GenAI e l'interoperabilità tra Pubbliche Amministrazioni

Conferenza Nazionale di Statistica

Fabrizio Davide – ISTAT (fabrizio.davide@istat.it) | Pietro Torre – ISTAT (pietro.torre@istat.it) |

Introduzione

La Piattaforma Digitale Nazionale Dati (PDND) è la base dell'ecosistema dell'interoperabilità tra le PP.AA. Italiane, quanto strumento per gestire l'autenticazione, l'autorizzazione e il tracciamento dei soggetti abilitati per garantire la sicurezza delle informazioni. Fornisce un insieme di regole condivise per semplificare gli accordi di interoperabilità snellendo i processi di istruttoria, riducendo oneri e procedure amministrative.

La PDND mette a disposizione il catalogo API, che riporta tutti gli e-service pubblicati sulla piattaforma dagli enti e tramite cui è possibile effettuare la richiesta di dati e successivamente integrare i propri servizi per i cittadini.

In questo quadro, il gruppo NDC di ISTAT si occupa della realizzazione, gestione e manutenzione del Catalogo nazionale della semantica dei dati (NDC), che è un componente chiave della PDND in quanto pubblica ontologie, vocabolari controllati e schemi dati collegati ai servizi della PDND. Il Catalogo è già esposto in www.schema.gov.it

In ambito NDC esistono vari temi aperti alla ricerca, tra cui ha particolare rilievo quello della correttezza dei materiali semantici che vengono catalogati e di consistenza del Catalogo. Infatti, Schema aggiunge continuamente nuovi materiali, man mano che le PP.AA. li producono in relazione agli e-services che espongono. Le ontologie aggiunte dovrebbero riusare concetti già definiti e registrati con URI sul w3id, ma potrebbero anche creare nuove definizioni che rischiano di diminuire la consistenza dell'universo concettuale del catalogo [1]. Questo aspetto si collega alla necessità di aumentare le possibilità di ricerca sul Catalogo che attualmente è possibile come ricerca testuale su alcuni metadati delle ontologie.

Tra i numerosi vantaggi dell'Intelligenza Artificiale per il settore pubblico è l'utilizzo dell'AI per favorire l'interoperabilità del settore pubblico (*AI4IOP*) [2]: l'IA può essere uno strumento essenziale per strutturare, curare, standardizzare e collegare meglio i dati della pubblica amministrazione, al fine di renderli più interoperabili all'interno e all'esterno di un singolo ente.

```
<?xml version="1.0"?>
<rdf:RDF xmlns="https://w3id.org/italia/onto/CPV/"
     xml:base="https://w3id.org/italia/onto/CPV/"
     xmlns:l0="https://w3id.org/italia/onto/l0/"
     xmlns:CLV="https://w3id.org/italia/onto/CLV/"
     xmlns:dct="http://purl.org/dc/terms/"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:dcat="http://www.w3.org/ns/dcat#"
     xmlns:foaf="http://xmlns.com/foaf/0.1/"
     xmlns:prov="http://www.w3.org/ns/prov#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
     xmlns:admsapit="https://w3id.org/italia/onto/ADMS/">
    <owl:Ontology rdf:about="https://w3id.org/italia/onto/CPV">
        <owl:versionIRI rdf:resource="https://w3id.org/italia/onto/CPV/0.8"/>
        <owl:imports rdf:resource="https://w3id.org/italia/onto/CLV"/>
        <owl:imports rdf:resource="https://w3id.org/italia/onto/TI"/>
        <owl:imports rdf:resource="https://w3id.org/italia/onto/l0"/>
        <dct:accrualPeriodicity rdf:resource="http://publications.europa.eu/r
        <dct:creator rdf:resource="https://w3id.org/italia/data/public-organi
        <dct:creator rdf:resource="https://w3id.org/italia/data/public-organi
        <dct:description xml:lang="it">Questa è l&apos;ontologia del profilo
L&apos;ontologia modella informazioni generali sulle persone fisiche come per
nascita, luogo di nascita, codice fiscale, residenza, cittadinanza e relazion
conoscenza sia di generiche relazioni di parentela ma non come registrate pre
l&apos;anagrafe.</dct:description>
        <dct:description xml:lang="en">This is the ontology of the Italian ap
ontology models general information on natural persons such as first name, su
tax code, residence, citizenship and relations between persons both generic o
relations but not as registered with the civil registry.</dct:description>
```

Immagine 1. Estratto dell'ontologia CPV-AP_IT.rdf

Materiali e metodi

Gli strumenti GenAI consentono di esplorare le soluzioni per le necessità di ricerca menzionate. In una prima fase la ricerca ha comportato lo sviluppo di tre scenari:

1. *Summarization* dei materiali semantici: una singola ontologia viene analizzata dal LLM secondo uno schema di analisi predefinito dai ricercatori [3], che riporta, tra l'altro, una quantificazione delle classi e delle relazioni, del numero di connessioni che un concetto ha con altri nell'ontologia per individuare i più rilevanti, ed altre metriche di grafo, ed una descrizione qualitativa dell'ontologia [4].
2. *Validazione* dei materiali semantici: sono stati prodotti dei dataset composti da ontologie corrette, e presenti nel catalogo, e loro variazioni affette da errori di vario genere [5] [6]. La tecnologia GenAI viene utilizzata per riconoscere le ontologie corrette ed apporre la label di correttezza.
3. *Correzione* dei materiali semantici affetti da errore: abbiamo richiesto la generazione di ontologie corrette a partire dalle ontologie riconosciute come errate nel passo precedente, distinguendo i casi di input secondo una scala di errore presente, articolata in cinque livelli di gravità [7] [8].

L'ambiente GenAI utilizzato è Google Gemini Pro, adottando quando necessario tecniche RAG. Sono state utilizzate, fra le altre, le ontologie gestite dal gruppo di lavoro del nostro stesso istituto e pubblicate su github italia [9].

Risultati e conclusioni

Scenario 1. *Summarization*: i risultati ottenuti sul data set preliminare di dieci ontologie sono ottimi secondo la valutazione di un esperto. È anche stato utilizzato un approccio di analisi automatica delle descrizioni che utilizza la soglia di affidabilità (sistema QERAS di Google), un punteggio di confidenza che determina quali previsioni restituire, e che è stato sempre superiore al 90%.

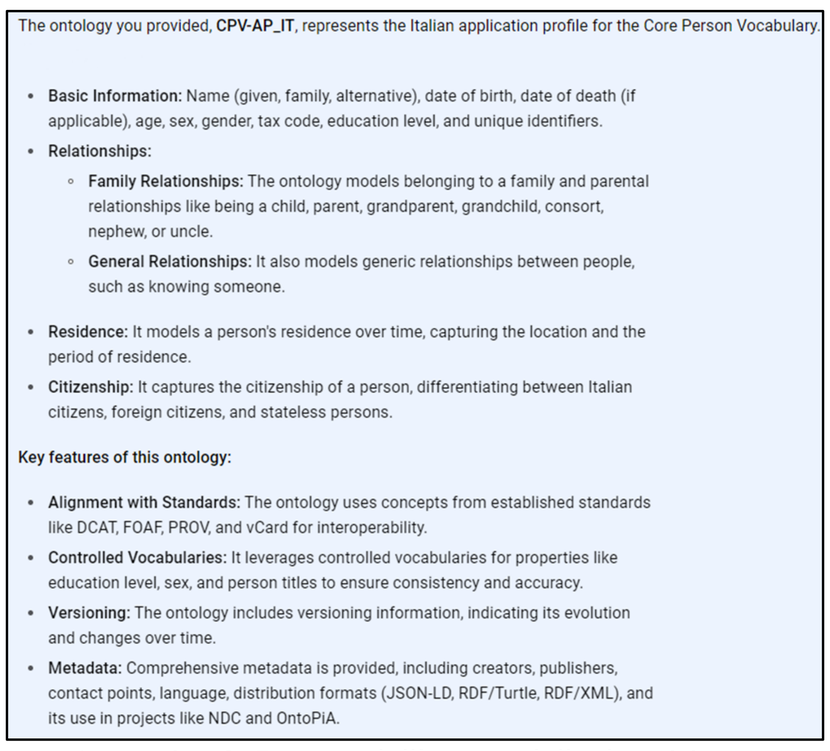
Nel caso 2. *Validazione*: i risultati di classificazione sono stati molto significativi (accuratezza del 95%, sensibilità del 93%) con un data set che riporta ontologie errate affette da errori di gravità medio alta (livelli 3, 4)

Nel caso 3. *Correzione*: la qualità delle ontologie corrette è stata valutata con la soglia di affidabilità, con risultati del 90% che decrescono a seconda del livello di gravità degli errori. Anche per gli errori gravi la proposta di ontologia corretta è risultata accettabile (66%) e quindi sufficiente.

Come prime conclusioni si può affermare che l'applicazione degli strumenti GenAI si dimostra molto utile ad avanzare alcuni temi aperti di ricerca sulla semantica nell'ambito del Catalogo Nazionale Dati, e presenta il potenziale per migliorare in efficacia la piattaforma di interoperabilità tra le PA.

```
+ Code  + Text  Commands

[ ] import base64
    import vertexai
    from vertexai.generative_models import GenerativeModel, Part, FinishReason
    import vertexai.preview.generative_models as generative_models

    def generate():
      vertexai.init(project="vocal-entity-412509", location="us-central1")
      model = GenerativeModel(
        "gemini-1.5-pro-001",
      )
      responses = model.generate_content(
          [document1, """can you describe this onthology?"""],
          generation_config=generation_config,
          safety_settings=safety_settings,
          stream=True,
      )

      for response in responses:
        print(response.text, end="")

    document1 = Part.from_data(
        mime_type="text/plain",
        data=base64.b64decode("""PD94bWwgdmVvc2lvbiI0iMS4wIi8
```

Immagine 2. Estratto di codice per l'interrogazione



The ontology you provided, **CPV-AP_IT**, represents the Italian application profile for the Core Person Vocabulary.

- **Basic Information:** Name (given, family, alternative), date of birth, date of death (if applicable), age, sex, gender, tax code, education level, and unique identifiers.
- **Relationships:**
  - **Family Relationships:** The ontology models belonging to a family and parental relationships like being a child, parent, grandparent, grandchild, consort, nephew, or uncle.
  - **General Relationships:** It also models generic relationships between people, such as knowing someone.
- **Residence:** It models a person's residence over time, capturing the location and the period of residence.
- **Citizenship:** It captures the citizenship of a person, differentiating between Italian citizens, foreign citizens, and stateless persons.

Key features of this ontology:

- **Alignment with Standards:** The ontology uses concepts from established standards like DCAT, FOAF, PROV, and vCard for interoperability.
- **Controlled Vocabularies:** It leverages controlled vocabularies for properties like education level, sex, and person titles to ensure consistency and accuracy.
- **Versioning:** The ontology includes versioning information, indicating its evolution and changes over time.
- **Metadata:** Comprehensive metadata is provided, including creators, publishers, contact points, language, distribution formats (JSON-LD, RDF/Turtle, RDF/XML), and its use in projects like NDC and OntoPiA.

Immagine 3. Estratto dell'output della descrizione

Bibliografia

[1]. Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05
[2] Tangi, L., Combetto, M., Martin Bosch, J. and Rodriguez Müller, P., Artificial Intelligence for Interoperability in the European Public Sector, Publications Office of the European Union, Luxembourg, 2023
[3]. Ciatto, G., Agiollo, A., Magnini, M., & Omicini, A. (2024). Large language models as oracles for instantiating ontologies with domain-specific knowledge. arXiv preprint arXiv:2404.04108
[4]. Carta, S., Giuliani, A., Piano, L., Podda, A. S., Pompianu, L., & Tiddia, S. G. (2023). Iterative zero-shot LLM prompting for knowledge graph construction. arXiv preprint arXiv:2307.01128
[5]. Gangemi, A., Catenacci, C., Ciaramita, M., & Lehmann, J. (2005, December). A theoretical framework for ontology evaluation and validation. In SWAP (Vol. 166, p. 16)
[6]. Raatikka, V., & Hyvönen, E. (2002). Ontology-based semantic metadata validation. 46 h Fg E 9 i VIE 7@ GPQ qhrhrRq@ rh, 6
[7]. Noy, N. F., & Musen, M. A. (2000, August). Algorithm and tool for automated ontology merging and alignment. In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831 (Vol. 115). sn
[8]. Solimando, A., Jiménez-Ruiz, E., & Guerrini, G. (2014). Detecting and correcting conservativity principle violations in ontology-to-ontology mappings. In The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13 (pp. 1-16). Springer International Publishing
[9] https://schema.gov.it/lodview/data/public-organization/ISTAT