Data Quality
Standard e Applicazioni

18 Aprile 2018 – DAUIN – Politecnico di Torino

Marco Torchiano

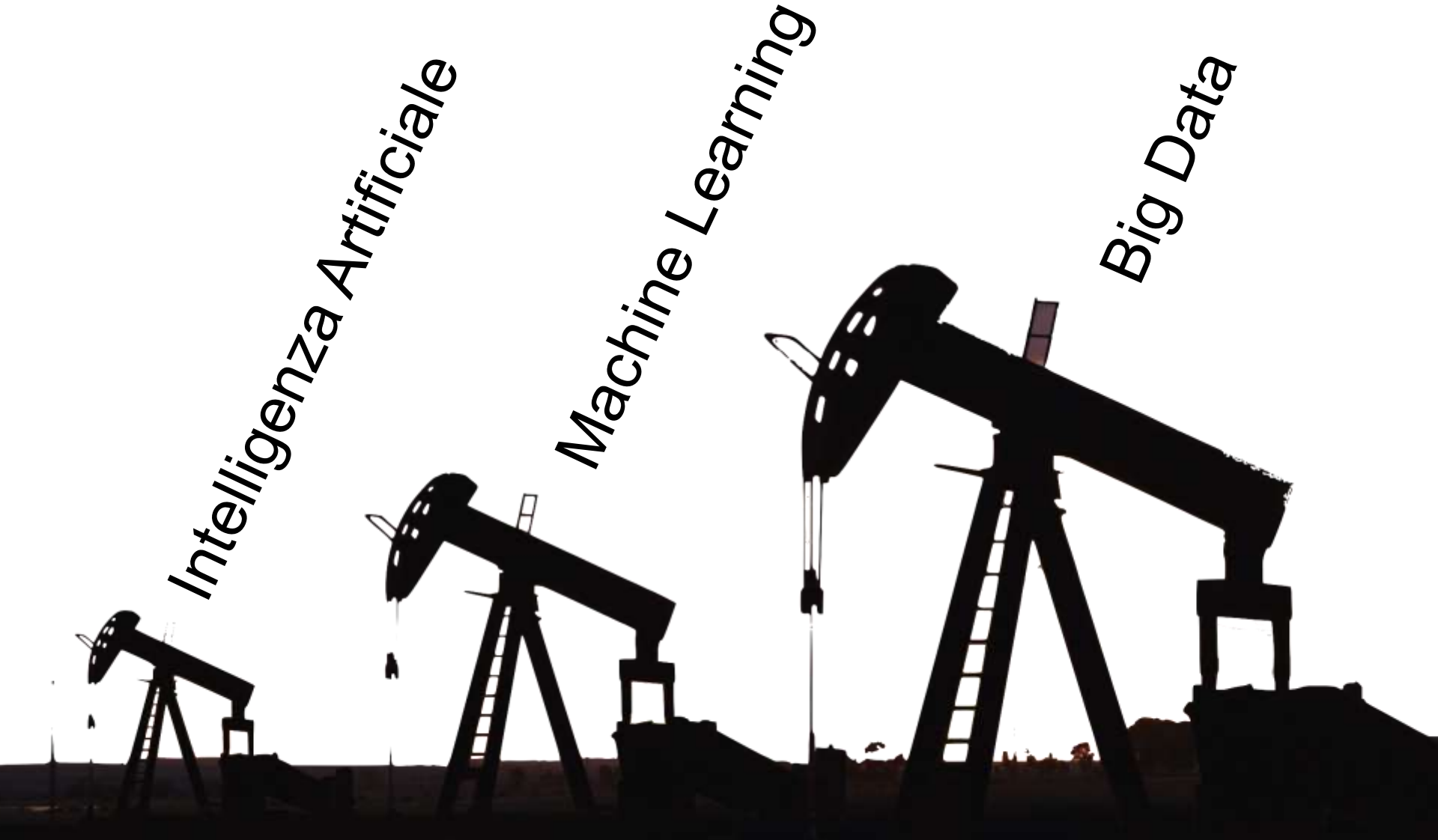
marco.torchiano@polito.it

SoftEng
http://softeng.polito.it

Nexa Center
for Internet & Society

PERCHÈ LA QUALITÀ DEI DATI È IMPORTANTE?

Data is The New Oil!



Garbage-In-Garbage-Out

From xkcd: https://xkcd.com/1838/

Pubblica Amministrazione

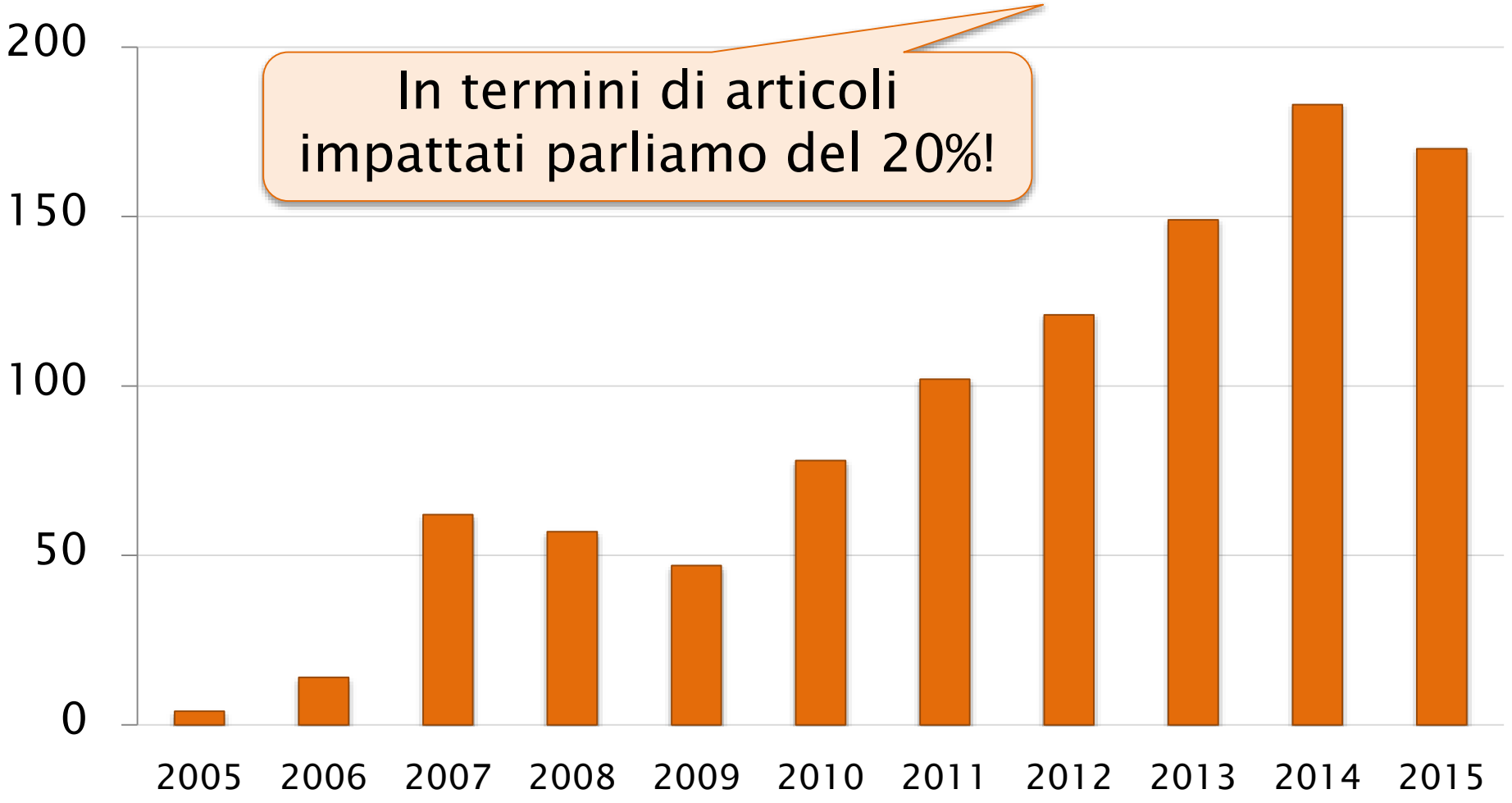
- Decreto Trasparenza (14/3/2013 n.33)
  - Contratti Pubblici (Art.37 & Art 9.)
  - Formato Standard XML (ANAC)
- Osservando un ateneo "cugino"…
  - CIG è uguale a "0000000" nel 6% dei casi
  - Codice Fiscale manca nel 3% dei contratti
  - Pagato più del dovuto nel 3% dei casi

Dati del 2014

Per i più curiosi , i dati "linked" e ripuliti sono su **`https://contrattipubblici.org`** by synapta

Ricerca

File dati errati da articoli di Genomica



Ziemann et al. Genome Biology, (2016):17(1)

Accademia

- Noi accademici siamo valutati sulla base di numeri
  - Di pubblicazioni prodotte e
  - Di citazioni ricevute per le stesse

- (Quasi) nessuna commissione legge in dettaglio le nostre pubblicazioni

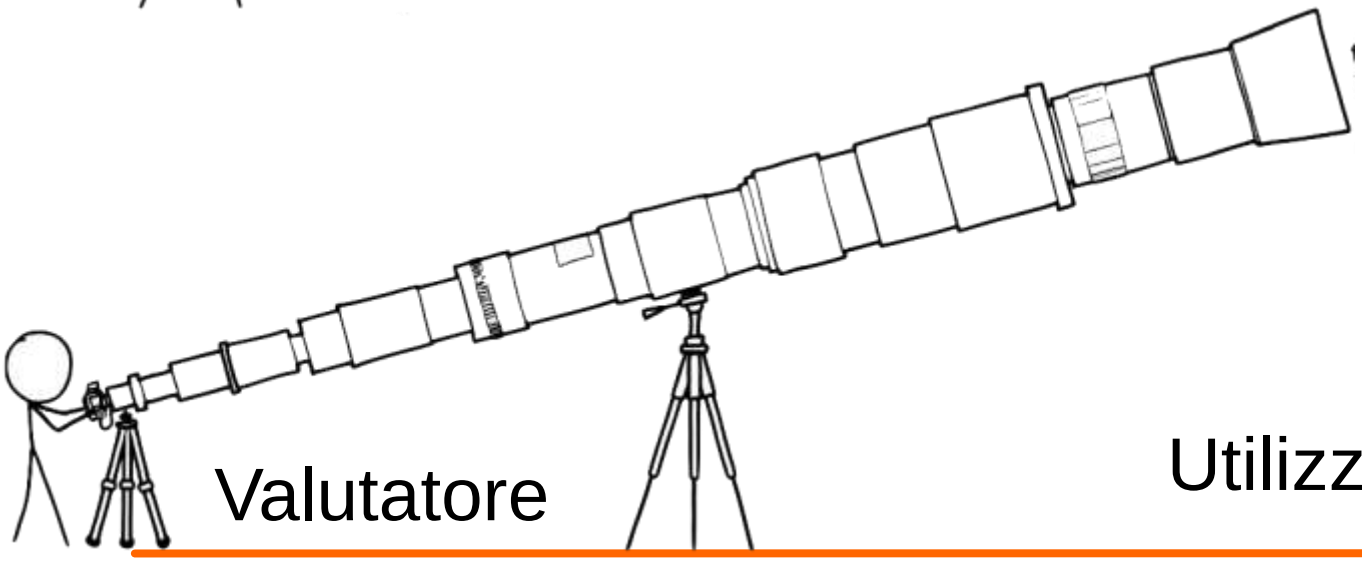
Cosa è La Qualità dei Dati?

Dipende… dal punto di vista

Produttore

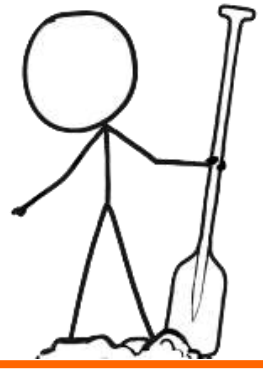
Acquirente

Integratore



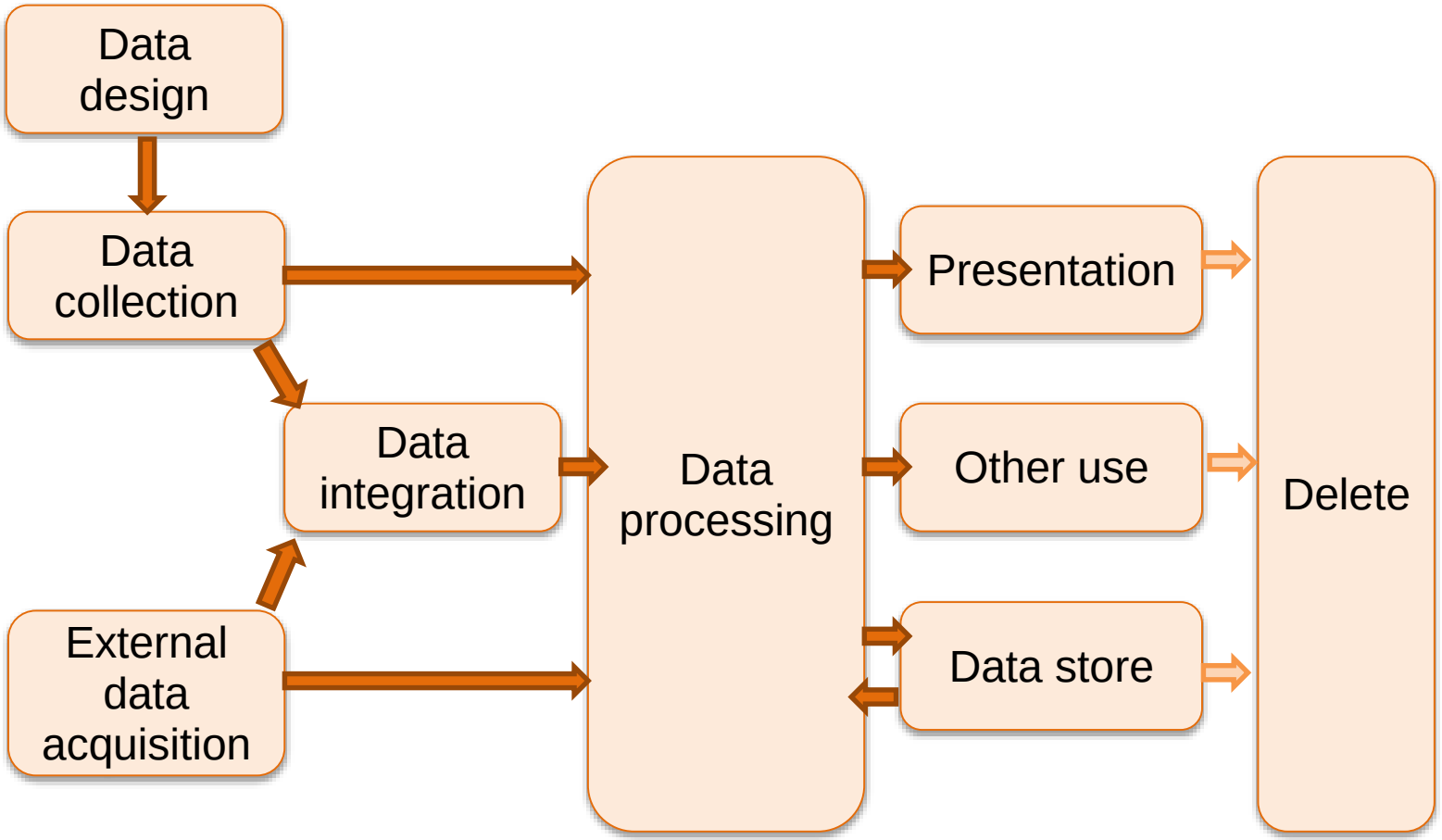
Valutatore

Utilizzatore



Dipende…

…dalla fase del Ciclo Di Vita (CDV)



Dipende…

- …dal tipo di caratteristiche
- Inerenti al dato
- Contingenti al sistema
  - Memorizzazione
  - Elaborazione
  - Presentazione

Come misurare la qualità dei dati?

STANDARD ISO SU QUALITÀ DEI DATI

Caratteristiche ISO-25012

Inerenti

- Accuratezza
- Completezza
- Coerenza
- Credibilità
- Attualità

- Accessibilità
- Conformità
- Riservatezza
- Efficienza
- Precisione
- Tracciabilità
- Comprensibilità

- Disponibilità
- Portabilità
- Ripristinabilità

Dipendenti dal sistema

Accuratezza – Misura ISO 25024

ID	Nome	Descrizione	Funzione di misurazione	CVD Entità target Proprietà
Acc-I-1	Accuratezza sintattica dei dati	Rapporto di vicinanza dei valori dei dati a un insieme di valori definiti in un dominio	X=A/B A= numero di dati elementari correlati a valori sintatticamente accurati B= numero di dati elementari per cui è richiesta la accuratezza sintattica	Tutto il CVD eccetto progettazione dei dati. File di dati. Dati elementari, valore dei dati.

NOTA 1 Un singolo valore è considerato "sintatticamente accurato" quando coincide con il valore di una fonte identificata di informazioni convalidate: il risultato è "sì" o "no".

NOTA 2 Un esempio di basso grado di accuratezza sintattica è quando la parola Mary è memorizzata come Marj.

Limitandoci a caratteristiche inerenti,

COME SI MISURA LA QUALITÀ DEI DATI IN PRATICA?

Mondo Chiuso o Aperto?

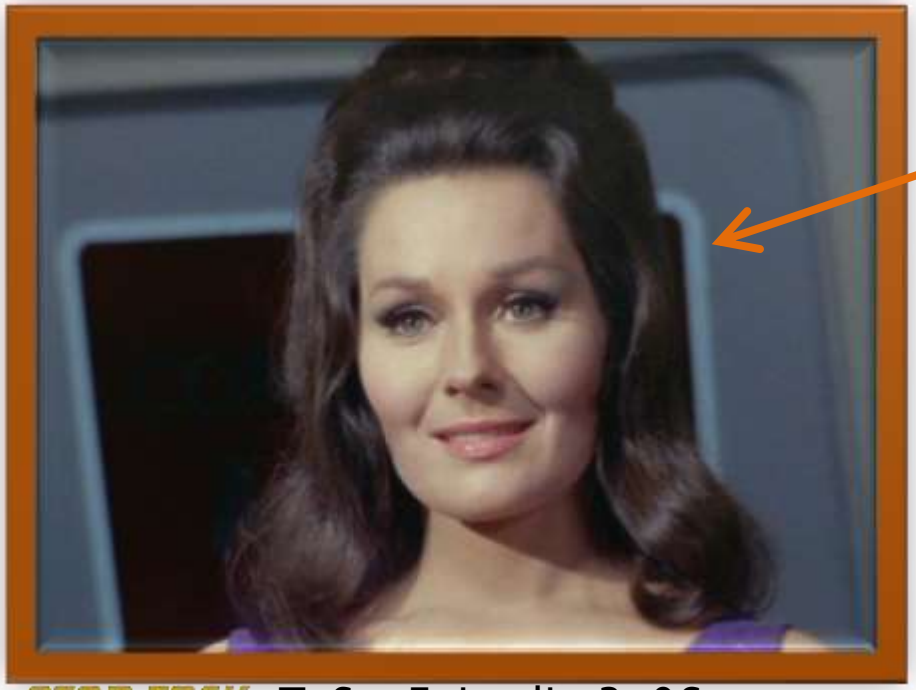
- Mondo Chiuso (CWA):
  - La conoscenza rappresentata nei dati (e nei loro schemi) è completa.
  - Es. se un nome compare nell'elenco dei nomi validi è corretto, altrimenti è errato.
- Mondo Aperto (OWA):
  - La conoscenza rappresentata nei dati è (consapevolmente) incompleta
  - Es. se un nome compare nell'elenco di quelli validi è corretto, altrimenti non è possibile decidere con certezza

CWA-Accuratezza: Genomica

- I geni umani noti sono codificati e ognuno ha un simbolo predefinito
- Qualunque codice non compreso in quelli definiti costituisce un errore di accuratezza sintattica
- Il codice ‘SEPT2’(Septin-2) quando viene importato in  si trasforma in ‘2 Febbraio’

CWA-Accuratezza: Nomi propri

*NOTA 2 Un esempio di basso grado di accuratezza sintattica è quando la parola **Mary** è memorizzata come **Marj**.*



Marj Dusay

STAR TREK ToS - Episodio 3x06

OWA – Accuratezza

Come decidere cosa è accurato?

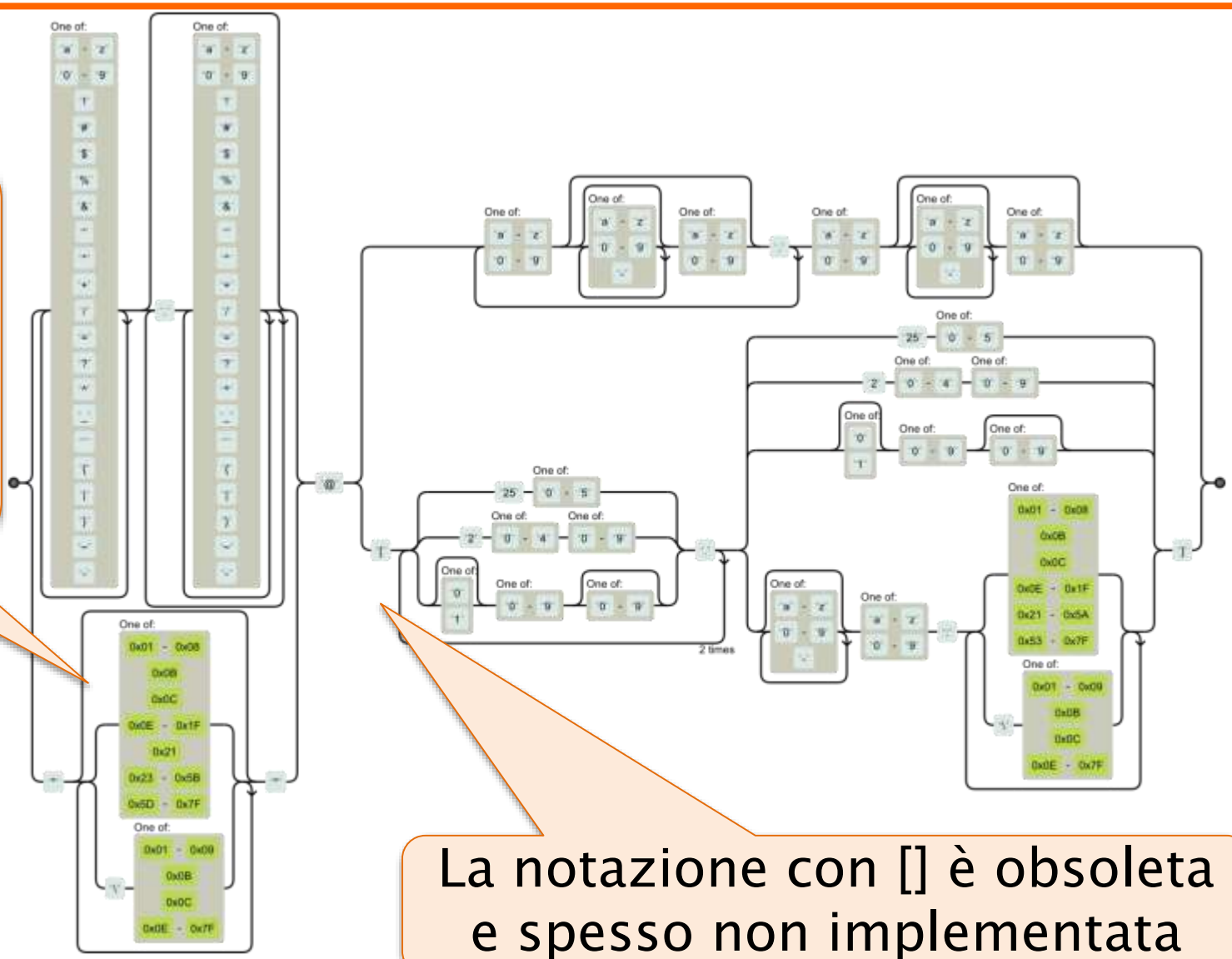
- Regole che indicano cosa è sintatticamente corretto
  - Es. Espressioni regolari
- Vincoli per indicare quali valori sono semanticamente accettabili
  - Es. Intervalli di validità

Da dove derivano le regole?

- Standard
- Conoscenza del Dominio
- Dati analoghi
- Dati passati

OWA: Email secondo RFC 5322

I caratteri non stampabili sono un problema per i client



La notazione con [] è obsoleta e spesso non implementata

OPENCOESIONE

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

Progetti | Soggetti | Risorse

Progetti | Cerca tra i progetti

VERSO LA PROGRAMMAZIONE 2014-2020

RISORSE | PROGRAMMI | BANDI | PROGETTI

FINANZIAMENTI MONITORATI (INCLUSE RISORSE ATTRATTE)
102,0 MILIARDI DI EURO

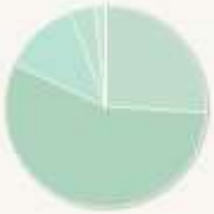
2014 OPEN GOVERNMENT AWARDS

OpenCoesione è il portale sull'attuazione dei progetti finanziati dalle politiche di coesione in Italia. Sono navigabili dati su risorse assegnate e spese, localizzazioni, ambiti tematici, soggetti programmatori e attuatori, tempi di realizzazione e pagamenti dei singoli progetti. Tutti possono così valutare come le risorse vengono utilizzate rispetto ai bisogni dei territori. I dati pubblicati sono aggiornati al 29/02/2016 e riguardano 103.790 soggetti

RISORSE TOTALI 2007-2013
99,286 MILIARDI DI EURO

NATURA DELL'INVESTIMENTO

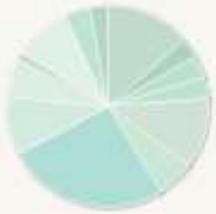
Cosa si fa con i progetti?



Acquisto beni e servizi	26.473.673.962
Infrastrutture	57.159.238.912
Incentivi alle imprese	12.571.185.885
Contributi a persone	4.084.443.619
Conferimenti capitale	1.712.313.939
Non disponibile	26.442.273

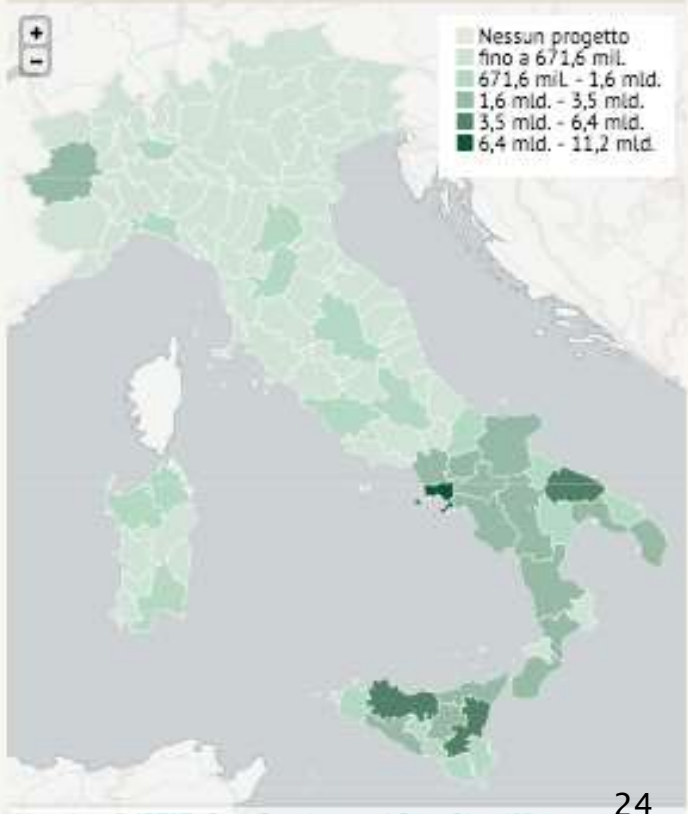
TEMI

In quali settori si interviene?



Ricerca e innovazione	13.792.637.465
Agenda digitale	3.343.061.119
Competitività imprese	3.617.632.179
Energia	3.430.288.657
Ambiente	11.562.174.366
Cultura e turismo	6.228.189.871
Trasporti	26.685.631.683
Occupazione	9.490.699.983
Inclusione sociale	6.520.601.319
Infanzia e anziani	684.938.953
Istruzione	9.500.368.579
Città e aree	4.184.287.627

TOTALI | PRO CAPITE | REGIONI | PROVINCE



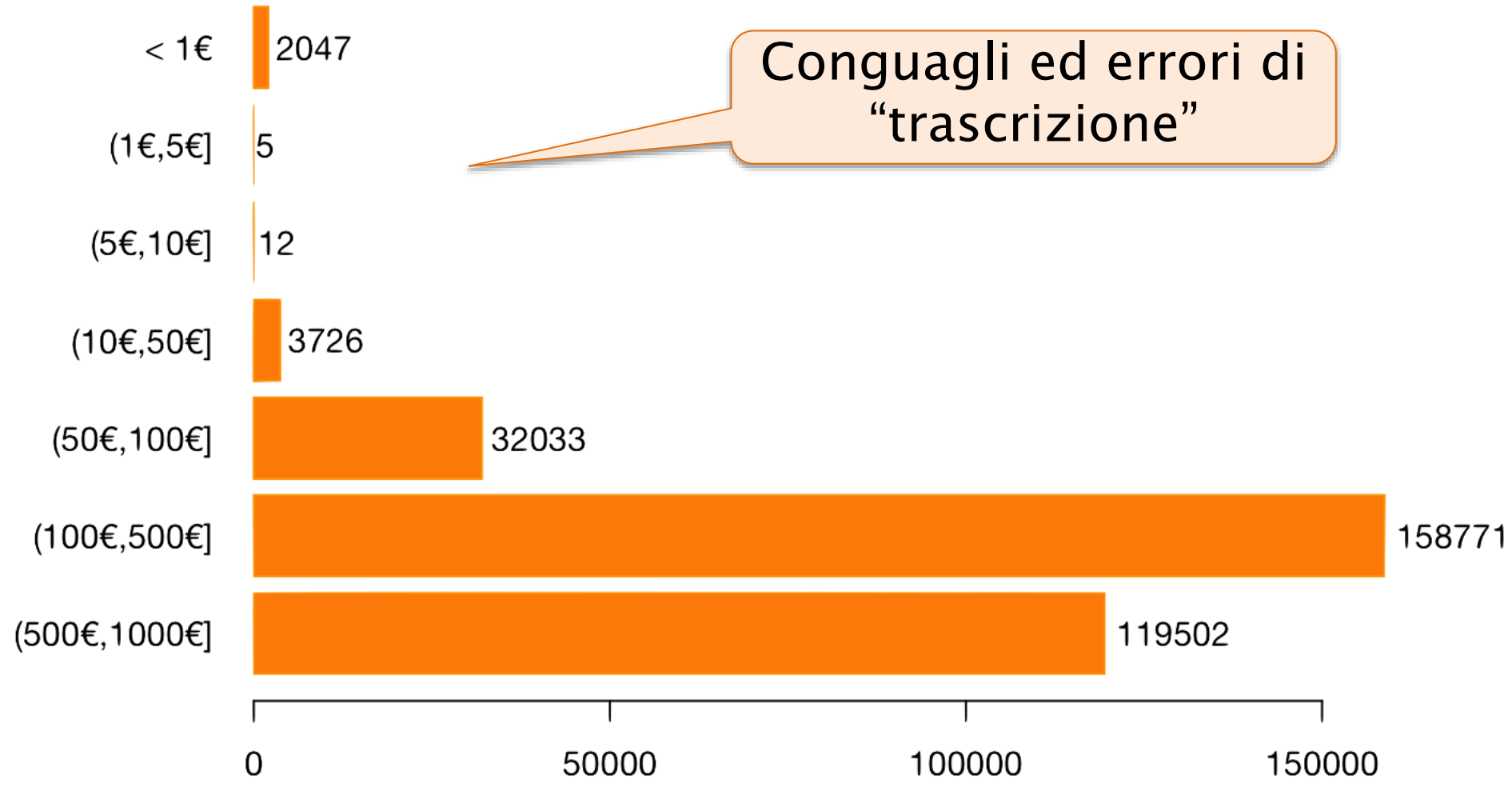
Map data © ISTAT, OpenCoesione and OpenStreetMap contributors

OWA: Finanziamenti Europei

- Open Coesione
  - Dati aperti liberamente accessibili
  - Descrizioni, Finanziamenti, Costi
- Dato: costi rendicontabili
- Come valutare l'accuratezza semantica?
  - Valori positivi
  - Valori almeno pari a…

Costi Rendicontabili

Distribuzione solo per progetti con costi fino a 1000€



DBPedia

- Ha lo scopo di estrarre informazioni da Wikipedia e pubblicarle su Web come Linked Open Data
- Base di conoscenza in formato RDF
  - Possiamo interrogare un *endpoint SPARQL* per ottenere informazioni strutturate
  - Es. sapere il tempo trascorso in orbita di tutti gli astronauti

Basi di Conoscenza

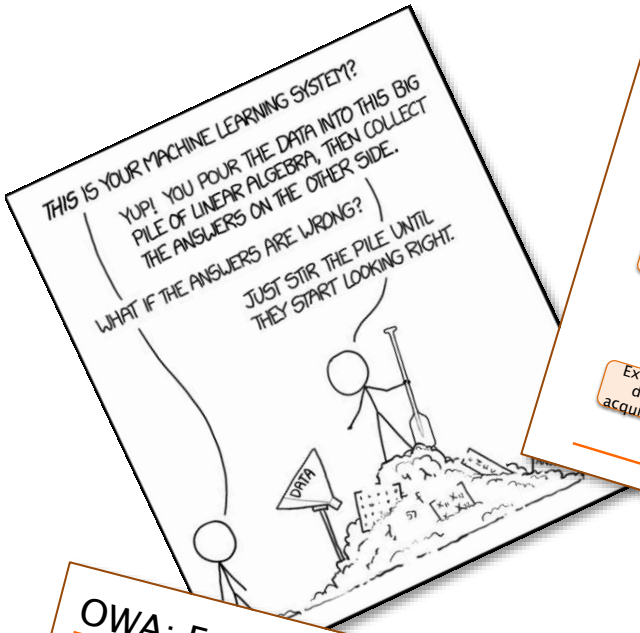
- Molte basi di conoscenza non sono le "*depositarie del sapere*"
- Spesso sono costruite estraendo dati da diverse fonti ed aggregandoli
- Sono soggette a continua evoluzione
  - Quando le fonti pubblicano nuovi dati devono essere estratti ed integrati
- L'evoluzione può essere la chiave per identificare problemi di qualità

CWA-Completezza: DBPedia

- Abdul Ahad Mohmand
  - Primo Afgano nello spazio
- DBPedia di Aprile 2016 non riporta il dato sul tempo nello spazio
- Tale informazione era presente nella versione di Ottobre 2015

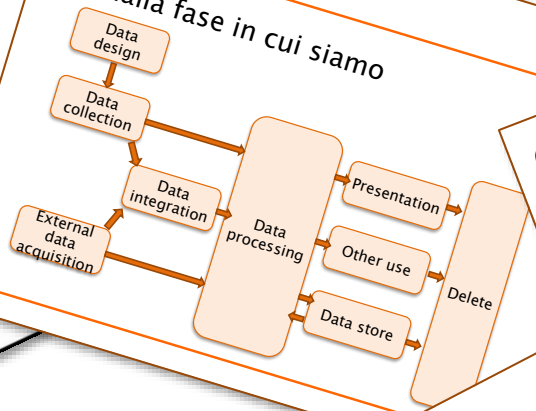
Abdul Ahad Momand	
Cosmonauta dell'Intercosmos	
Nazionalità	Afghanistan
Status	Ritirato
Data di nascita	1º gennaio 1959
Selezione	1992
Altre attività	Pilota
Tempo nello spazio	8 giorni, 20 ore e 26 minuti
Missioni	▪ Soyuz TM-5 ▪ Soyuz TM-6
Data ritiro	1988

In conclusione



Dipende...

...dalla fase in cui siamo



Caratteristiche ISO-25012

Inerenti

- Accuratezza
- Completezza
- Coerenza
- Credibilità
- Attualità

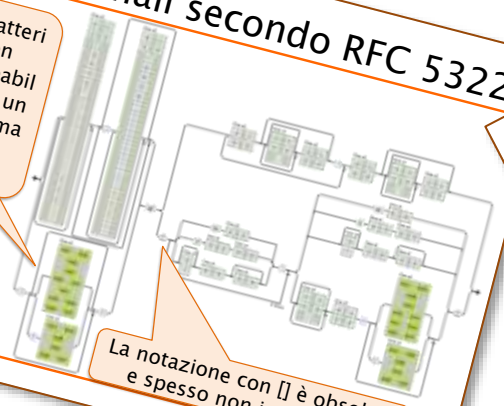
Dipendenti dal sistema

- Accessibilità
- Conformità
- Riservatezza
- Efficienza
- Precisione
- Tracciabilità
- Comprensibilità
- Disponibilità
- Portabilità
- Ripristinabilità

15

OWA: Email secondo RFC 5322

I caratteri non stampabili sono un problema per i client

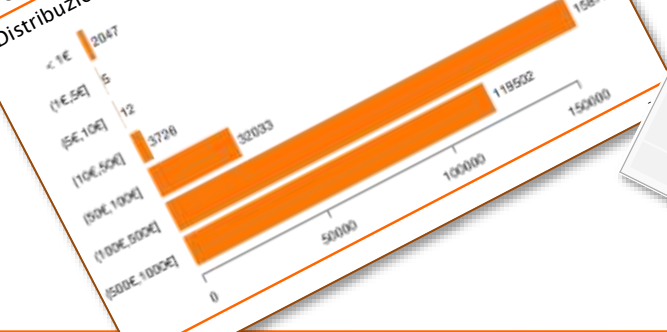


La notazione con [] è obsoleta e spesso non implemtata

30

Costi Rendicontabili

Distribuzione solo per progetti con costi fino a 1000€



Abdul Ahad Momand

Cosmonauta dell'Intercosmos

Nazionalità	Afghanistan
Status	Ritirato
Data di nascita	1° gennaio 1959
Selezione	1992
Altre attività	Pilota
Tempo nello spazio	8 giorni, 20 ore e 26 minuti
Missioni	Soyuz TM-5 Soyuz TM-6
Data ritiro	1988

Ringraziamenti

A tutti coloro con cui ho condiviso questo e percorso e da cui ho imparato molto

- Antonio Vetrò
- Domenico Natale
- Andrea Trenta
- Rifat Rashid
- e anche:
  - L.Canova, R.Iemma, F.Iuliano, A.Melandri, F.Morando, C.Orozco Minotas, G.Procaccianti, G.Rizzo

RIFERIMENTI

Riferimenti

- ISO/IEC 25012:2008, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model
- ISO/IEC 25024:2015, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality
- Vetrò, Antonio; Canova, Lorenzo; Torchiano, Marco; Orozco Minotas, Camilo; Iemma, Raimondo; Morando, Federico “Open Data Quality Measurement Framework: Definition and Application to Open Government Data”GOVERNMENT INFORMATION QUARTERLY, Vol.33, pp.325-337, ISSN:0740-624X
- Torchiano, Marco; Vetro', Antonio; Iuliano, Francesca “Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study” in IEEE 41st Annual Computer Software and Applications Conference (COMPSAC 2017)

Riferimenti

- M.Ziemann, Y. Eren, A. El-Osta. "Gene name errors are widespread in the scientific literature" Genome Biology 17(1), 2016, p.177
  - https://doi.org/10.1186/s13059-016-1044-7
- How to Find or Validate an Email Address
  - http://www.regular-expressions.info/email.html
- Open Coesione
  - https://opencoesione.gov.it/it/
- DBPedia
  - http://wiki.dbpedia.org