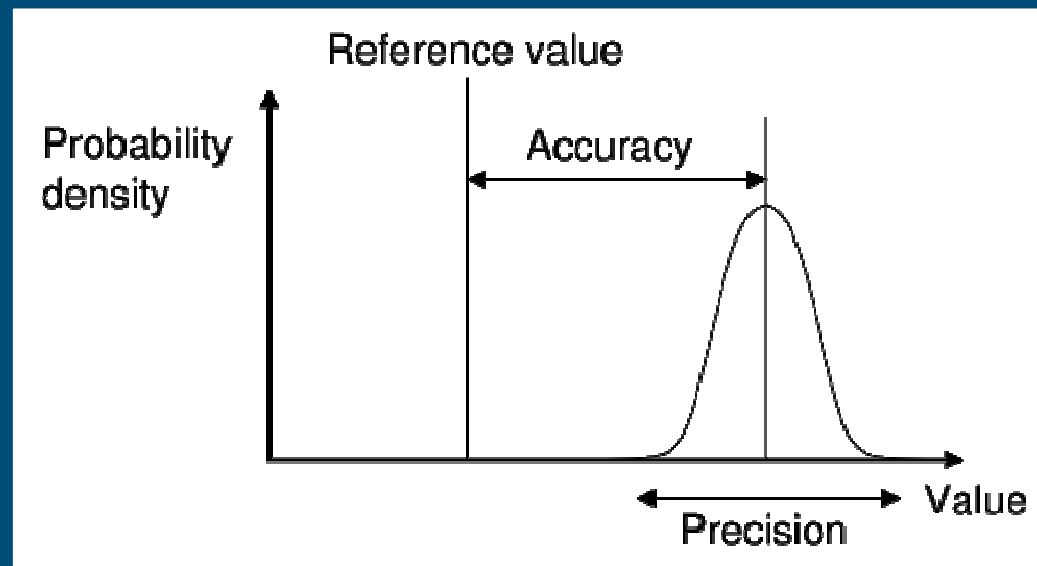
■ what do we mean with 'Data Quality'

- availability / completeness
 - is the data there?
- accuracy
 - closeness of measured values, observations or estimates to the real or true value
- precision
 - also called reproducibility or repeatability
 - the degree to which further measurements or calculations show the same or similar results
 - characterized in terms of the standard deviation of the measurements



Data Quality Indexes

- what do we mean with 'Data Quality'



Data Quality Indexes

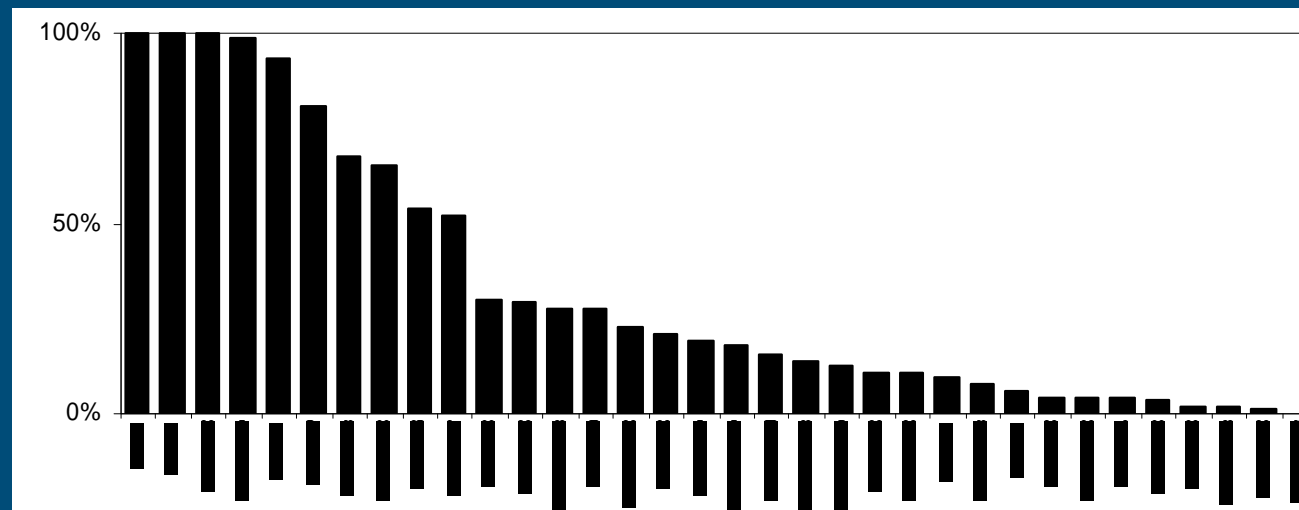
- what do we mean with 'Data Quality' for passport data?
 - availability / completeness
 - is the data there?
 - consistency
 - do the data follow the rules (format, coding, values)?
 - correctness
 - do the data reflect reality

Data Quality Indexes

- how do we quantify 'Data Quality' for passport data?
 - availability / completeness
 - how many descriptors have values?
 - consistency
 - how many errors in terms of format, coding and undeclared values can be found?
 - correctness
 - how many incorrect data can be found?

Data Quality Indexes

- focus on availability / completeness
 - completeness per descriptor
 - data EURISCO 2003



Data Quality Indexes

- focus on availability / completeness
 - completeness per descriptor doesn't tell the whole story
 - accession name is irrelevant if it concerns a wild accession
 - an index was created to assess completeness in its context: the Documentation Quality Index

Data Quality Indexes

- focus on availability / completeness
 - Documentation Quality Index (DQI)
 - based on EURISCO
 - calculation based on the occurrence of values in descriptors which were rewarded systematically
 - ranged from zero to a maximum of 150
 - minimum is ascribed if only the mandatory descriptors (NICODE, INSTCODE, ACCENUMB and GENUS) have values
 - not all descriptors are relevant for each sample status (SAMPSTAT), sub sets of SAMPSTAT-independent and SAMPSTAT-dependent descriptors were distinguished for calculation of the DQI - e.g. a collection site is not relevant for a breeder's variety

Data Quality Indexes

- focus on availability / completeness

- Documentation Quality Index (DQI)

- SAMPSTAT independent : maximum total 100

Descriptor	DQI-value	Constraints/Remarks	Descriptor	DQI-value	Constraints/Remarks
0 NICODE	0	Mandatory field	27 STORAGE	5	If '99' check [28]
1 INSTCODE	0	Mandatory field	28 REMARKS	0	
2 ACCENUMB	0	Mandatory field	31 DONORDESCR	2	[23] = null
5 GENUS	0	Mandatory field	32 DUPLDESCR	2	[26] = null
6 SPECIES	25	if null then [7] and [8] and [9] = 0	33 ACCEURL	5	
7 SPAUTHOR	2	[6] not null			
8 SUBTAXA	5	[6] not null			
9 SUBTAUTHOR	1	[6] not null, [8] not null			
10 CROPNAME	10				
12 ACQDATE	3				
20 SAMPSTAT	15	If '999' check [28]			
23 DONORCODE	7	[31] = null			
24 DONORNUMB	10 / 7	10 if ([23] or [31] not null); else 7			
25 OTHERNUMB	7				
26 DUPLSITE	5	[32] = null			

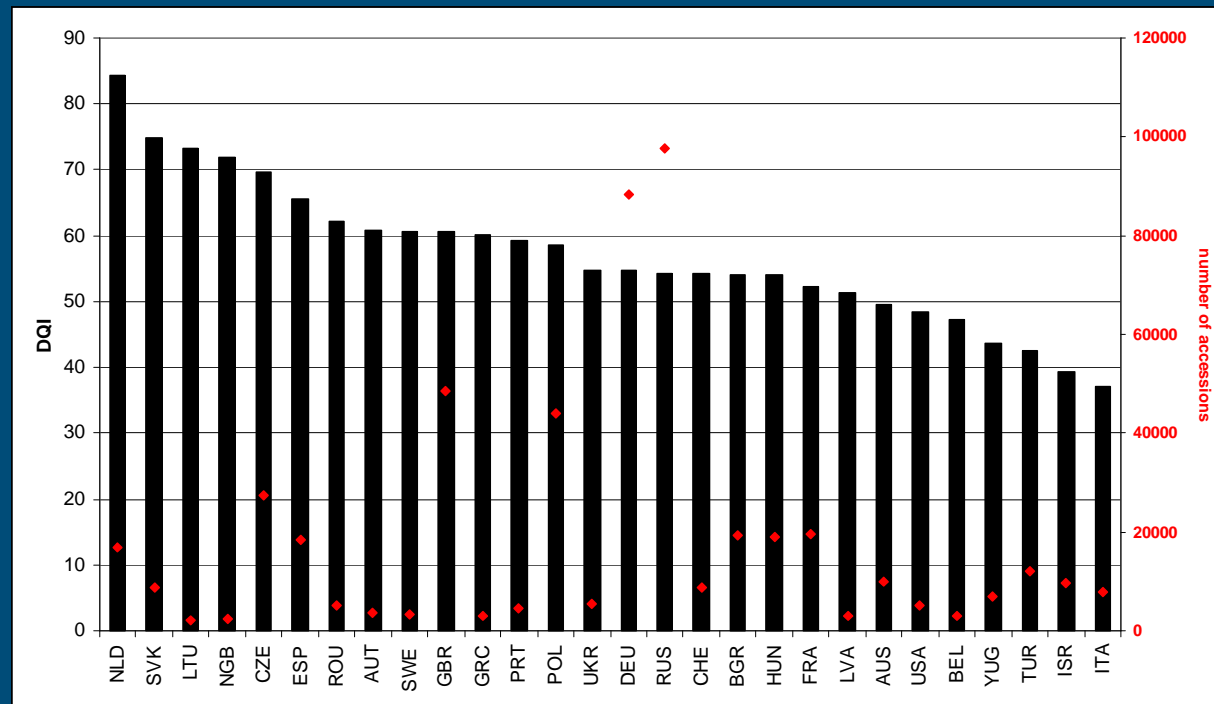
Data Quality Indexes

- focus on availability / completeness
 - Documentation Quality Index (DQI)
 - SAMPSTAT dependent : maximum total 50

	Descriptor	Wild / weedy	Trad. cult.	Research material	Advanced cultivar	999 or Null
3	COLLNUMB	7	7	0	0	2
4	COLLCODE	5 if [29] not null	5 if [29] not null	0	0	2
11	ACCENAME	0	5	10	20	5
13	ORIGCTY	10	10	6	6	6
14	COLLSITE	12 / 3	7 / 2	0	0	3
15	LATITUDE	7 if [16] not null	5 if [16] not null	0	0	2
16	LONGITUDE	7 if [15] not null	5 if [15] not null	0	0	2
17	ELEVATION	2	2	0	0	1
18	COLLDATE	4	4	0	0	2
19	BREDCODE	0	0	15 if [30] null	10 if [30] null	2
21	ANCEST	0	2	16	11	5
22	COLLSRC	5 if '99' check [28]	3 if '99' check [28]	3 if '99' check [28]	3 if '99' check [28]	2
29	COLLDESCR	3 if [4] not null	3 if [4] not null	0	0	1
30	BREDESCR	0	0	4 if [19] null	3 if [19] null	1

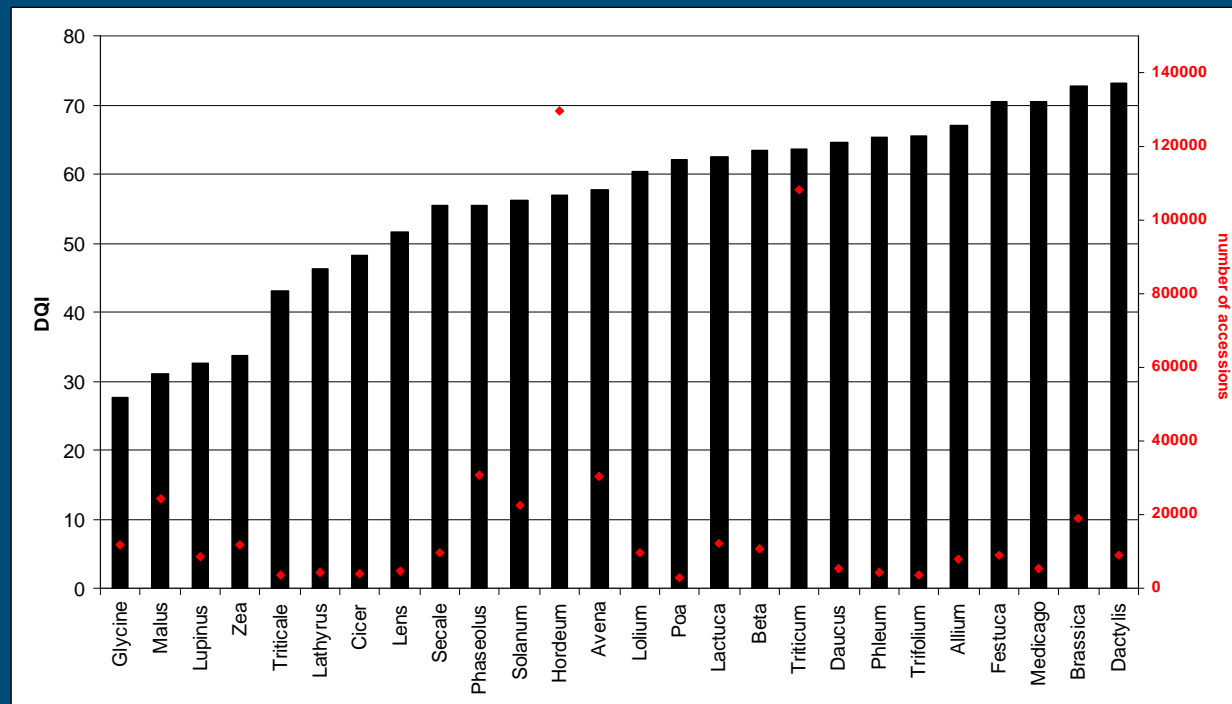
Data Quality Indexes

- focus on availability / completeness
 - Documentation Quality Index (DQI)
 - data EURISCO 2003: DQI vs. NI country



Data Quality Indexes

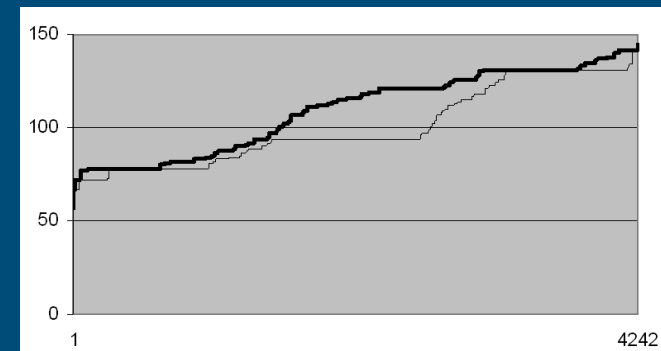
- focus on availability / completeness
 - Documentation Quality Index (DQI)
 - data EURISCO 2003: DQI vs. crop



Data Quality Indexes

- practical use Data Quality Indicator
 - in selecting material for utilisation DQI can be one of the factors
 - CGN uses the DQI to set PRIORITY in its on-line 'core selector'
 - in targeting, monitoring and reporting improvement projects DQI or similar tools can be used
 - example from CGN project to improve passport quality

DQI	Before	After	Increase	Increase %
Pepper	114.9	116.7	1.8	1.2
Eggplant	98.3	119.8	21.5	14.3
Cucumber	83.8	94.4	10.6	7.1
Tomato	107.0	115.8	8.8	5.9



Data Quality Indexes

- indices for availability / completeness
 - can be quantified
 - e.g. DQI – concept should be worked out further and applied in setting priorities

Data Quality Indexes

- indices for (internal) consistency
 - taxonomical nomenclature
 - % 'accepted names'
 - origin data
 - % long/lats within country borders
 - address codes
 - % existing codes
 - '999'codes
 - % properly decoded in remarks

Data Quality Indexes

- indices for correctness
 - origin data
 - % collected samples within distribution area
 - requires additional external data