Nexa Center for Internet & Society
Politecnico di Torino

Studying the Internet, exploring its potential & experimenting new ideas

La qualità degli Open Data

Master in Ingegneria dei dati

Torino, 9 ottobre 2014

Sommario

- Introduzione
  - Gli Open Data
  - Perché è importante la qualità degli Open Data
- Una prima analisi sulla qualità degli degli OD pubblicati dalle PA
  - Caso di studio: La trasparenza nella PA - il Decreto Trasparenza
  - Analisi fattuale: analisi dei formati, errori di pubblicazione, quantità di informazioni
  - I risultati nei capoluoghi di regione italiani
- La qualità intrinseca del dato
  - Un modello per valutarla
  - Le dimensioni prese in considerazione
  - Caso di studio: Open Coesione - best practice Italiana
  - Alcuni risultati dell'analisi dei dati
- Conclusioni

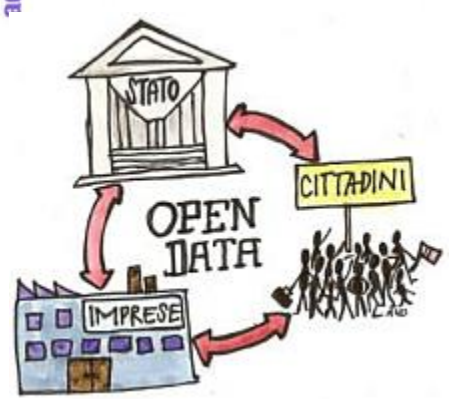
Gli Open Data

Dati che possono essere **usati**, riusati e **ridistribuiti** da chiunque, soggetti al massimo ai requisiti di **attribuzione** e ***share-alike***

Open Definition 2.0: I dati e i contenuti aperti possono essere **usati, modificati e condivisi liberamente** da **chiunque** e per **qualsiasi scopo** (soggetti al massimo a requisiti che preservino la provenienza e l'apertura)

Vantaggi

- Trasparenza nei governi (OGD)
- Partecipazione dei cittadini
- Stimolo all'innovazione attraverso il riuso dei dati.
- Esempi:



OPENCOESIONE
Verso un migliore uso delle risorse: scopri, segui, sollecita.

P TO

DataMarket

DATA.POLICE.UK

2

La qualità dei Dati

- **Bassa qualità** ➡ poco potenziale di **riuso** e alti costi associati al riuso (a volte troppo alti)
- **Esempio:** ParcheggiaTO ➡ dati riutilizzabili se agiornati, completi e standardizzati
- **Cause** della bassa qualità, punto di vista dell'utilizzatore:
  - dati anche di alta qualità all'interno dell'organizzazione (memorizzati in sistemi che non prevedono la pubblicazione) sono pubblicati senza seguire una **procedura di apertura formalizzata**:
    - Metadati mancanti, poca comprensibilità
    - Visualizzazione statica di un database: problemi di attualità, coerenza, accuratezza (duplicazioni)

La qualità dei Dati

- **Strumenti** già esistenti per aprire i dati:
  - CKAN: ha integrato Open Refine per controlli sulla qualità dei dati
  - SOCRATA: fornisce warning su dati con problemi relativi ai metadati

Come analizzare la qualità dei Dati aperti

- **Verifica «fattuale»:** il file è pubblicato? È in formato machine processable? Contiene abbastanza informazioni?
- **Analisi della qualità intrinseca del dato**: il dataset è completo? È accurato? È attuale? È descritto con dei metadati appropriati?

Sommario

- Introduzione
  - Gli Open Data
  - Perché è importante la qualità degli Open Data
- Una prima analisi sul possibile riuso degli OD pubblicati
  - Caso di studio: La trasparenza nella PA - il Decreto Trasparenza
  - Analisi fattuale: analisi dei formati, errori di pubblicazione, quantità di informazioni
  - I risultati nei capoluoghi di regione italiani
- La qualità intrinseca del dato
  - Un modello per valutarla
  - Le dimensioni prese in considerazione
  - Caso di studio: Open Coesione - best practice Italiana
  - Alcuni risultati dell'analisi dei dati
- Conclusioni

Verifica fattuale
Caso di studio: La Trasparenza nelle PA

- Decreto Trasparenza (d.lgs. n.33, 14 marzo 2013) disciplina gli obblighi di pubblicità, trasparenza e diffusione delle informazioni da parte delle Pubbliche Amministrazioni (PA) ed attua la legge anticorruzione (190/2012)
- Molti dataset sono stati pubblicati, tuttavia sussistono diversi problemi:
  - Dataset non presenti
  - Quantità di informazioni insufficiente
  - Dataset difficili da interpretare
  - Formati non aperti
  - Dati troppo aggregati

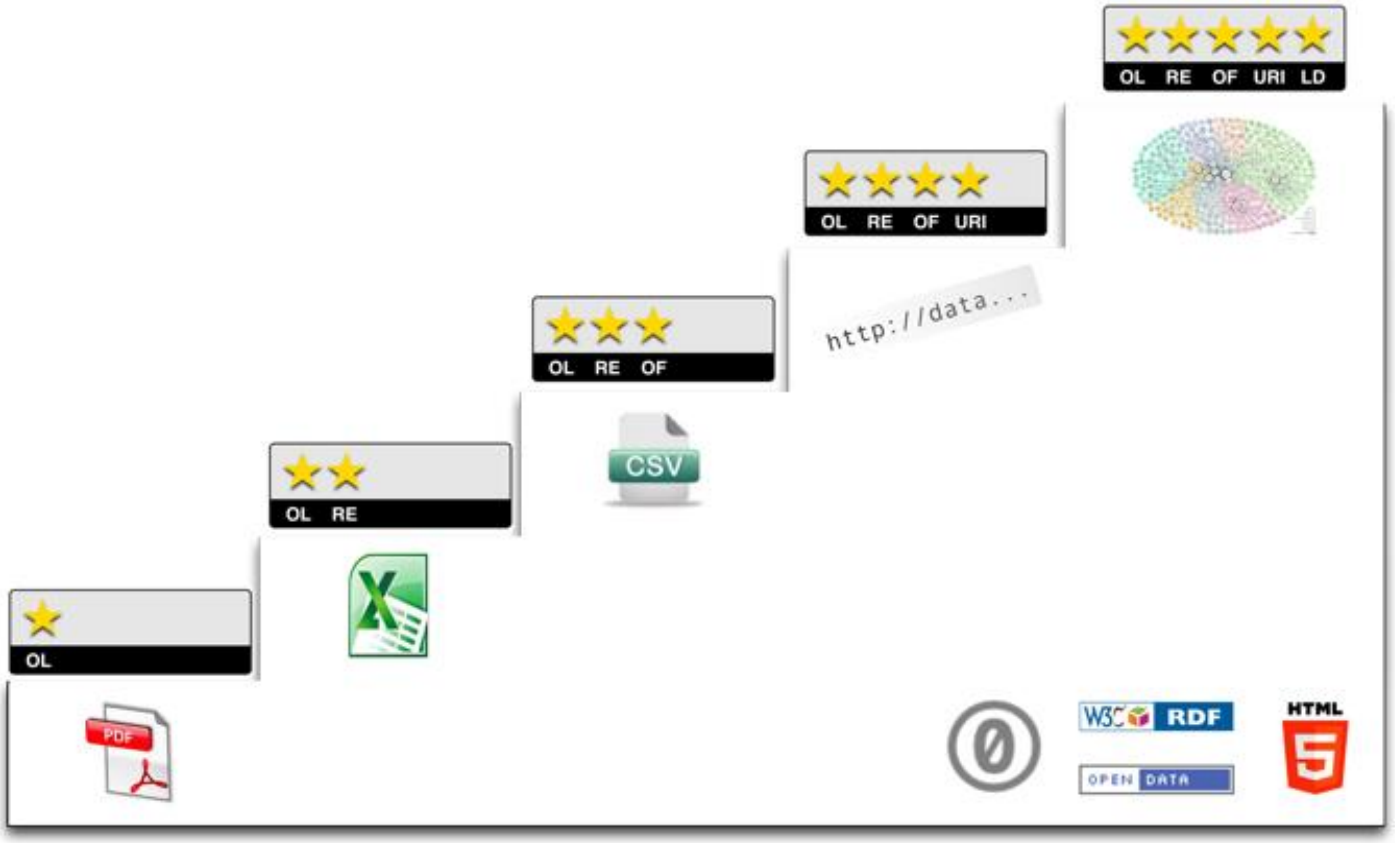
Trasparenza della Pubblica Amministrazione: i primi risultati di un'analisi di dettaglio

- Censiti i dataset dei 20 capoluoghi di Regione Italiani
- Sezioni scrutinate
  - Sovvenzioni, Sussidi e contributi (Art.26-27 d.lgs. n. 33/2013)
    - Albo beneficiari
    - Atti di concessione
  - Beni Immobili e patrimonio immobiliare (Art. 30, d.lgs. n. 33/2013)
    - Patrimonio immobiliare,
    - Canoni di fitto attivo,
    - Canoni di fitto passivo
- Censiti 100 dataset

Le dimensioni per una prima analisi

- **Formato del file**
  - Il file è open e machine readable oppure è raster?
- **Errori di pubblicazione**
  - Il file è stato pubblicato? Se è pubblicato è di qualche utilità (non troppo aggregato, comprensibile, in formato tabellare?)
- **Quantità di informazioni**
  - Quanti e quali attributi sono presenti in ciascun dataset? Forniscono una quantità di informazione sufficiente?

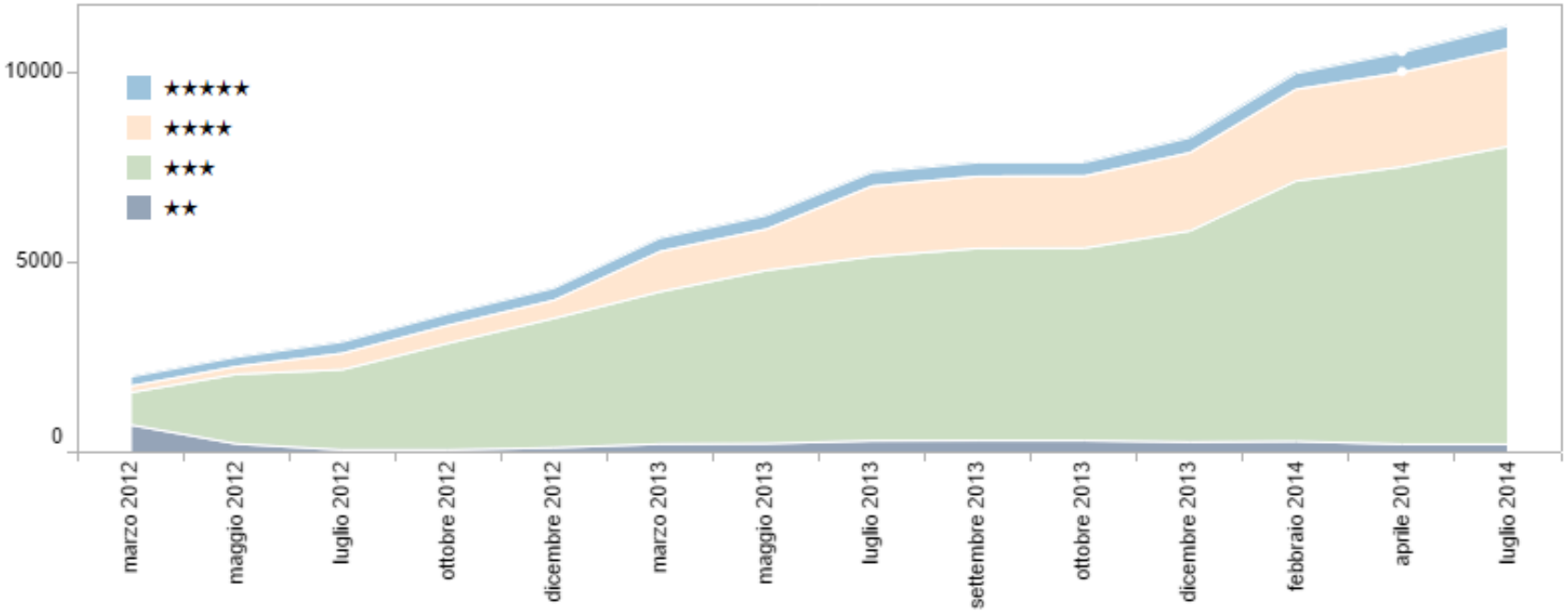
Il formato dei file – Five Star Open Data (FSOD)



Fonte: http://5stardata.info/

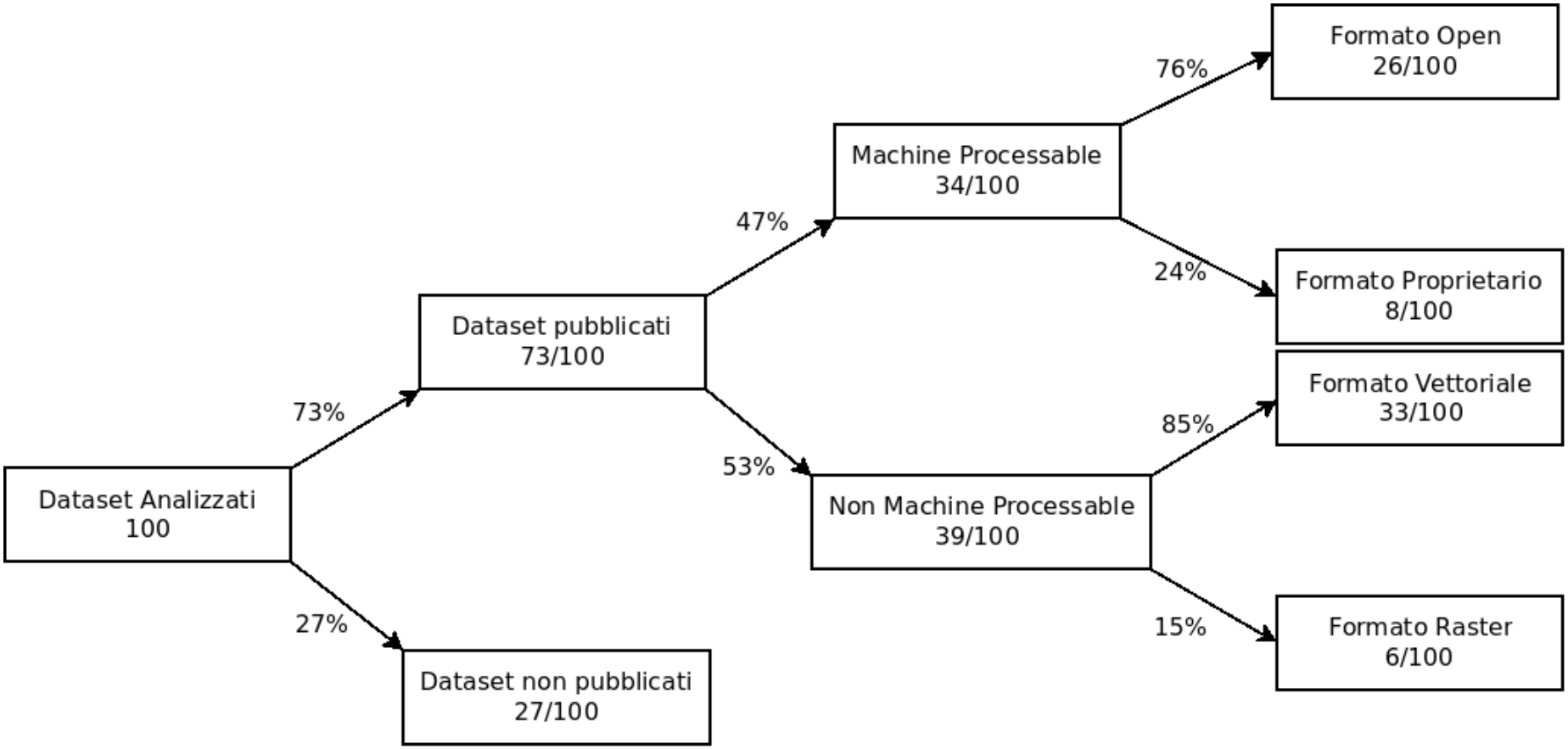
I formati degli Open Data in Italia

Circa 11200 dataset rilasciati in formato aperto, analisi da marzo 2012 fino a giugno 2014

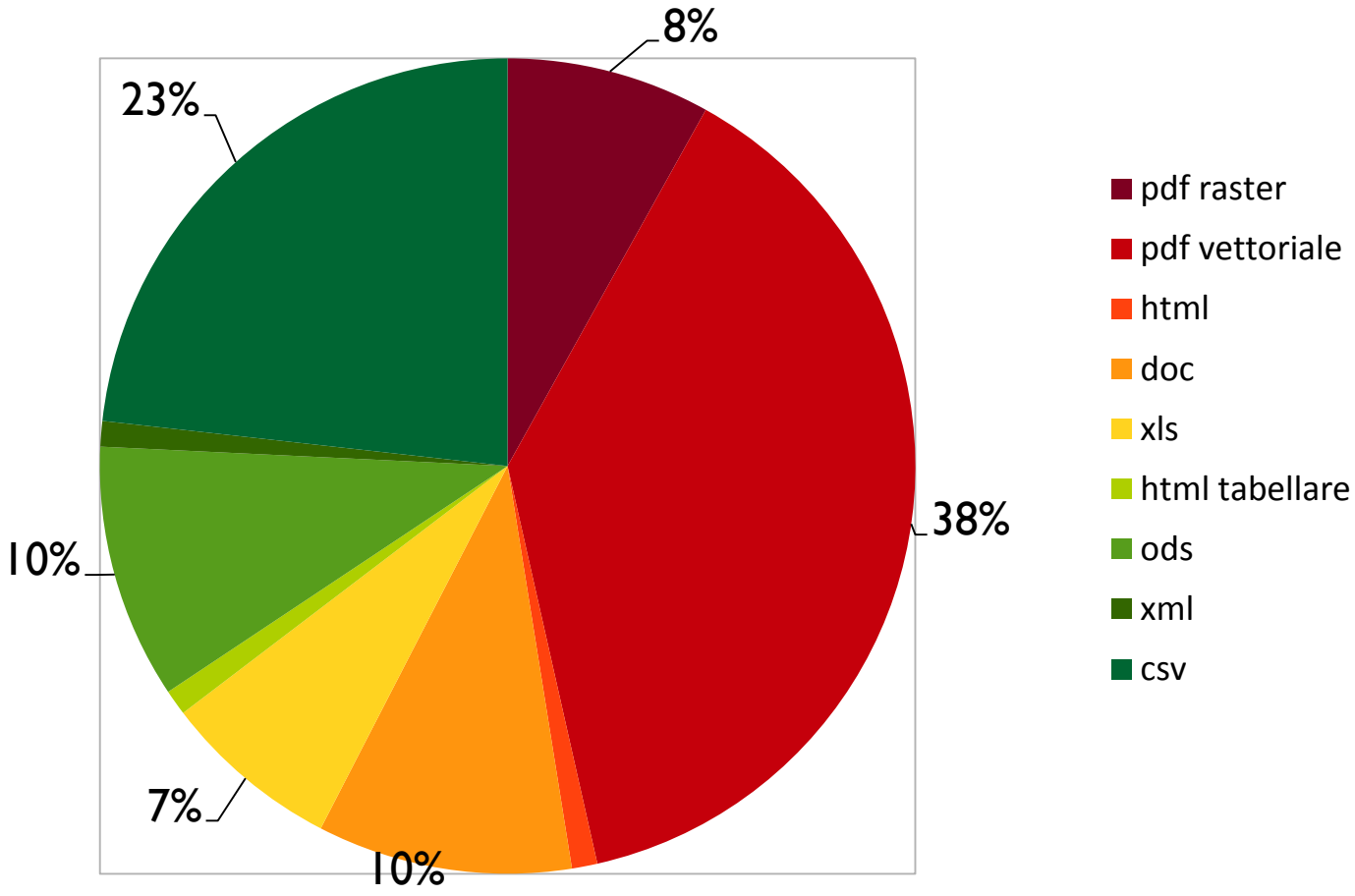


Fonte: http://www.dati.gov.it/content/infografica#Quanti sono i dati aperti in Italia?

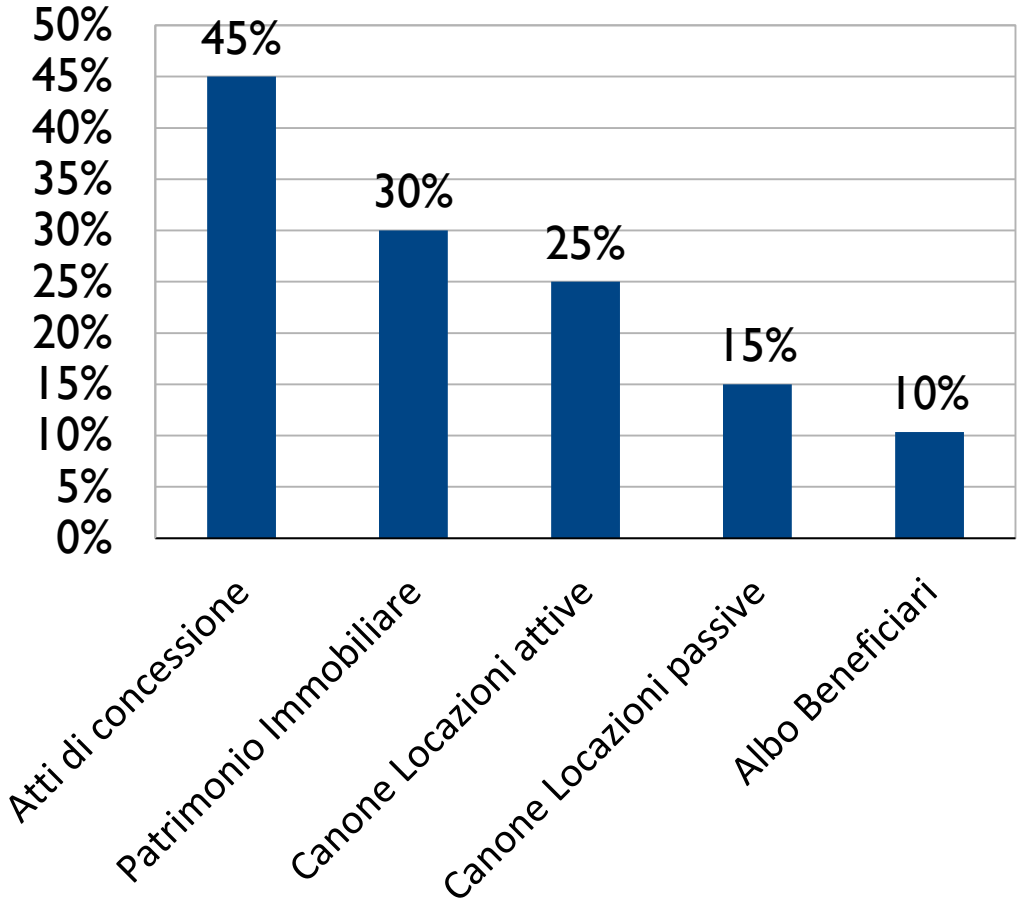
Analisi dei formati – Capoluoghi di Regione



Analisi dei formati – Capoluoghi di Regione



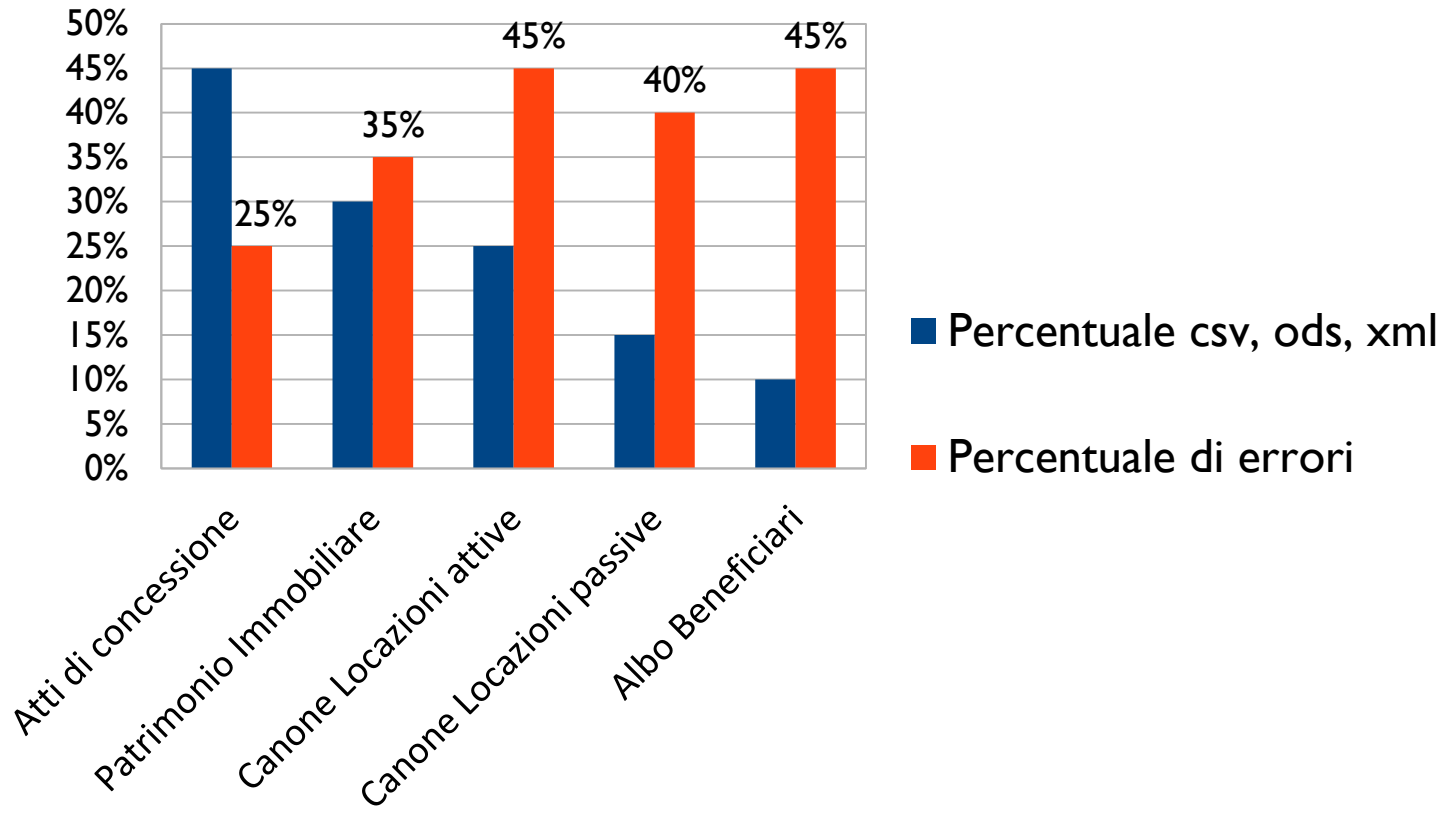
Formati aperti e machine processable - Sezione



Errori di pubblicazione

- 38% dei file analizzati sono di fatto inutilizzabili
  - File non pubblicati = 27
  - File con errori di pubblicazione (es: non tabellare, non comprensibile, troppo aggregato) = 11

Errori per sezione



Quantità di informazioni

- Numero di attributi pubblicati su 10 prescritti per legge (art 27 c. 1 d.lgs 33/2013)
- Media: 4,7/10
- Varianza molto alta

Sommario

- Introduzione
  - Gli Open Data
  - Perché è importante la qualità degli Open Data
- Una prima analisi sul possibile riuso degli OD pubblicati
  - Caso di studio: La trasparenza nella PA - il Decreto Trasparenza
  - Analisi fattuale: analisi dei formati, errori di pubblicazione, quantità di informazioni
  - I risultati ottenuti analizzando i capoluoghi di Regione italiani
- La qualità intrinseca del dato
  - Un modello per valutarla
  - Le dimensioni prese in considerazione
  - Caso di studio: Open Coesione - best practice Italiana
  - Alcuni risultati ottenuti dall'analisi della qualità dei dati
- Conclusioni

Analisi della qualità intrinseca del dato

- Verificabile una volta che il file sia stato **pubblicato**, che contenga **informazioni** e che, possibilmente, sia **machine processsable**
- Il modello FSOD cattura solo un aspetto della qualità del dato (la sua processabilità in modo automatico)

- Come analizzarla?
  - Con un set di **metriche** applicabili, oggettive, fattibili, automatizzabili, che considerino i metadati e che analizzino il dato in tutti i suoi aspetti

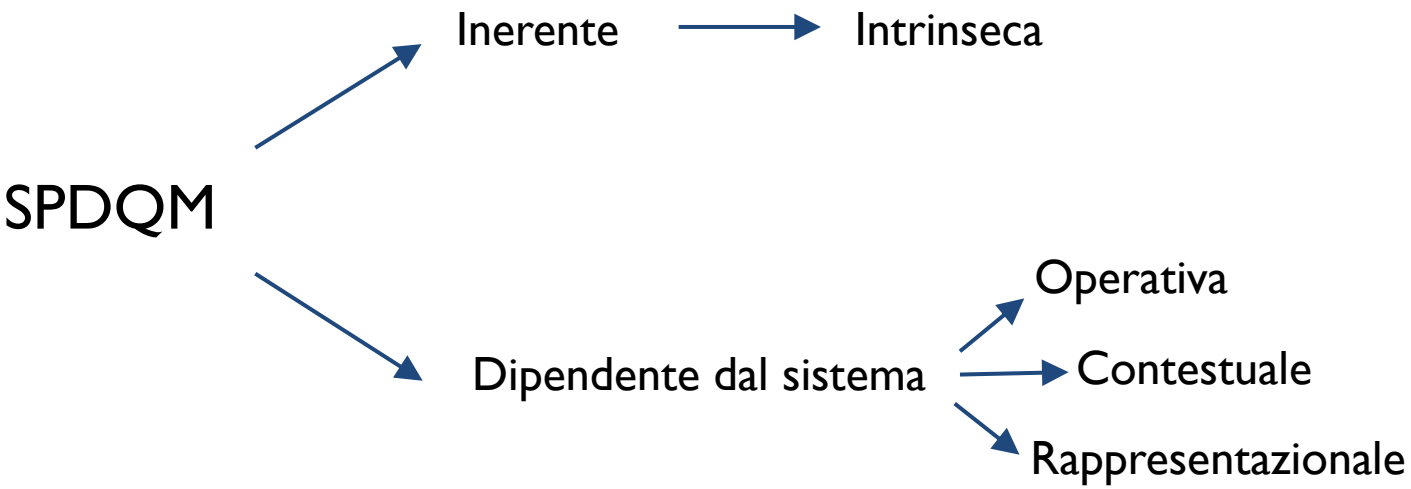
Modelli per la valutazione della qualità dei dati a confronto

5 criteri:

1. Applicabilità
2. Oggettività
3. Fattibilità
4. Automatizzabilità
5. Valutazione dei metadati

	# intrisec	score	Score medio	metriche condivise con SPDQM	% metriche condivise con SPDQM
SPDQM	12	45	3,75	12	100
CDQ	5	18	3,6	3	60
DQA	4	16	4	3	75
TDQM	4	16	4	2	50
QAFD	4	16	4	2	50
TIQM	5	15	3	2	40
DWQ	5	15	3	0	0
COLDQ	3	12	4	3	100
DaQuinCIS	3	12	4	3	100
ISTAT	2	8	4	2	100
AMEQ	2	8	4	2	100
AIMQ	5	8	1,6	1	20
IQM	3	6	2	0	0
CIHI	4	4	1	0	0

Modello per valutare la qualità: SPDQM

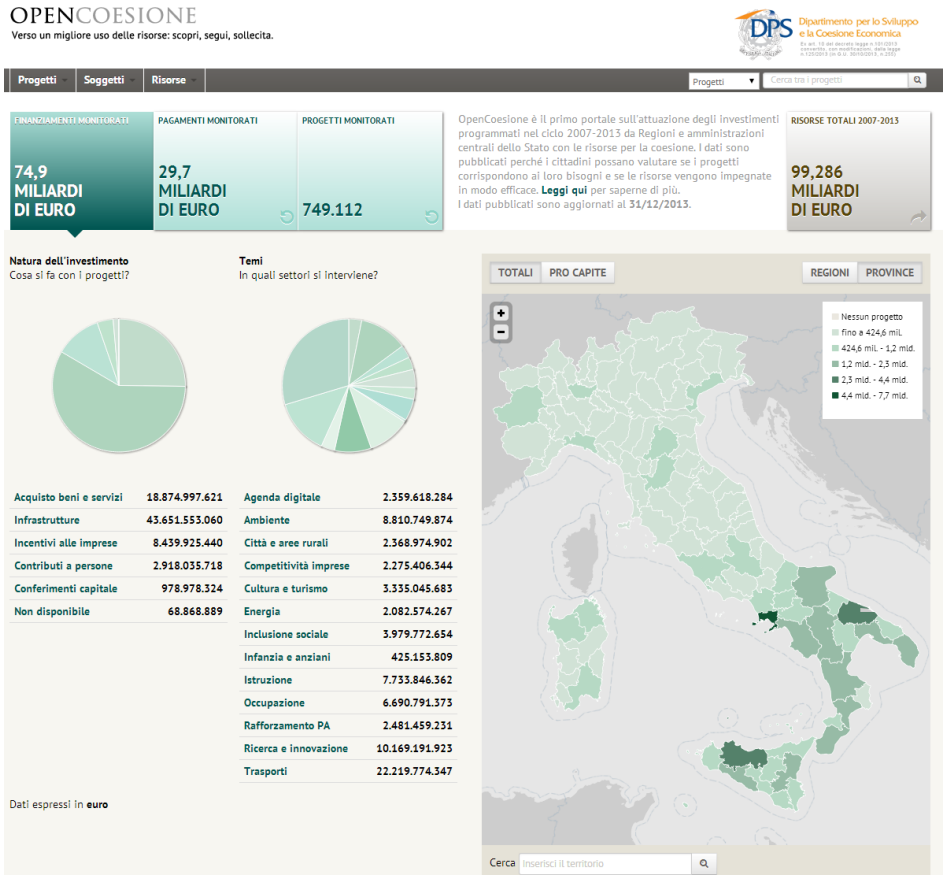


Modello inizialmente creato per valutare la qualità delle informazioni nei portali web.

Dimensione	Metrica	Sigla
Accuratezza	Percentuale celle corrette	pcvc
	Correttezza delle aggregazioni	ea
Completezza	Percentuale celle complete	pcc
	Percentuale righe complete	pcrp
Tracciabilità	Traccia creazione	tc
	Traccia modifiche	tam
Attualità	Percentuale righe correnti	prc
	Ritardo di pubblicazione	rp
Scadenza	Data scadenza definita	ds
	Ritardo dalla scadenza	rds
Standardizzazione	Colonne aderenti a uno standard	pcs
	Egms compliance	egmsc
	Five star open data	fsod
Comprensibilità	Colonne con metadati	pcm
	Colonne in formato comprensibile	pcfc

Caso di studio: Open Coesione

- Portale che pubblica i dati dei progetti finanziati tramite i **Fondi Europei per la Coesione 2007-2013**
  - Interfaccia Interattiva
  - Dataset .csv scaricabili
- 75 miliardi di Euro tracciati, 750K progetti
- OpenCoesione è la best practice in Italia
- 4° posto agli Open Government Awards del 2014 (assegnati a settembre)



monitoring marathon

Maratona di monitoraggio civico delle politiche pubbliche in Italia basata sui dati di OPENCOESIONE

Scarica il toolkit

HOME | COME FUNZIONA MONITHON | I REPORT DI MONITORAGGIO | BLOG | CREDITS

Invia il tuo report

Partecipa alla discussione sulla nostra **Mailing List**! E' open e aperta a tutti. Oppure scrivici a redazione@monithon.it

I pallini colorati nella mappa rappresentano i report di monitoraggio gia' inviati, su vari progetti tracciati sul portale pubblico **OpenCoesione**.

Per creare un nuovo report:

1. **Registrati o accedi** a Monithon.it
2. **Clicca su una delle seguenti tipologie di progetto** su cui sono gia' avviate collaborazioni con associazioni (es. Libera, Action Aid) o amministrazioni (@PONREC):

RIUSO DEI BENI CONFISCATI ALLE MAFIE

INNOVAZIONE SOCIALE MIUR

PROGETTI PER LA RICOSTRUZIONE IN EMILIA ROMAGNA

I PROGETTI "BLOCCATI" CITATI DA RIZZO E STELLA

POTENZIAMENTO STRUTTURALE DELLE UNIVERSITÀ - MIUR @PONREC

3. Sulla mappa appariranno nuovi progetti. Clicca sul progetto poi su "Crea un report!"

...OPPURE...

Scegli un progetto su **OpenCoesione** e **crea un report partendo da zero** su qualasiasi progetto finanziato con fondi pubblici (ricordati di copiare nel report il titolo e il codice del progetto)

La qualità dei dati in OpenCoesione: esempi di problematiche

Problema semantico

Errore?

CUP: F85C09006320009

ANTICRISI POLITICHE ATTIVE-MOBILITA'

Data di aggiornamento: ~~31.12.2013~~

ACQUISTO BENI E SERVIZI - CORSI DI FORMAZIONE
INFANZIA E ANZIANI - MISURE PER MIGLIORARE L'ACCESSO ALL'OCCUPAZIONE E AUMENTARE LA PARTECIPAZIONE SOSTENIBILE E IL PROGRESSO DELLE DONNE'

SEI UN PROTAGONISTA DI QUESTO PROGETTO? RACCONTACELO QUI.

SOGGETTI

PROGRAMMATORE
trasversale srl

ATTUATORE
trasversale srl

TEMPI

INIZIO PREVISTO
25 ottobre 2012

INIZIO EFFETTIVO
Dato non disponibile

FINE PREVISTA
30 giugno 2013

FINE EFFETTIVA
Dato non disponibile

PRIORITÀ QSN
Competitività dei sistemi produttivi e occupazione

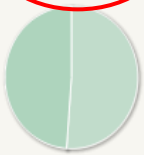
OBIETTIVO GENERALE QSN
Qualificare e finalizzare in termini di occupabilità e adattabilità gli interventi e i servizi di politica attiva del lavoro, collegandoli alle prospettive di sviluppo del territorio

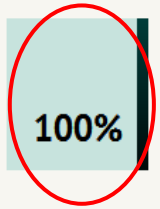
TERRITORIO ROMA

RISORSE PUBBLICHE

FINANZIAMENTO
45,00 euro

PAGAMENTI EFFETTUATI
45,00 euro





Unione europea	22 euro
Fondo di Rotazione (Co-finanziamento nazionale)	21 euro
Regione	0 euro

Visualizza l'andamento dei pagamenti

FONDO STRUTTURALE EUROPEO (FSE)
Fondi Strutturali relativi alla programmazione 2007/2013

PROGRAMMA
POR CRO FSE LAZIO

ASSE
Occupabilità

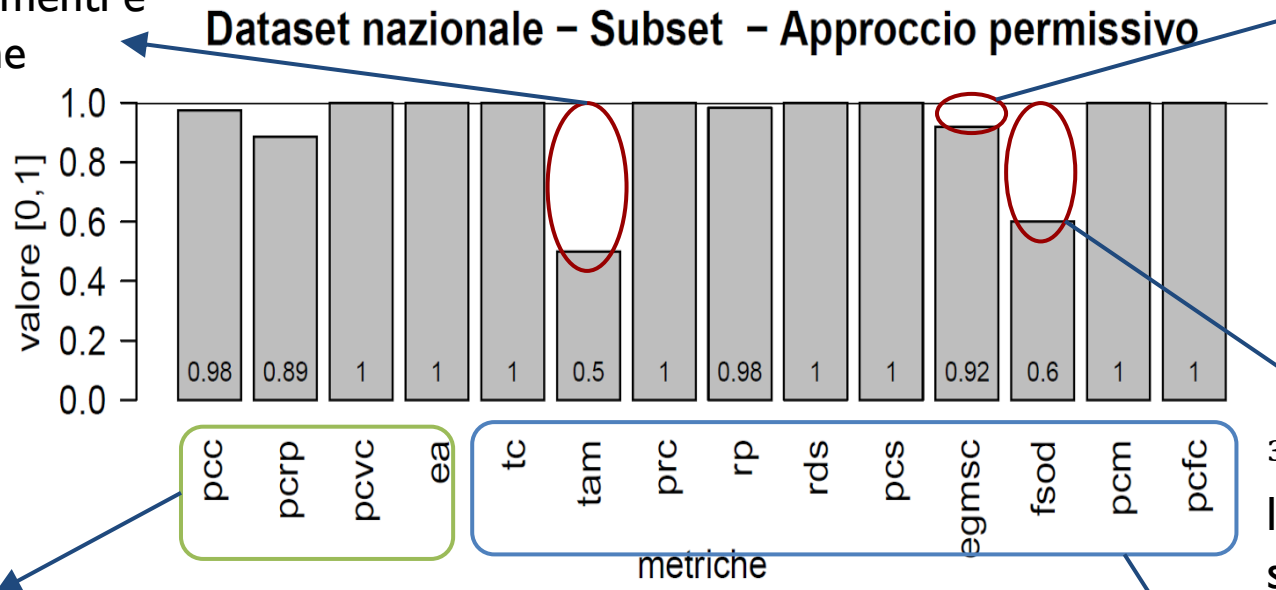
OBIETTIVO
Occupabilità - Aumentare l'efficienza, l'efficacia, la qualità e l'inclusività delle istituzioni del mercato del lavoro

43 !

Dati mancanti

Risultati Complessivi

manca traccia aggiornamenti e modifiche



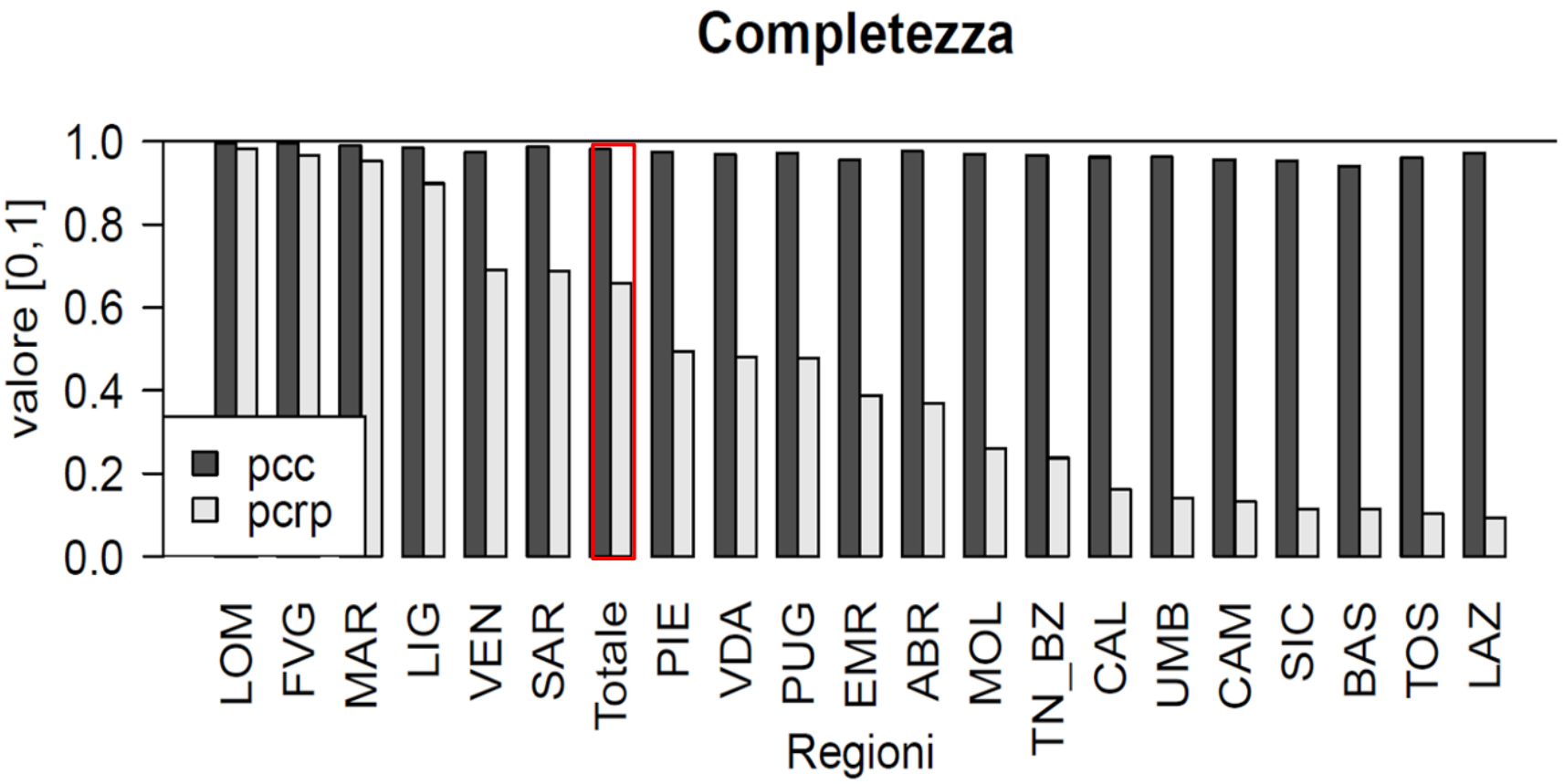
manca "pubblicatore" e "lingua" nei metadati

⅗ stelle, IRI e linked data per più stelle

Problema dei valori NA e Zero nei finanziamenti

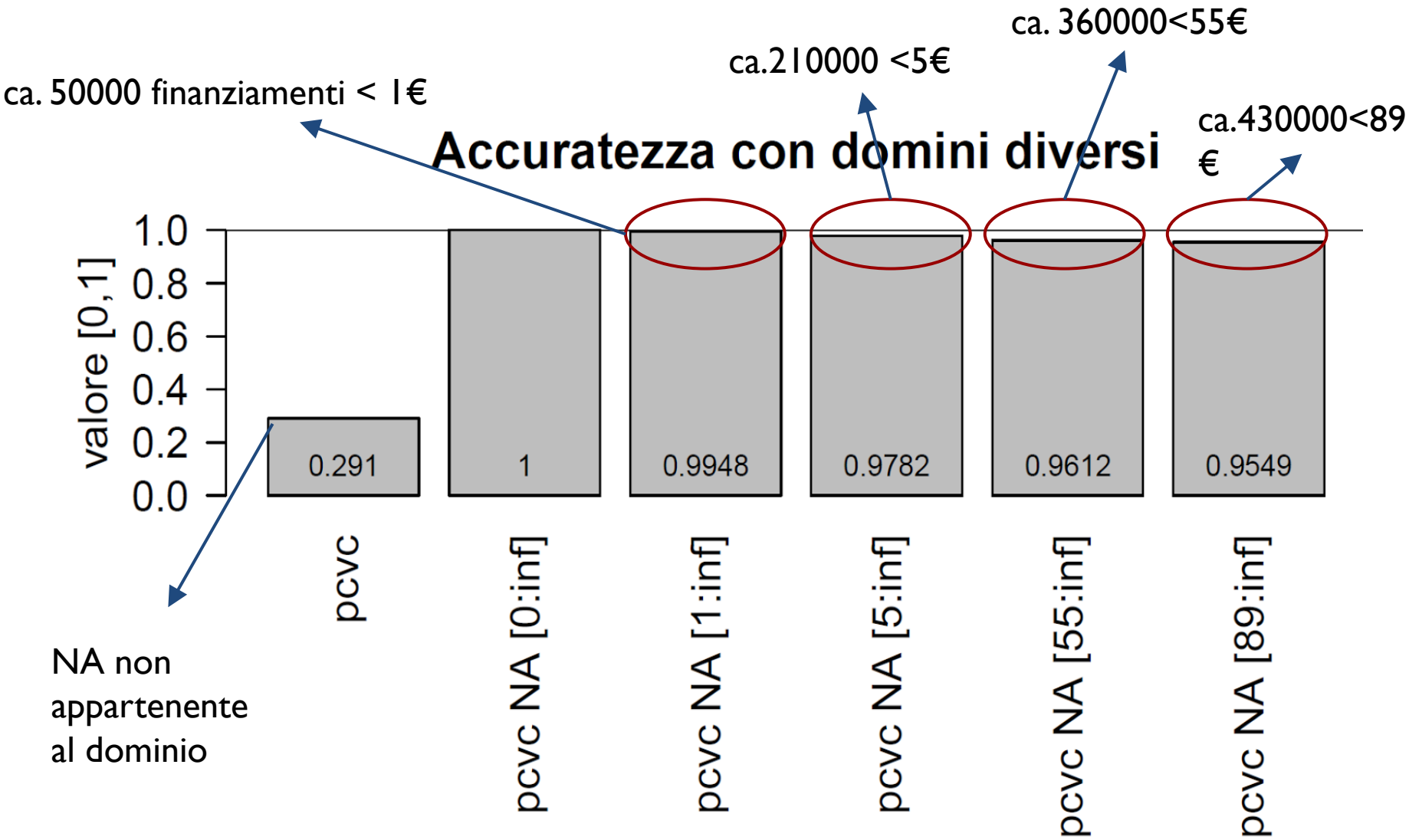
metriche a livello di dataset, non cambiano

Completezza – breakdown Regionale



- Risultati condizionati da **mancanza date e descrizioni** (Ateco)
- Sono stati considerati come errori di completezza i dati per cui non era ammessa l'assenza nei metadati

Accuratezza – Dominio e outliers



per studi futuri..

catturato da accuratezza

per studi futuri..

CUP: F85C09006320009

ANTICRISI POLITICHE ATTIVE-MOBILITA'

Data di aggiornamento: **31.12.2013**

ACQUISTO BENI E SERVIZI - CORSI DI FORMAZIONE
INFANZIA E ANZIANI - MISURE PER MIGLIORARE L'ACCESSO ALL'OCCUPAZIONE E AUMENTARE LA PARTECIPAZIONE SOSTENIBILE E IL PROGRESSO DELLE DONNE'

SEI UN PROTAGONISTA DI QUESTO PROGETTO? RACCONTACELO QUI.

SOGGETTI

PROGRAMMATORE
trasversale srl

ATTUATORE
trasversale srl

TEMPI

INIZIO PREVISTO
25 ottobre 2012

INIZIO EFFETTIVO
Dato non disponibile

FINE PREVISTA
30 giugno 2013

FINE EFFETTIVA
Dato non disponibile

PRIORITÀ QSN

Competitività dei sistemi produttivi e occupazione

OBIETTIVO GENERALE QSN

Qualificare e finalizzare in termini di occupabilità e adattabilità gli interventi e i servizi di politica attiva del lavoro, collegandoli alle prospettive di sviluppo del territorio

TERRITORIO ROMA

RISORSE PUBBLICHE

FINANZIAMENTO
45,00 euro

Unione europea	22 euro
Fondo di Rotazione (Co-finanziamento nazionale)	21 euro
Regione	0 euro

PAGAMENTI EFFETTUATI
45,00 euro

Visualizza l'andamento dei pagamenti

FONDO STRUTTURALE EUROPEO (FSE)
Fondi Strutturali relativi alla programmazione 2007/2013

PROGRAMMA
POR CRO FSE LAZIO

ASSE
Occupabilità

OBIETTIVO
Occupabilità - Aumentare l'efficienza, l'efficacia, la qualità e l'inclusività delle istituzioni del mercato del lavoro

catturato da completezza

Suggerimenti consegnati al Ministero (MEF)

- Descrivere meglio il dominio nei metadati (null o zero?)
- Aggiungere traccia modifiche
- Aggiungere pubblicatore e lingua nei metadati
- IRI + linked per 5 Star (cfr. Linee Guida sez. 4.1).
- Metadati associati al dato (cfr. Linee Guida sez. 4.2)

Possibilità di studi futuri:

- Euristiche per la definizione della chiusura del progetto
- Analisi sulla categorizzazione dei progetti

Conclusioni

- Verifica fattuale (errori, formato file, quantità di informazione):
  - I dati come sono pubblicati ora dalle PA sono molto difficili da riutilizzare. Sarebbe necessaria più standardizzazione e più apertura nei formati.
  - Utile come primo controllo sullo stato dei dati pubblicati
- Uso di un modello formale (SPDQM):
  - Possibile in modo automatico su dati già pubblicati secondo determinati standard (machine processable, schema standard)
  - Efficace: utile introspezione sulla qualità del dato, cattura diverse problematiche.

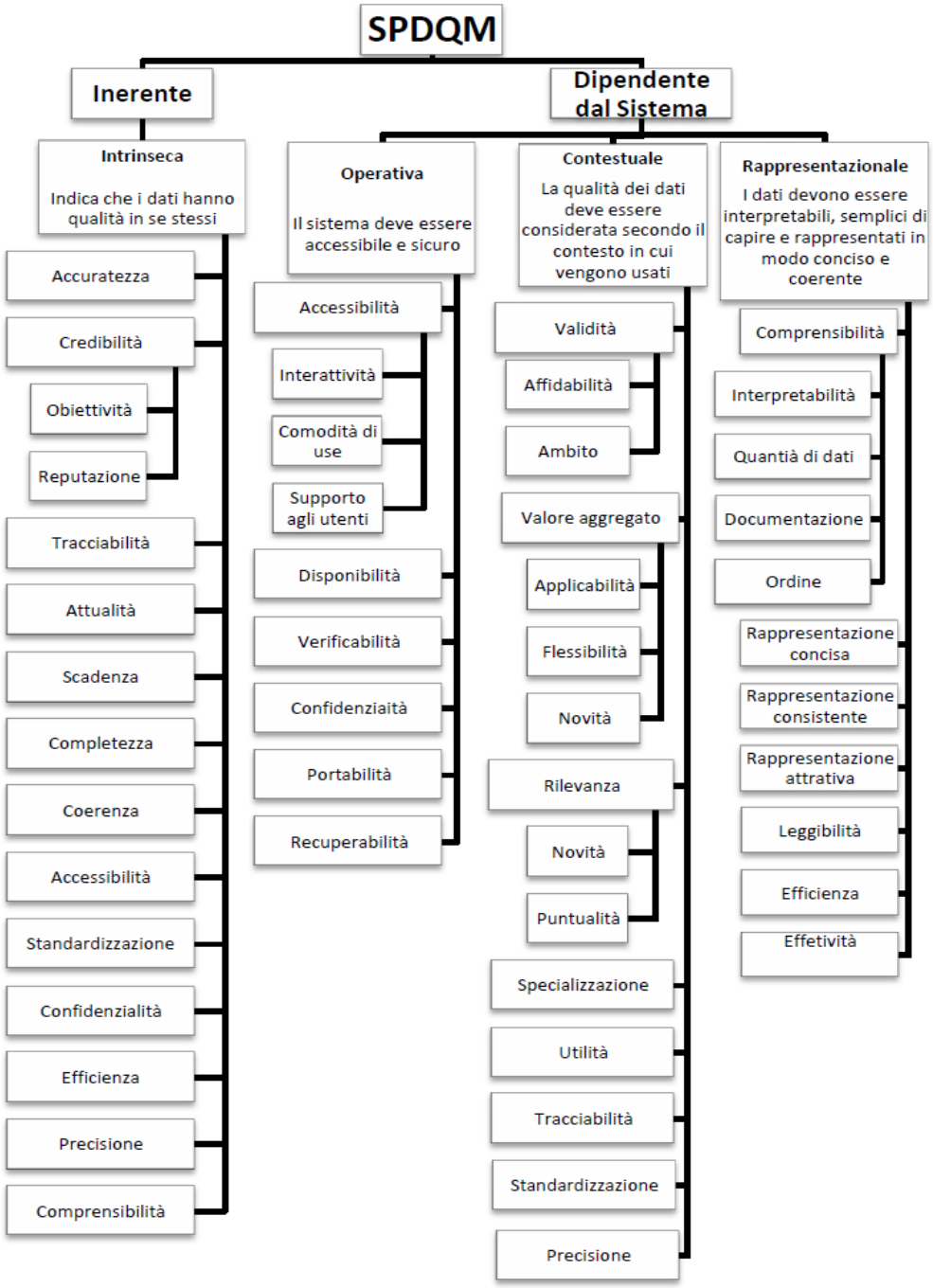
Grazie per l'attenzione!

Master in Ingegneria dei dati

Torino, 9 ottobre 2014

Albero SPDQM

SPDQM = Categorie definite in PDQM + punti di vista definiti nello standard ISO/IEC 25012 (SQuaRE)



Accuratezza

- Nel caso degli Open Data è praticamente impossibile possedere dei valori reali con il quale verificare l'accuratezza del dato
  - Es: http://opencoesione.gov.it/progetti/3pipban-13586/
- Si calcola verificando l'appartenenza al dominio di un determinato valore (accuratezza sintattica, Batini 2006)
- Disponibili in rete programmi che tramite algoritmi deduttivi scovano gli errori di accuratezza quali typo e duplicazioni di entrate. Es: Open Refine

- Formula: *correttezza = 1- (numero tot errori / numero totale di dati)*

CUP: J15C13006730007

(0254000072) PRAGMATICA DEGLI OPEN DATA NELLA PA - OP4PA

Data di aggiornamento: **30.06.2014**

INCENTIVI ALLE IMPRESE - ATTIVITA' DI RICERCA
RICERCA E INNOVAZIONE - ASSISTENZA ALLA RST, IN PARTICOLARE NELLE PMI (INCLUSO L'ACCESSO AI SERVIZI DI RST NEI CENTRI DI RICERCA)

SEI UN PROTAGONISTA DI QUESTO PROGETTO? RACCONTACELO QUI.

SOGGETTI

PROGRAMMATORE
REGIONE PIEMONTE

ATTUATORE
COSVIFOR SRL

TEMPI

INIZIO PREVISTO
01 ottobre 2013

INIZIO EFFETTIVO
01 ottobre 2013

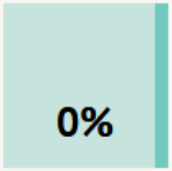
FINE PREVISTA
01 ottobre 2013

FINE EFFETTIVA
01 ottobre 2013

RISORSE PUBBLICHE

FINANZIAMENTO
5.100,00 euro

PAGAMENTI EFFETTUATI
0,00 euro



Unione europea	**2.018** euro
Fondo di Rotazione (Co-finanziamento nazionale)	**2.371** euro
Regione	**709** euro

FONDO EUROPEO DI SVILUPPO REGIONALE (FESR)
Fondi Strutturali relativi alla programmazione 2007/2013

PROGRAMMA
POR CRO FESR PIEMONTE

ASSE
Innovazione e transizione produttiva

OBIETTIVO

Completezza

- Nel caso degli Open Data bisogna assumere che solo i dati presenti nel dataset e nessun altro valore rappresenti i fatti nel mondo reale (assunzione del mondo chiuso, manca solitamente una tabella di riferimento).
- I valori nulli, se non diversamente specificato nei metadati sono considerati come valori mancanti.
- Vengono calcolate il numero di celle mancanti e il numero di righe non complete (quanta informazione fornisce una tupla rispetto al suo massimo potenziale informativo?)

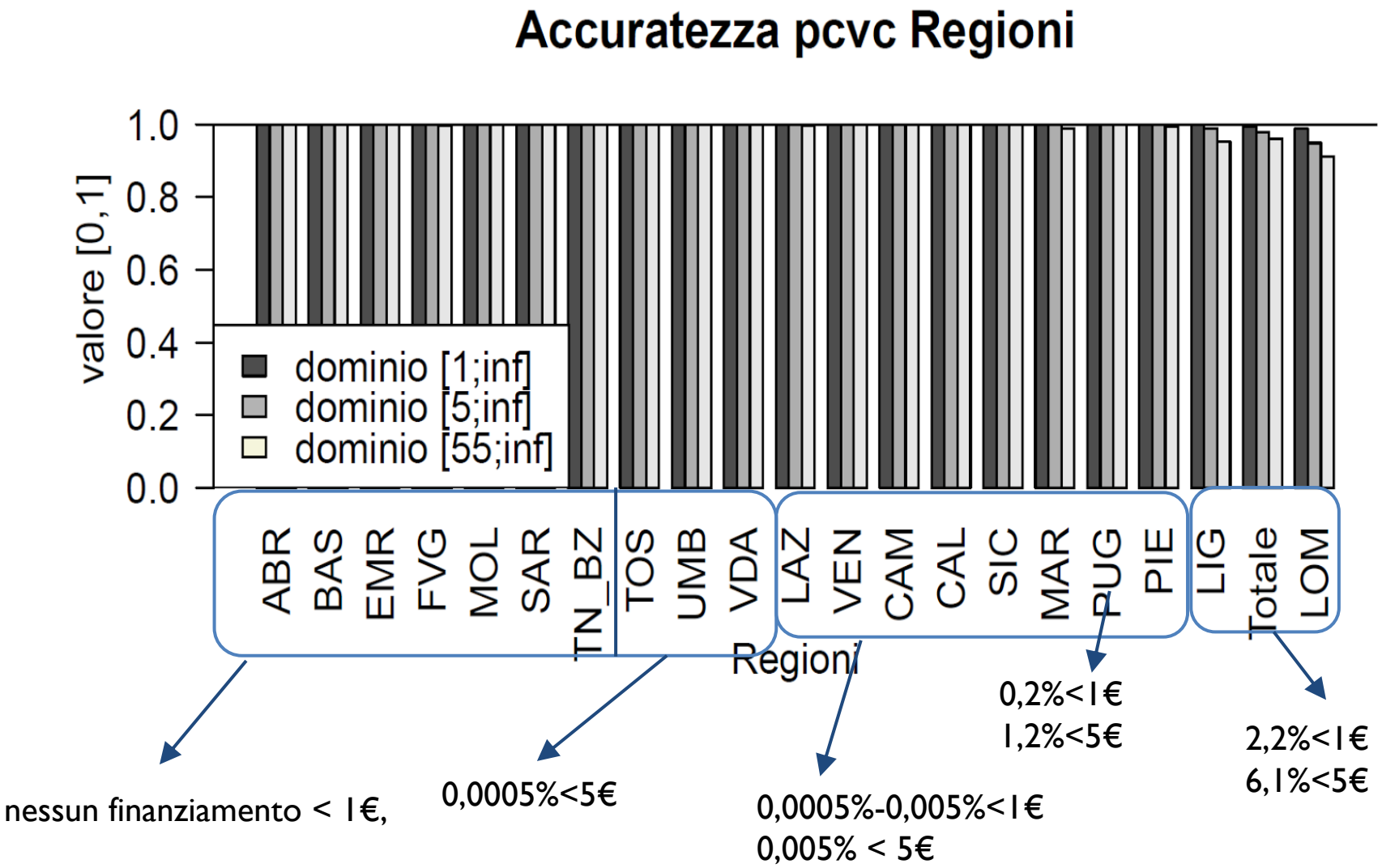
- Formula: *completezza* = $Q_{\mathrm{c}}(\mathrm{record}) = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}[field_{\mathrm{i}}=\mathrm{null}]}{\mathrm{n}}$

Dimensioni legate al tempo: Attualità, Scadenza

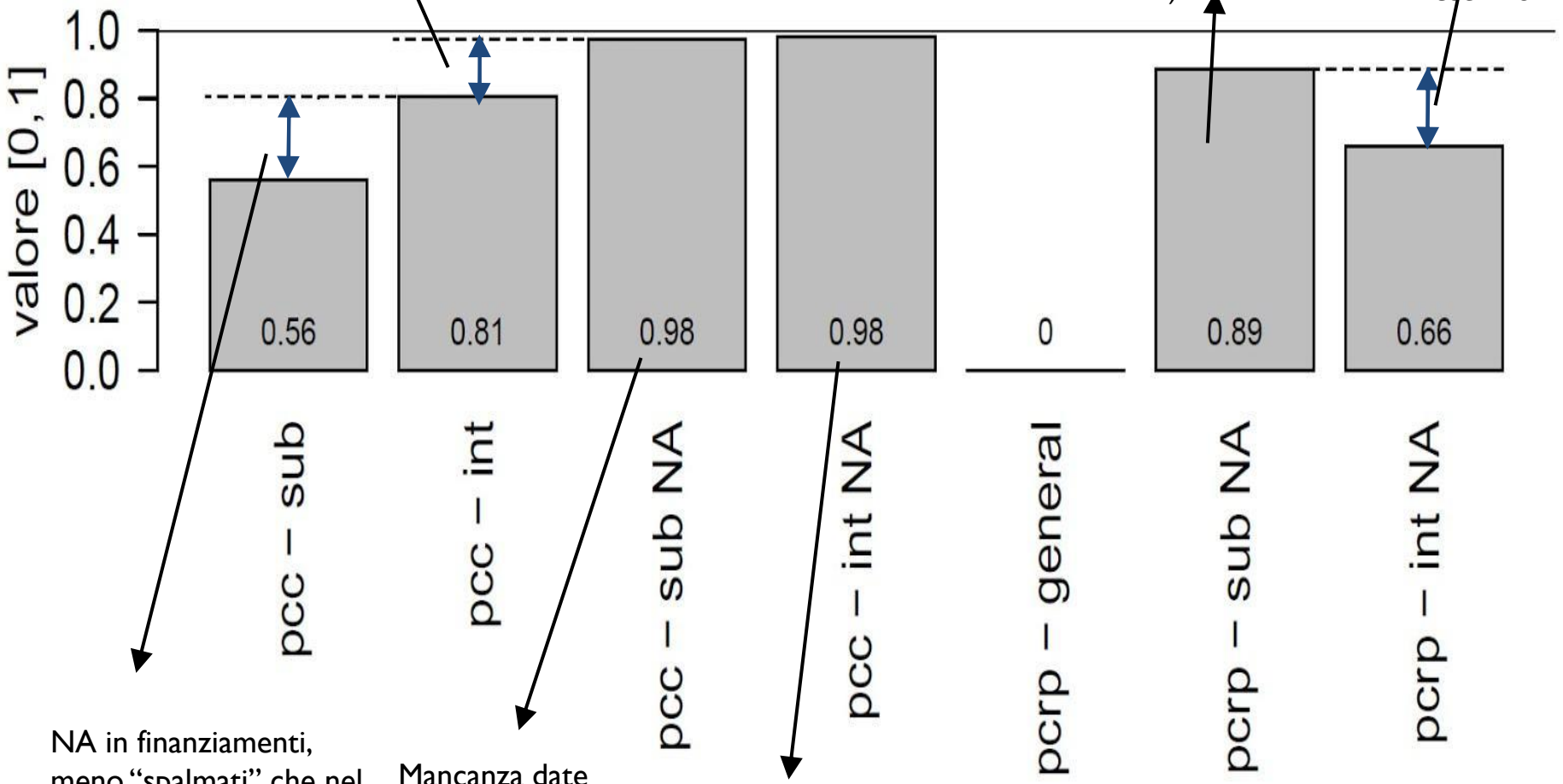
- Attualità
  - Percentuale di righe correnti
    - Procedura: (i) definire il periodo di validità del dataset, (ii) se ci sono attributi con date queste colonne sono da controllare, (iii) percorre tutte le righe del dataset contando quelle che hanno almeno un attributo con valore non corrente (iv) calcolare la percentuale di righe correnti sul totale delle righe
  - Ritardo di pubblicazione
    - Quanto tempo passa dalla disponibilità dei dati alla loro pubblicazione (rapportato al periodo di tempo a cui si riferisce il dataset). Es: orario lezioni semestre o orario di una conferenza (di 1 gg)
- Scadenza
  - Data di scadenza definita: verifica se la data di scadenza è stata definita
  - Ritardo dalla scadenza: se la scadenza è definita, quanto tempo passa prima che una nuova versione del dataset sia pubblicata

Dimensioni relative ai metadati: tracciabilità, comprensibilità, standardizzazione

- Tracciabilità
  - Da quale ente proviene una certa informazione? Chi gestisce i dati?
    - Traccia di creazione del file
    - Traccia delle modifiche – è disponibile uno storico delle modifiche?
- Comprensibilità
  - Per comprendere il significato degli attributi (descrizione, insieme dei valori ammissibili, unità di misura)
    - Percentuale di colonne con metadati
    - Percentuale di colonne in formato comprensibile: in base al formato di rappresentazione definire se le colonne sono comprensibili
- Standardizzazione
  - Percentuale di colonne aderenti a uno standard (su quelle aderibili)
  - E-gms compliance: verifica se i metadati aderiscono al set di metadati definiti dall' e-government Metadata Standard (Sorgente, data di creazione, categoria, titolo, descrizione, identificatore, pubblicatore, copertura, lingua)
  - FSOD: verifica che standard di formato viene seguito, è adatto al paradigma Open Data?


Accuratezza pcvc Regioni
valore [0,1]
1.0
0.8
0.6
0.4
0.2
0.0
dominio [1;inf]
dominio [5;inf]
dominio [55;inf]
ABR
BAS
EMR
FVG
MOL
SAR
TN_BZ
TOS
UMB
VDA
LAZ
VEN
CAM
CAL
SIC
MAR
PUG
PIE
LIG
Totale
LOM
Regioni
nessun finanziamento < 1€,
0,0005%<5€
0,0005%-0,005%<1€
0,005% < 5€
0,2%<1€
1,2%<5€
2,2%<1€
6,1%<5€

Completezza



NA in finanziamenti e pagamenti, valori mancanti per descrizioni Ateco e altre descrizioni

Nell'89% dei progetti sono presenti le date (controllo non effettuato su "fine effettiva", vista la sua natura)

Mancanza codici e descrizioni Ateco e altre descrizioni

NA in finanziamenti, meno "spalmati" che nel dataset intero

Mancanza date

Mancanza date e codici/descrizioni Ateco

sub: analisi sui 22 attributi
int: analisi su intero dataset
general: stesso risultato sia in sub che int
NA: i valori NA appartengono al dominio