Categorizzazione degli open-data italiani sul profilo europeo dei metadati DCAT-AP

Simone Gasperoni
Matricola 489184
Corso di laurea magistrale in Ingegneria informatica

Relatore: Prof.ssa Carla Limongelli
Co-relatore: Dott. Paolo Starace

23/03/2017
Dipartimento di Ingegneria
Via Vito Volterra, 62, Roma

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.
LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

CSV **Pronto Soccorso - Accessi in tempo reale**
Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ
TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO
CREATO IL: 02/07/2015
AGGIORNATO IL: 02/07/2015
FREQUENZA DI AGGIORNAMENTO: 6 MINUTI
PERIODO TEMPORALE: 2015
OPENESS RATING: 1
FLUSSO DI ATTIVITÀ
SOSTENITORI: 0
LOCALIZZAZIONE

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.
LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

RISORSA FISICA

CSV **Pronto Soccorso - Accessi in tempo reale**
Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ
TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO
CREATO IL: 02/07/2015
AGGIORNATO IL: 02/07/2015
FREQUENZA DI AGGIORNAMENTO: 6 MINUTI
PERIODO TEMPORALE: 2015
OPENESS RATING: 1
FLUSSO DI ATTIVITÀ
SOSTENITORI: 0
LOCALIZZAZIONE

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

METADATI

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.

LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

CSV **Pronto Soccorso - Accessi in tempo reale**

Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ

TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO

CREATO IL: 02/07/2015

AGGIORNATO IL: 02/07/2015

FREQUENZA DI AGGIORNAMENTO: 6 MINUTI

PERIODO TEMPORALE: 2015

OPENESS RATING: 1

FLUSSO DI ATTIVITÀ

SOSTENITORI: 0

LOCALIZZAZIONE

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.

LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

CSV **Pronto Soccorso - Accessi in tempo reale**

Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ

TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO

CREATO IL: 02/07/2015

AGGIORNATO IL: 02/07/2015

FREQUENZA DI AGGIORNAMENTO: 6 MINUTI

PERIODO TEMPORALE: 2015

OPENESS RATING: 1

FLUSSO DI ATTIVITÀ

SOSTENITORI: 0

LOCALIZZAZIONE

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.
LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

CSV **Pronto Soccorso - Accessi in tempo reale**
Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ
TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO
CREATO IL: 02/07/2015
AGGIORNATO IL: 02/07/2015
FREQUENZA DI AGGIORNAMENTO: 6 MINUTI
PERIODO TEMPORALE: 2015
OPENESS RATING: 1
FLUSSO DI ATTIVITÀ
SOSTENITORI: 0
LOCALIZZAZIONE

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.
LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

CSV **Pronto Soccorso - Accessi in tempo reale**
Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ
TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO
CREATO IL: 02/07/2015
AGGIORNATO IL: 02/07/2015
FREQUENZA DI AGGIORNAMENTO: 6 MINUTI
PERIODO TEMPORALE: 2015
OPENESS RATING: 1
FLUSSO DI ATTIVITÀ
SOSTENITORI: 0
LOCALIZZAZIONE

DATASET OPEN DATA: dati e metadati

- INTRODUZIONE
- **DATASET OPEN DATA**
- MODELLO DATI DCAT-AP
- OPEN DATA HUB
- METADATO DI CATEGORIA
- MACHINE LEARNING
- SPERIMENTAZIONI
- SVILUPPI

Dataset | Flusso di attività | Correlazioni

Pronto Soccorso - Accessi in tempo reale

Il dataset riporta in tempo reale gli accessi nei Pronto Soccorso del Lazio. Si evidenzia che questi dati non possono essere utilizzati per misurare la qualità e tempestività dell'assistenza fornita nei Pronto Soccorso, ma esclusivamente per conoscere le eventuali attese presso i PPSS. In base alle condizioni d'urgenza, il livello di gravità e la priorità si attribuisce a ciascun paziente un codice Triage, corrispondente alla sua condizione: - codice rosso: molto critico, pericolo di vita, priorità massima, accesso immediato alle cure; - codice giallo: mediamente critico, presenza di rischio evolutivo, possibile pericolo di vita; - codice verde: poco critico, assenza di rischi evolutivi, prestazioni differibili; - codice bianco: non critico, pazienti non urgenti.
LICENZA: CREATIVE COMMONS ATTRIBUTION 4.0 (CC BY 4.0)

DATI E RISORSE

CSV **Pronto Soccorso - Accessi in tempo reale**
Aggiornamento dei dati al: 11-03-2017 alle ore: 18:54

API | RDF

CATEGORIA: SANITÀ
TAG: PRONTO SOCCORSO |

ORGANIZZAZIONE: REGIONE LAZIO
CREATO IL: 02/07/2015
AGGIORNATO IL: 02/07/2015
FREQUENZA DI AGGIORNAMENTO: 6 MINUTI
PERIODO TEMPORALE: 2015
OPENESS RATING: 1
FLUSSO DI ATTIVITÀ
SOSTENITORI: 0
LOCALIZZAZIONE

PROFILO DI APPLICAZIONE DEI DATI DCAT-AP-it

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

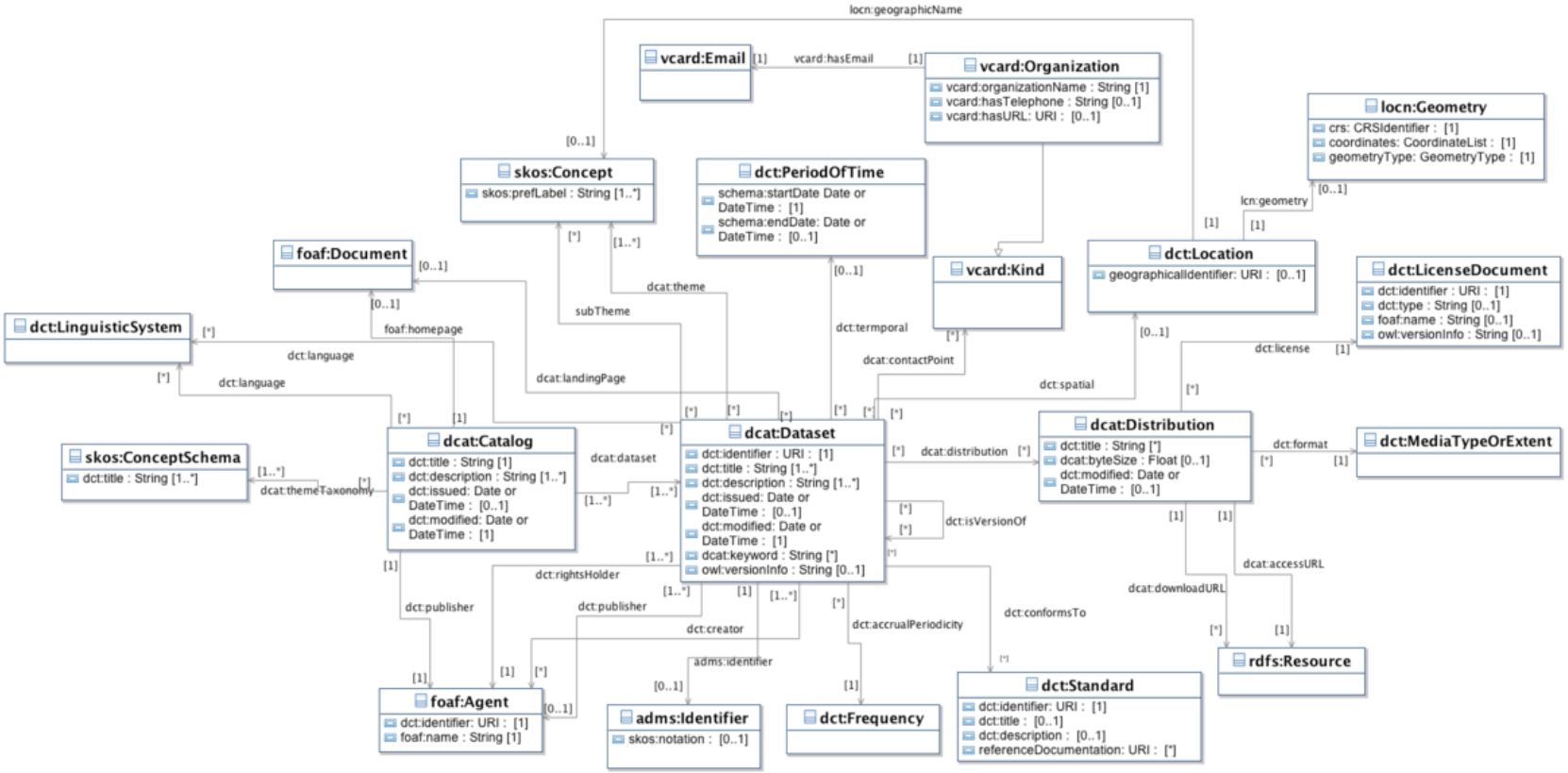
OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI



OPEN DATA HUB ITALY

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

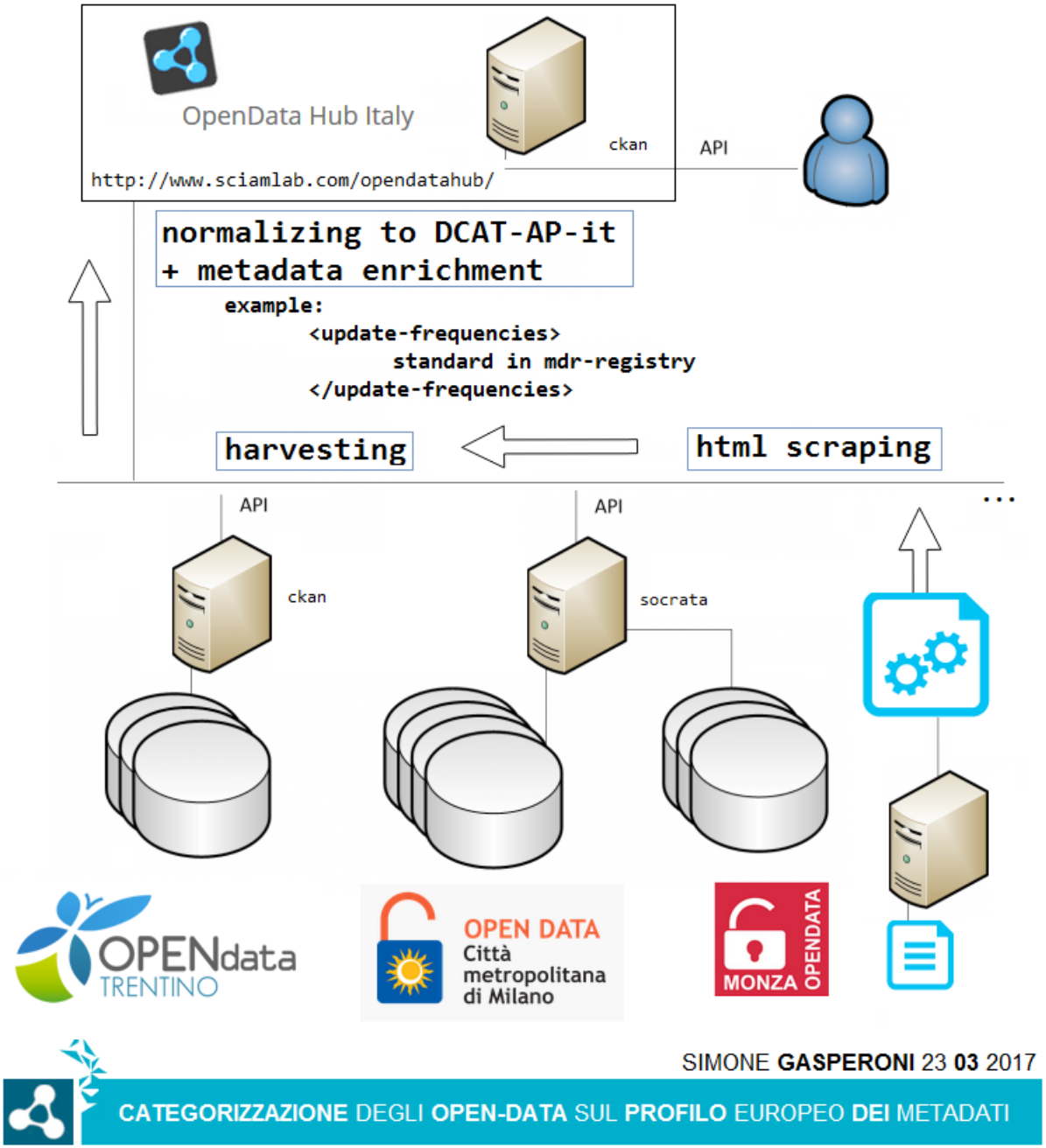
METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

SCIAMLAB



OPEN DATA HUB ITALY

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

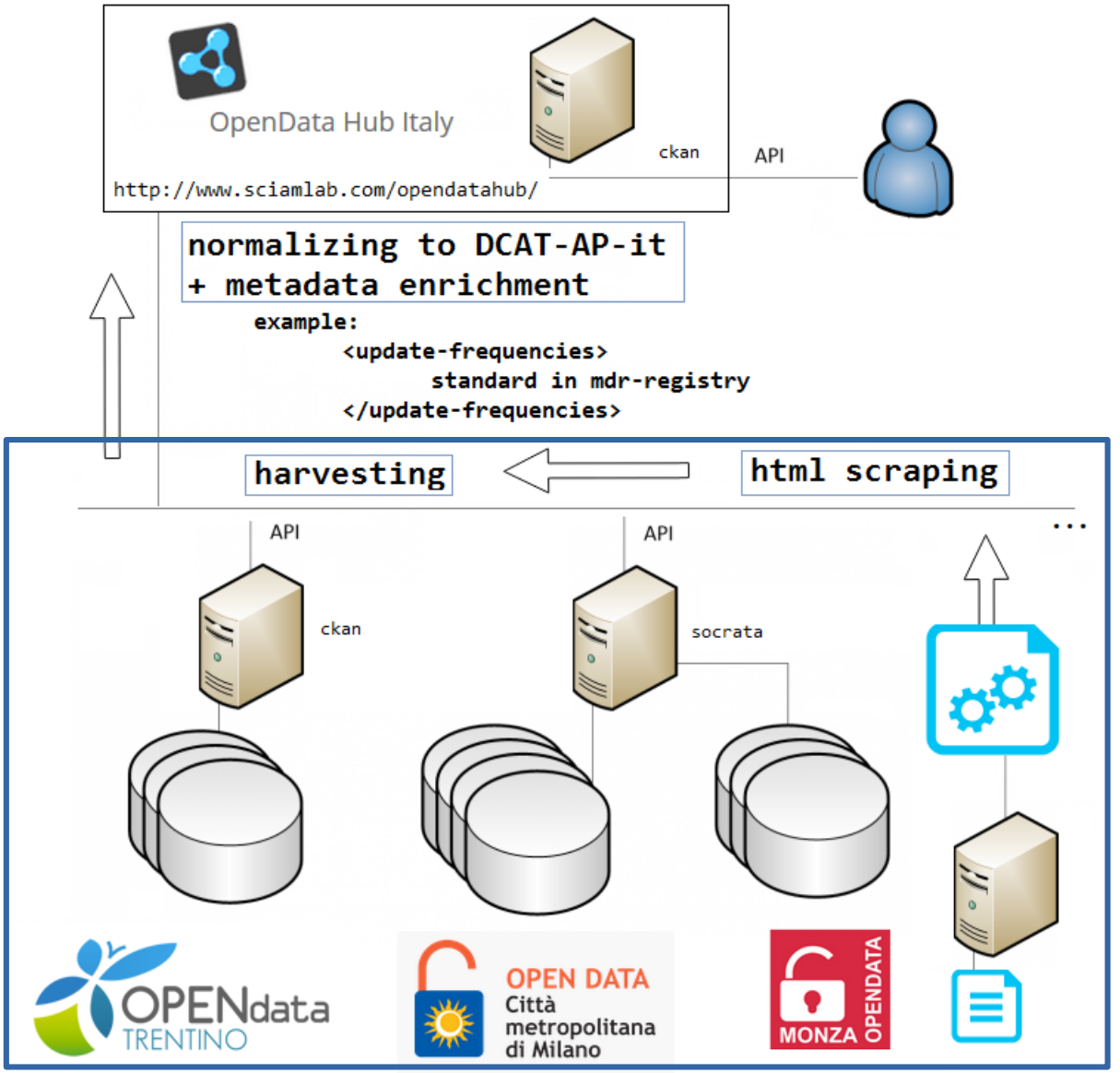
OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI



OPEN DATA HUB ITALY

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

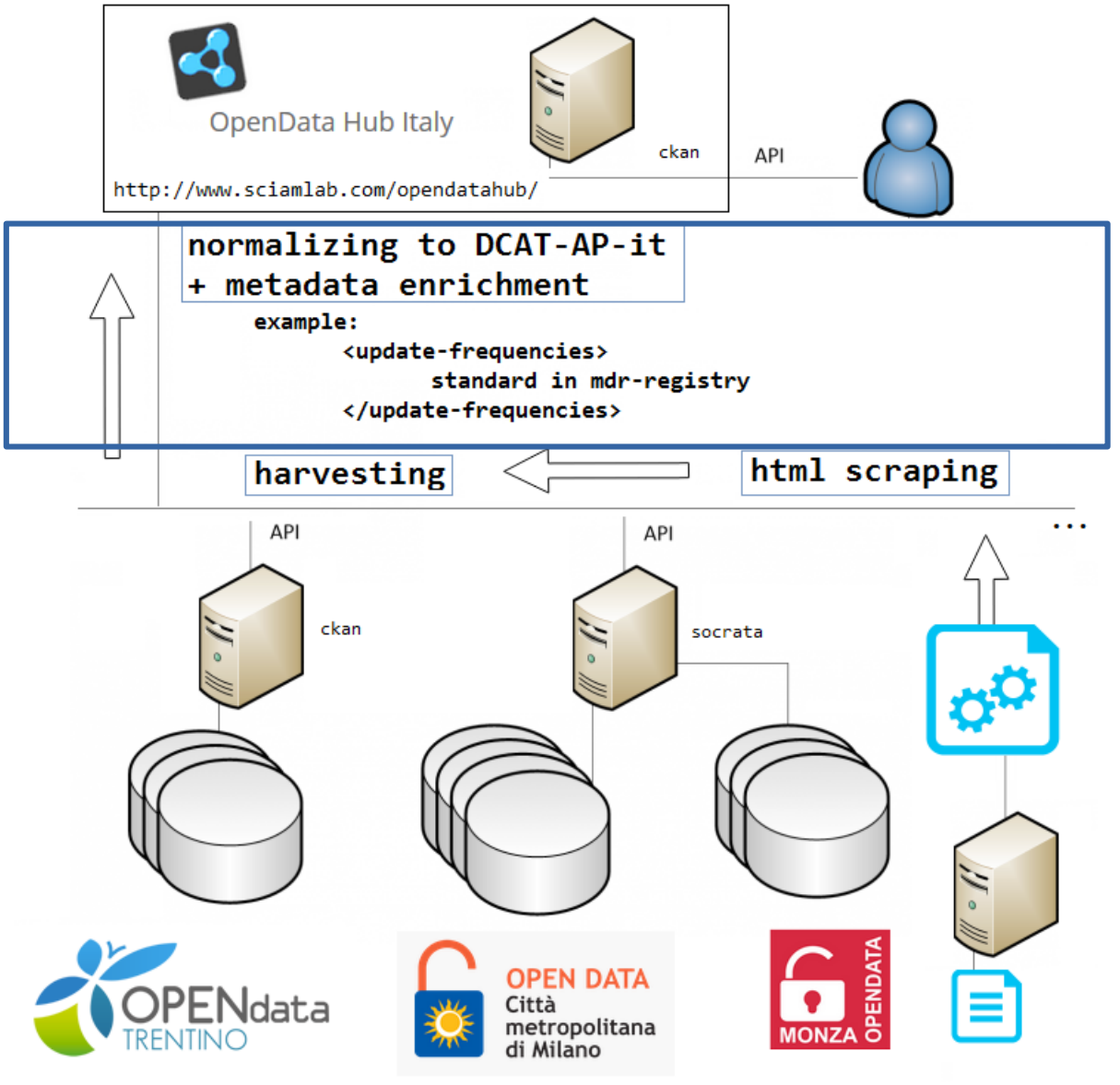
OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI



OPEN DATA HUB ITALY

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

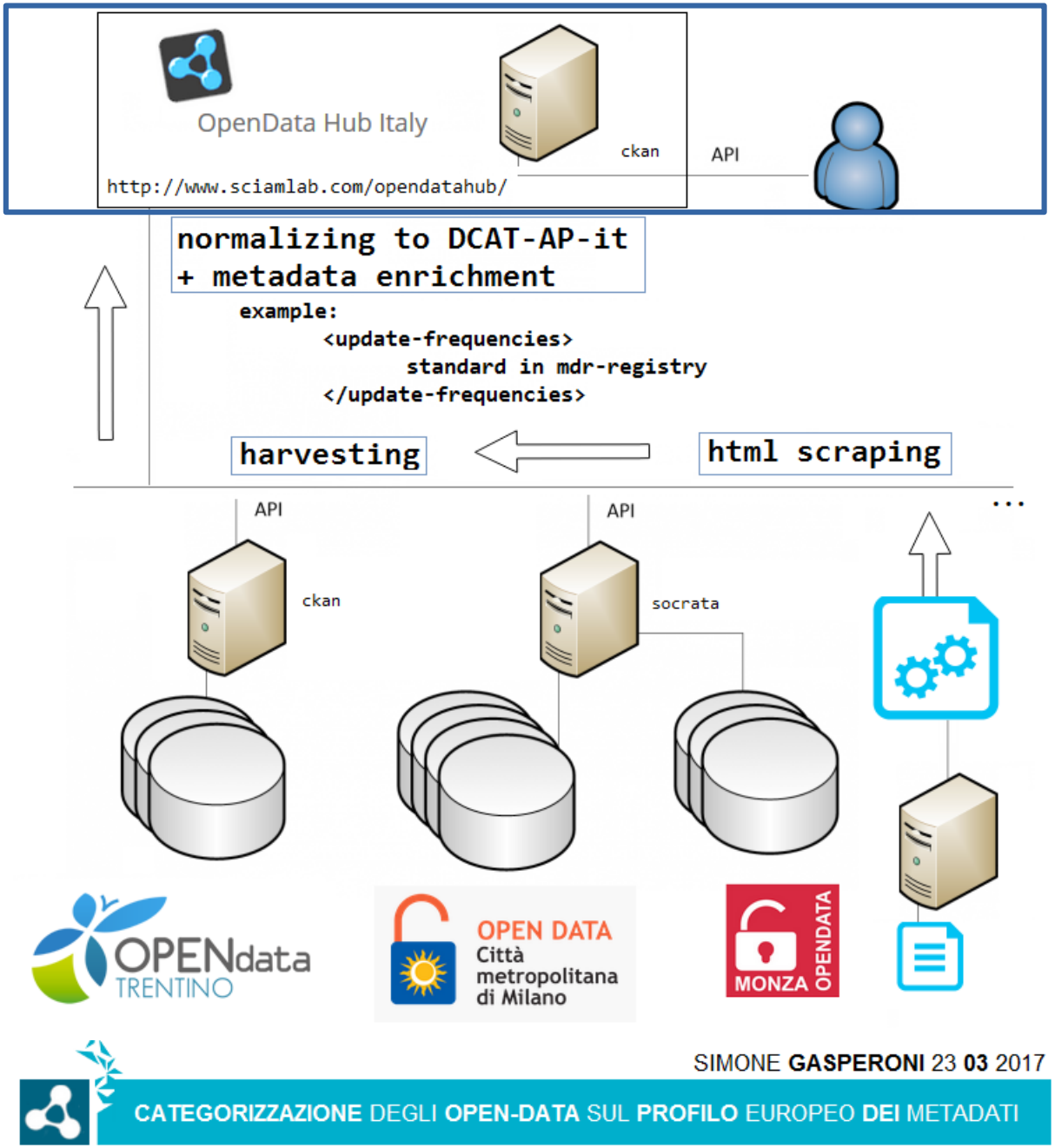
METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

SCIAMLAB



CATEGORIZZAZIONE DEI DATASET OPEN-DATA

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

- Molte amministrazioni non utilizzano il metadato di categoria nella pubblicazione dei dataset
- Molte amministrazioni utilizzano un catalogo di categorie diverso da quello indicato dal profilo dei metadati
- L’assegnazione delle categorie è errata perché chi categorizza manualmente i dataset non è documentato sulle tematiche coperte dalle categorie

BRIDGE CONCETTI DEL TESAURO EUROVOC – CATEGORIE DCAT-AP

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

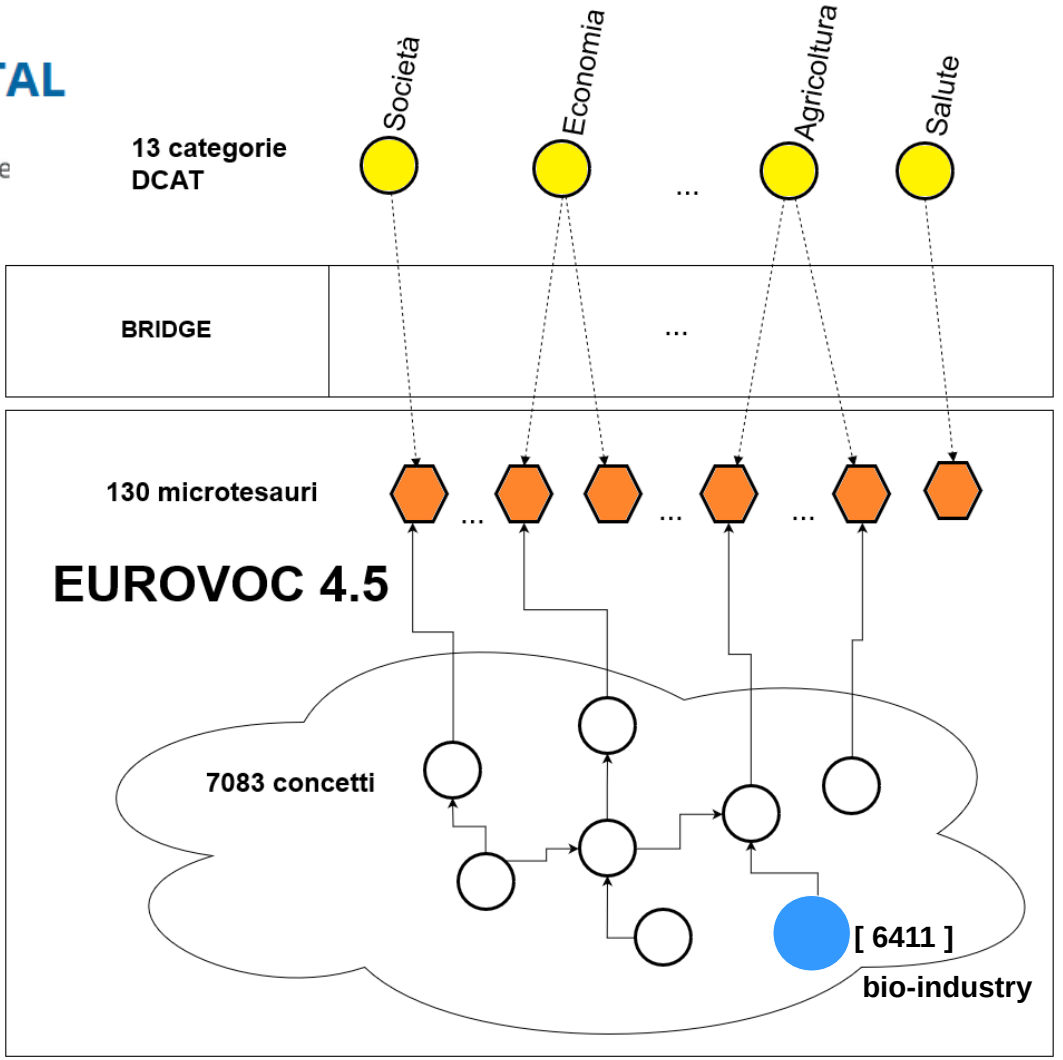
SVILUPPI

EUROPEAN DATA PORTAL

AgID Agenzia per l'Italia Digitale

Agricoltura
5636 forestry
5641 fisheries
6006 plant product
6011 animal product
6026 foodstuff
6031 agri-foodstuffs
6036 food technology
...

Energia
6606 energy policy
6616 oil industry
...



BRIDGE CONCETTI DEL TESAURO EUROVOC – CATEGORIE DCAT-AP

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

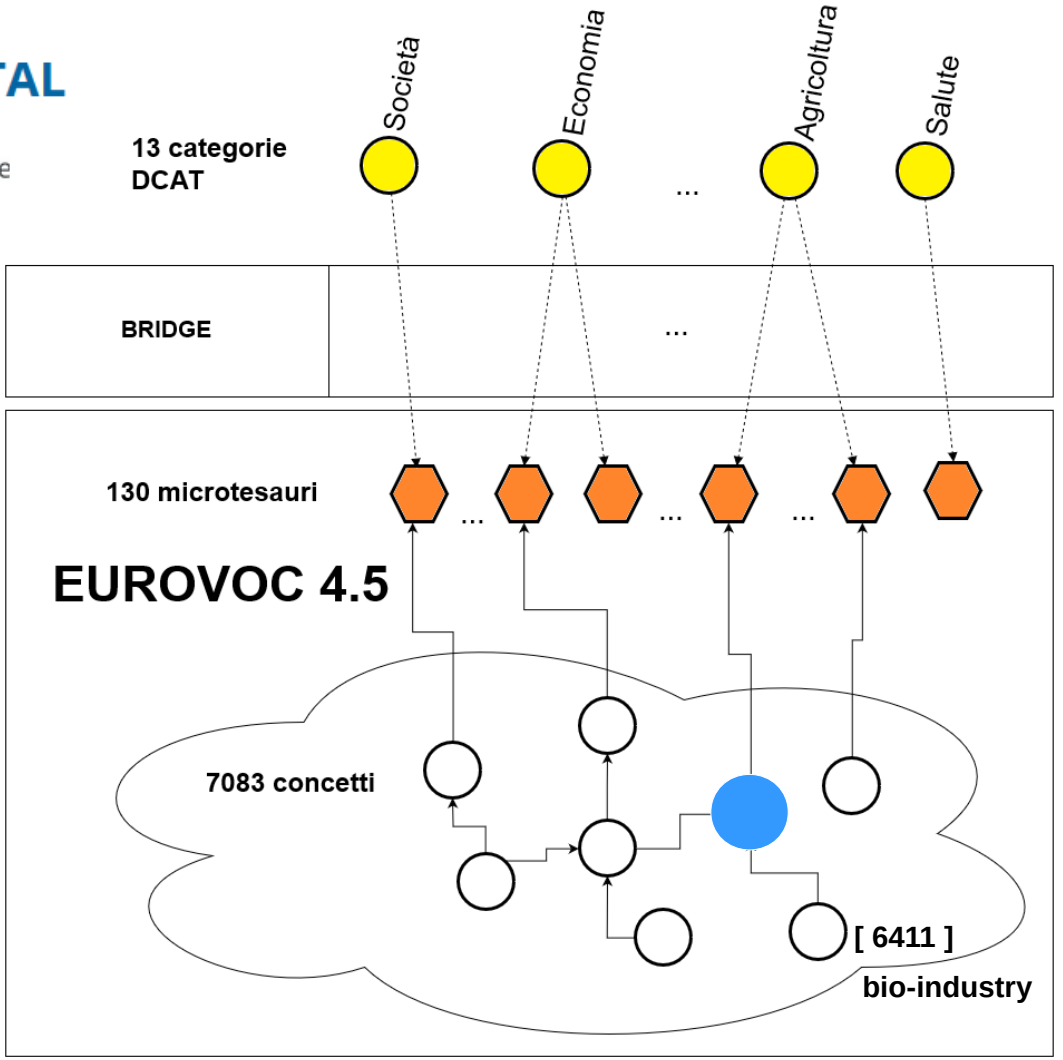
SVILUPPI

EUROPEAN DATA PORTAL

AgID Agenzia per l'Italia Digitale

Agricoltura
5636 forestry
5641 fisheries
6006 plant product
6011 animal product
6026 foodstuff
6031 agri-foodstuffs
6036 food technology
...

Energia
6606 energy policy
6616 oil industry
...



BRIDGE CONCETTI DEL TESAURO EUROVOC – CATEGORIE DCAT-AP

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

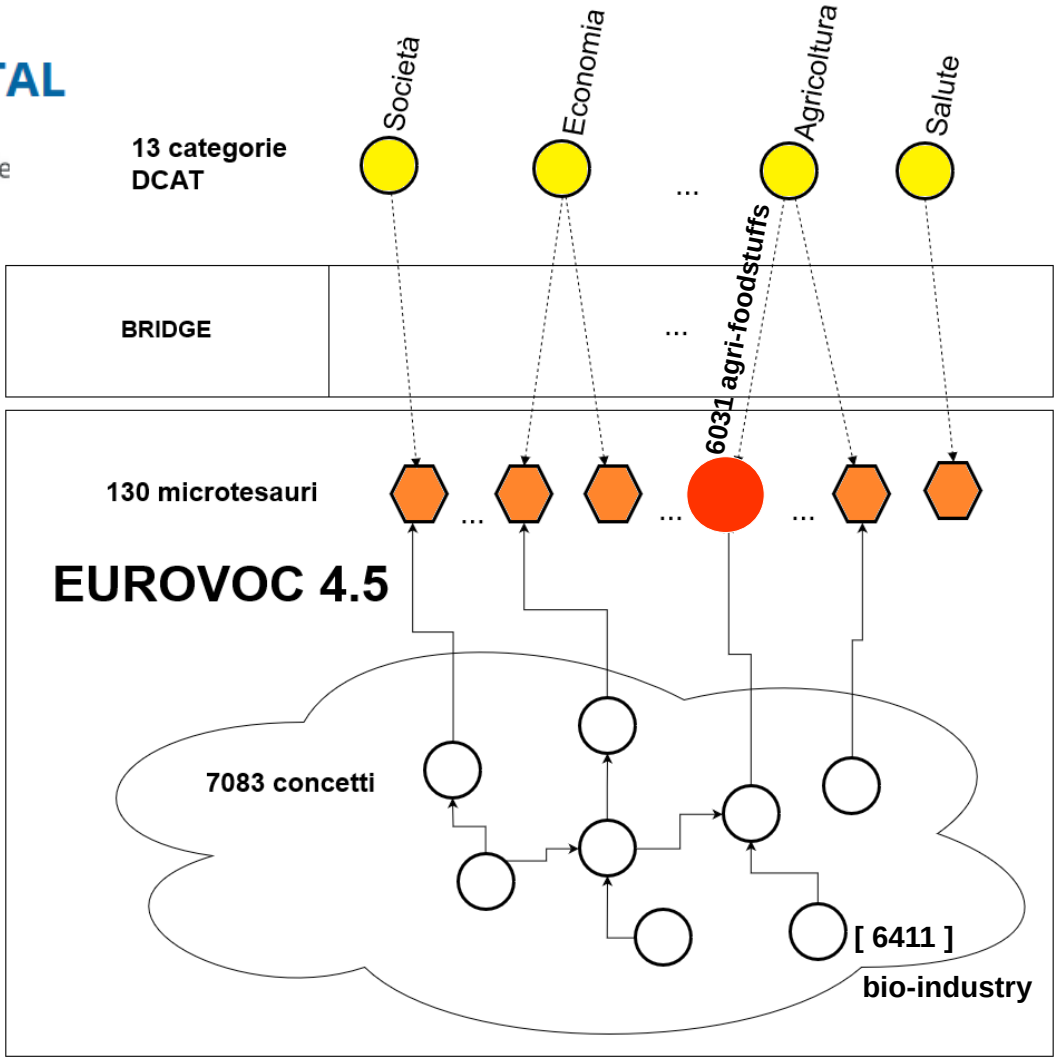
SVILUPPI

EUROPEAN DATA PORTAL

AgID Agenzia per l'Italia Digitale

Agricoltura
5636 forestry
5641 fisheries
6006 plant product
6011 animal product
6026 foodstuff
6031 agri-foodstuffs
6036 food technology
...

Energia
6606 energy policy
6616 oil industry
...



BRIDGE CONCETTI DEL TESAURO EUROVOC – CATEGORIE DCAT-AP

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

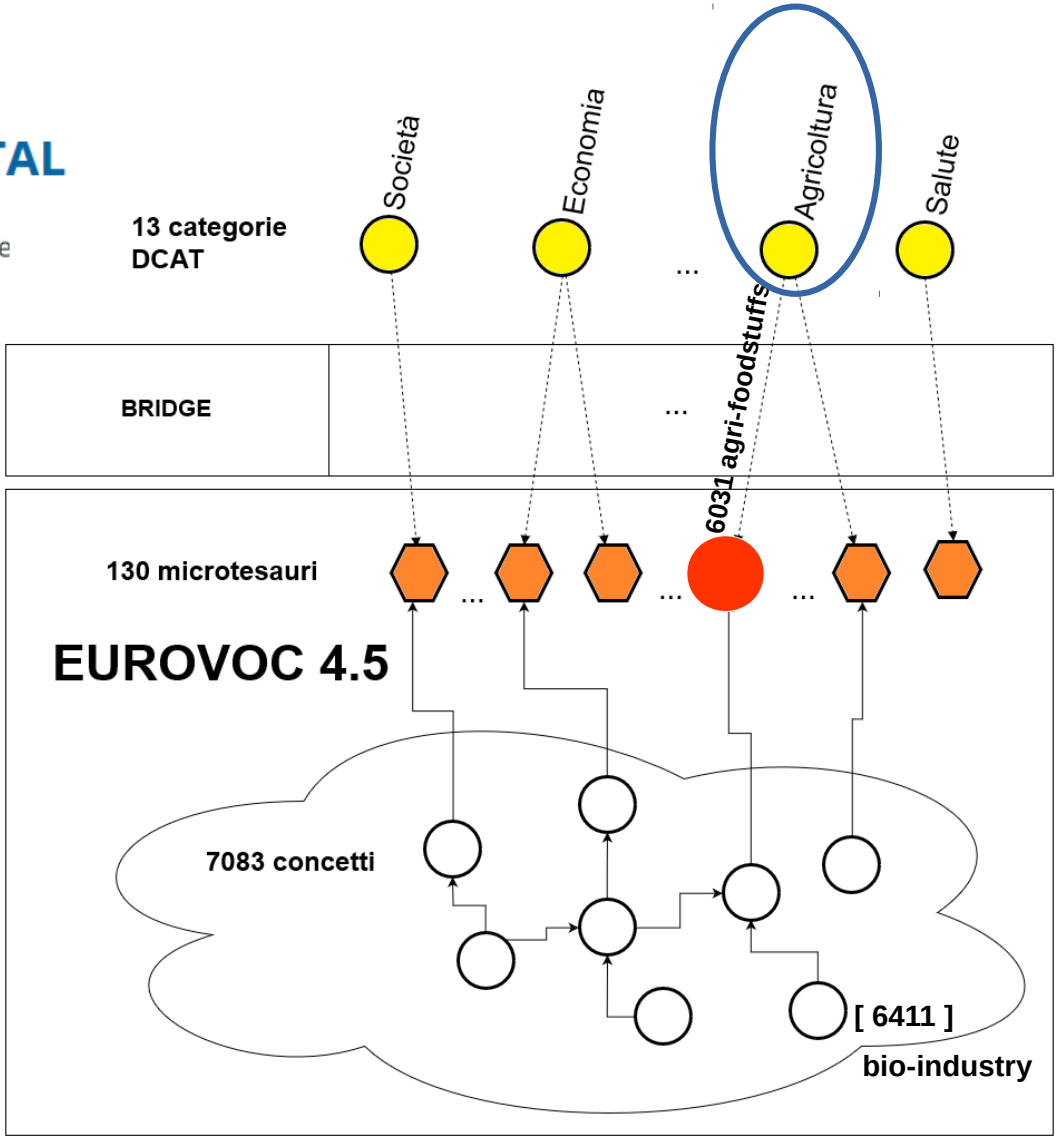
SVILUPPI

EUROPEAN DATA PORTAL

AgID Agenzia per l'Italia Digitale

Agricoltura
5636 forestry
5641 fisheries
6006 plant product
6011 animal product
6026 foodstuff
6031 agri-foodstuffs
6036 food technology
...

Energia
6606 energy policy
6616 oil industry
...



CLASSIFICATORI BAYESIANI & TEXT CATEGORIZATION

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

Formula di bayes

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Predicted value in TC

$$v = Max_i[P(C_i)\prod_{j=1}^{m} P(A_j = a_j|C_i)]$$

Modello di Bernoulli

$$\prod_{t=1}^{|V|}[b_{it} \cdot P(w_t|C) + (1 - b_{it}) \cdot (1 - P(w_t|C))]$$

Modello multinomiale

$$\prod_{h=1}^{len(D_j)} P(u_h|C)$$

Divergenza di Kullback-Leibler

$$\frac{1}{|d|} \cdot \log P(c_j) - \sum_{t=1}^{|V|} P(w_t|d) \cdot \log \frac{P(w_t|d)}{P(w_t|c_j)}$$

LOGICA DI ADDESTRAMENTO

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

JRC-ACQUIS

```
<head n="1">
        Decisione n. 2046/2002/CE
        del Parlamento europeo e del
        Consiglio,del 21 ottobre 2002
        che ...
</head>


<div type="body">
        ...
</div>

<textClass>
        <classCode scheme="eurovoc">
        206
        </classCode>
<classCode scheme="eurovoc">
        ...
</textClass>
```

BRIDGE EUROVOC-DCAT

BY EUROPEAN DATA PORTAL

```
put("0406", GOVE);
...
put("0416", EDUC);
...
put("3655", ECON);
```

LOGICA DI ADDESTRAMENTO

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

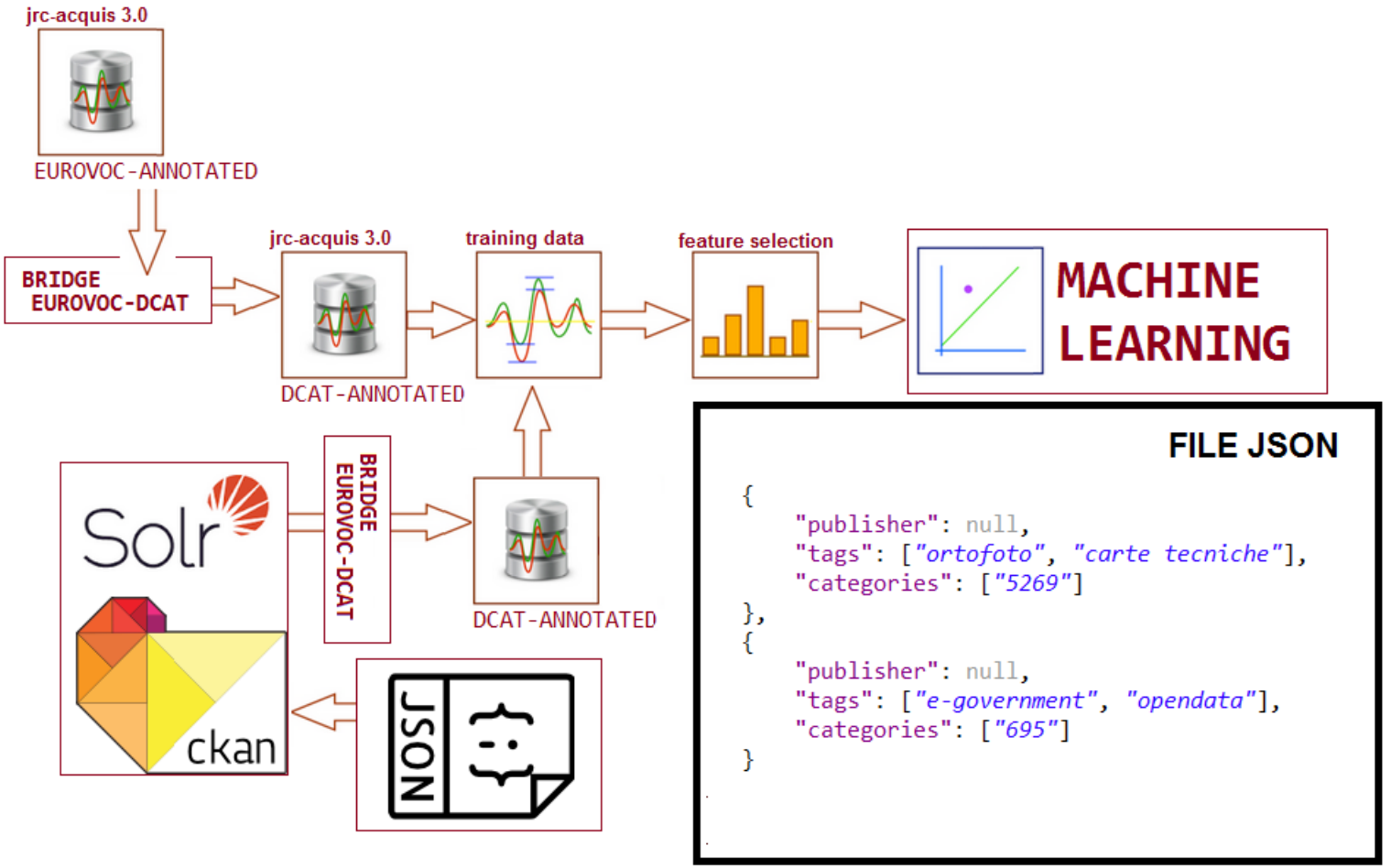
OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI



TEST FEATURE SELECTION

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

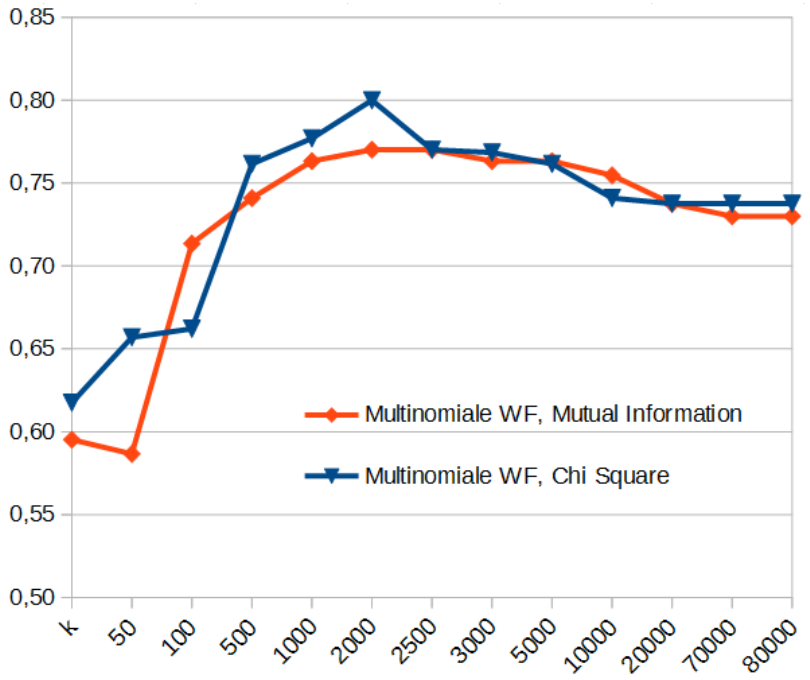
SVILUPPI

Mutual information

$$I(U,C) = \sum_{e_t \in [0,1]} \sum_{e_c \in [0,1]} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

Chi-square

$$\chi^2(t,c) = \sum_{t \in [0,1]} \sum_{c \in [0,1]} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$



TEST

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

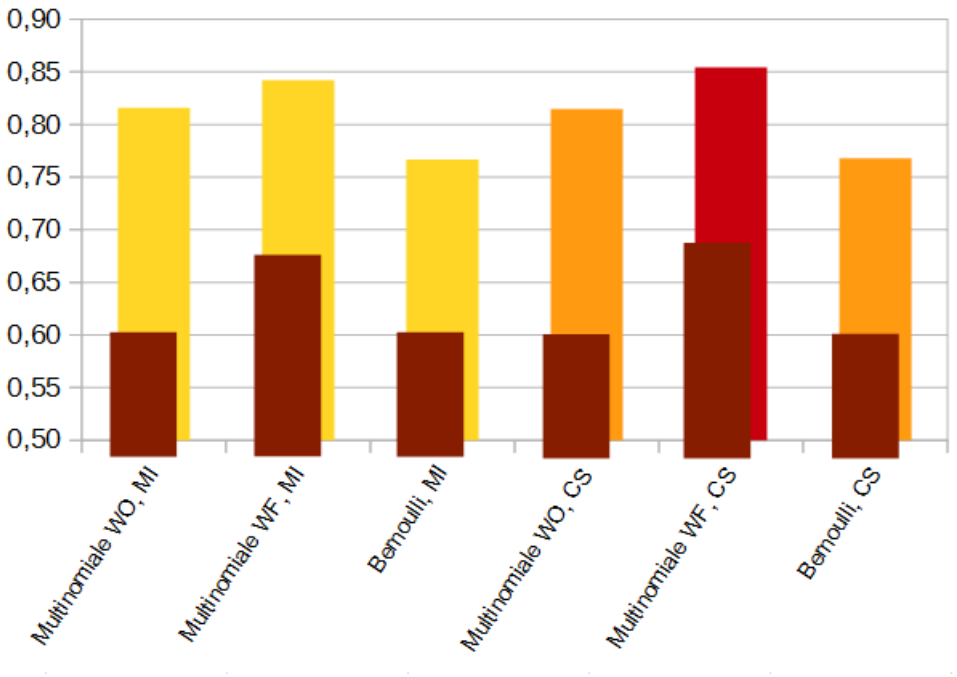
METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

Regione Lazio: 258 dataset
Provincia di Lecce: 325 dataset
Ministero delle infrastrutture e dei trasporti: 27 dataset
Regione Trentino: 615 dataset
Regione Lombardia: 1062 dataset
materia ortofotografica: 247 dataset
materia turistica: 133 dataset
materia scolastica: 224 dataset
trasporti ed autostrade: 42 dataset



SVILUPPI

INTRODUZIONE
DATASET OPEN DATA
MODELLO DATI DCAT-AP
OPEN DATA HUB
METADATO DI CATEGORIA
MACHINE LEARNING
SPERIMENTAZIONI
SVILUPPI

- **Outlier detection:**
  l'individuazione manuale degli outlier è dispendiosa,
  si potrebbero utilizzare algoritmi non supervisionati per la pattern recognition o la misura di confidenza del classificatore basato sulla divergenza KL

- **Potenziamento della feature extraction:**
  utilizzando tecniche di WSD, lemmatizzazione, ed entity recognition

- **API REST:**
  rilascio di un servizio

INTRODUZIONE

DATASET OPEN DATA

MODELLO DATI DCAT-AP

OPEN DATA HUB

METADATO DI CATEGORIA

MACHINE LEARNING

SPERIMENTAZIONI

SVILUPPI

GRAZIE

github.com/simonegasperoni/open-cata