

Datathon Progetto SMERALD
13 Giugno 2025 - Palermo

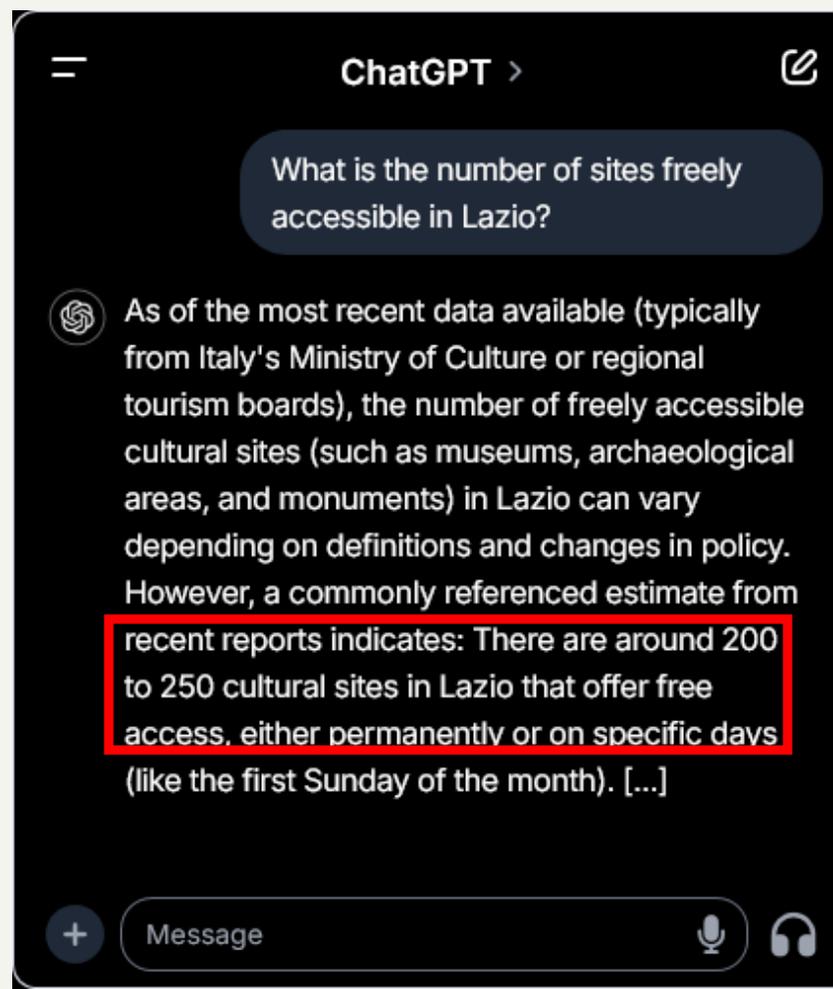
L'occasione dei dati (aperti) da non farsi sfuggire nell'era dell'IA

Giorgia Lodi

Prima Tecnologa presso l'Istituto di Scienze e Tecnologie della Cognizione del CNR (CNR-ISTC)
Socia della spinoff CNR BUP srl
Membro del Comitato Direttivo di OnData

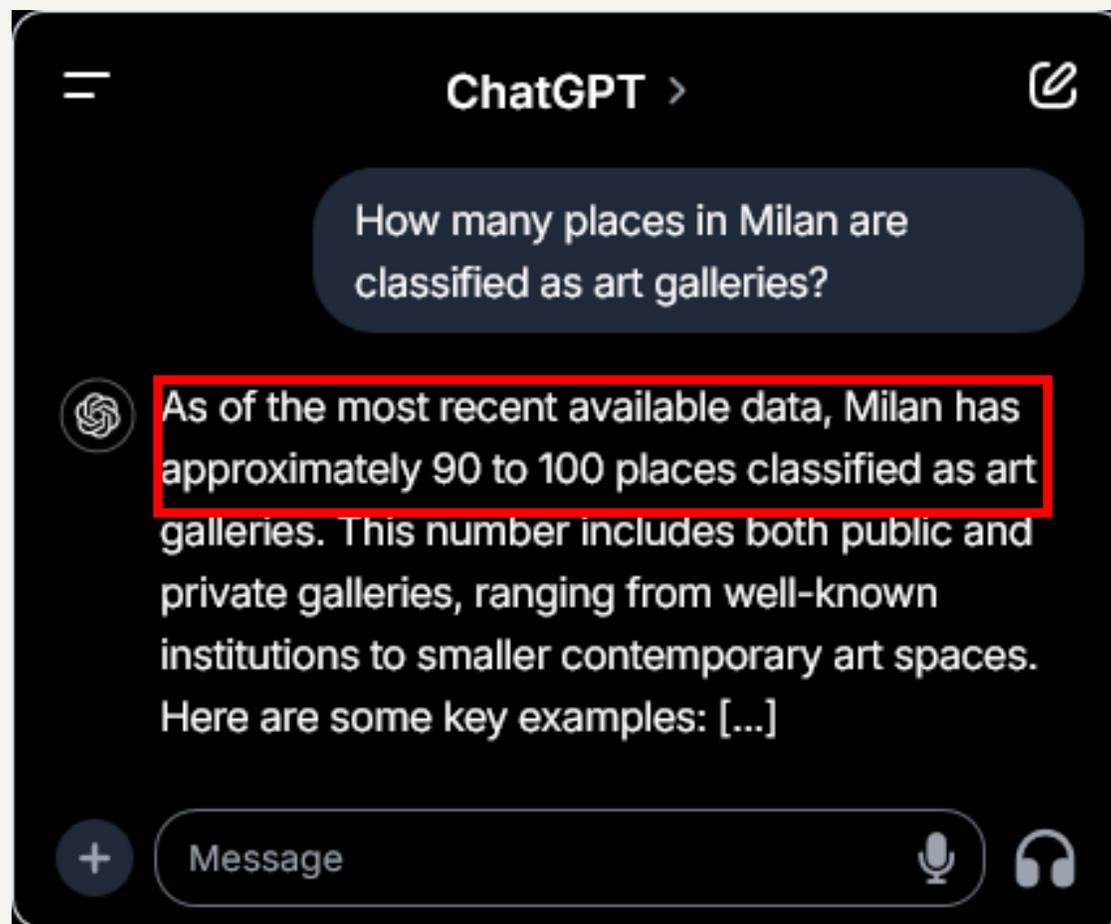
Ho chiesto a
ChatGPT
informazioni sui
beni culturali
italiani...

Domanda 1



Domanda 2

Ho chiesto a
ChatGPT
informazioni sui
beni culturali
italiani...



Ma noi abbiamo i linked open data dei beni culturali italiani... (chiedilo ad ArCo!)



Che accade se faccio una query SPARQL sui dati? Ho lo stesso risultato?

Domanda 1

Esegui la query su [Virtuoso](#)

```
</> Query SPARQL
1 SELECT COUNT(DISTINCT ?instituteorsite) AS ?numSites
2 WHERE {
3   ?instituteorsite rdf:type cis:CulturallInstituteOrSite ;
4   cis:hasSite ?site;
5   accessCondition:hasAccessCondition/rdfs:label ?accessLabel.
6   ?site cis:siteAddress/clvapit:hasRegion/rdfs:label ?regionLabel.
7   FILTER(REGEX(STR(?regionLabel), "^lazio$", "i"))
8   FILTER(REGEX(STR(?accessLabel), "ingresso libero", "i"))
9 }
```

Risultati della query SPARQL

numSites
432

Ma noi abbiamo i linked open data dei beni culturali italiani... (chiedilo ad ArCo!)



Che accade se faccio una query SPARQL sui dati? Ho lo stesso risultato?

Domanda 2

```
</> Query SPARQL
1 SELECT (COUNT(DISTINCT ?instituteorsite) AS ?numSites)
2 WHERE {
3   ?instituteorsite rdf:type cis:CulturalInstituteOrSite ;
4   dc:type ?type ;
5   cis:hasSite/cis:siteAddress/clvapit:hasCity/rdfs:label ?cityLabel.
6   FILTER(REGEX(STR(?cityLabel), "^milano$", "i"))
7   FILTER(REGEX(STR(?type), "galleria", "i"))
8 }
```

Risultati della query SPARQL

numSites
58

E se
proviamo a
istruire
ChatGPT
con i dati
aperti
disponibili?

Domanda 1

The screenshot displays a chatbot interface with a dark theme. On the left, a chat window shows a user asking "What is the number of sites freely accessible in Lazio?" and the chatbot replying "The query has found that there are 432 cultural institutes or sites in Lazio that are freely accessible." Below the chat is a "Domanda" section with a text input field containing "Scrivi qui la tua domanda..." and a "Chiedilo ad ArCo!" button. On the right, a "Query SPARQL" section shows the query code and a "Risultati della query SPARQL" section displaying the result "432".

Chatbot

What is the number of sites freely accessible in Lazio?

The query has found that there are 432 cultural institutes or sites in Lazio that are freely accessible.

Domanda

Scrivi qui la tua domanda...

Chiedilo ad ArCo!

Reset

Esegui la query su [Virtuoso](#)

```
</> Query SPARQL
1 SELECT COUNT(DISTINCT ?instituteorsite) AS ?numSites
2 WHERE {
3   ?instituteorsite rdf:type cis:CulturalInstituteOrSite ;
4   cis:hasSite ?site;
5   accessCondition:hasAccessCondition/rdfs:label ?accessLabel.
6   ?site cis:siteAddress/clvapit:hasRegion/rdfs:label ?regionLabel.
7   FILTER(REGEX(STR(?regionLabel), "^lazio$", "i"))
8   FILTER(REGEX(STR(?accessLabel), "ingresso libero", "i"))
9 }
```

Risultati della query SPARQL

numSites
432

E se
proviamo a
istruire
ChatGPT
con i dati
aperti
disponibili?

Domanda 2

The screenshot shows a chatbot interface with a dark theme. At the top left, it says "Chatbot". The user's question is "How many places in Milan are classified as art galleries?". The chatbot's response is "Based on the information provided, there are 58 art galleries located in Milan." Below this, there is a "Domanda" section with a text input field containing "Scrivi qui la tua domanda...". At the bottom, there are two buttons: "Chiedilo ad ArCo!" (with a rocket icon) and "Reset" (with a trash icon).

On the right side, there is a section titled "Esegui la query su Virtuoso" with a globe icon. Below it is a code editor showing a SPARQL query:

```
</> Query SPARQL
1 SELECT (COUNT(DISTINCT ?instituteorsite) AS ?numSites)
2 WHERE {
3 ?instituteorsite rdf:type cis:CulturalInstituteOrSite ;
4 dc:type ?type ;
5 cis:hasSite/cis:siteAddress/clvapit:hasCity/rdfs:label ?cityLabel.
6 FILTER(REGEX(STR(?cityLabel), "^milano$", "i"))
7 FILTER(REGEX(STR(?type), "galleria", "i"))
8 }
```

Below the query editor, it says "Risultati della query SPARQL" and shows a table with one row:

numSites
58

(1) Le allucinazioni

Uno dei grossi problemi è che producono cosiddette **allucinazioni**

Le allucinazioni sono **risposte non vere o fuorvianti** che **non** hanno **attinenza con la realtà**

Le risposte appaiono però verosimili

Negli esempi precedenti il fenomeno delle allucinazioni è stato mitigato



Stefano Zanero @raistolo.bsky.social · 2d

Cari amici de [@ilpost.it](#) apprezzo molto questa vostra spiegazione, che però manca un punto fondamentale: gli LLM non inventano "spesso" le cose, le inventano SEMPRE. Non hanno "ogni tanto" dei problemi di allucinazione. Ogni loro output È un'allucinazione, che OGNI TANTO ha attinenza con la realtà.

il Post @ilpost.it · 2d

Perché spesso Chatgpt si inventa le cose



Perché spesso Chatgpt si inventa le cose

Come date, leggi, titoli di libri e citazioni di canzoni, per non parlare dei risultati di calcoli matematici: i motivi sono principalmente due

[ilpost.link](#)



(2) Sono scatole nere

In base al dato in input si costruiscono **un modello all'interno** basato su fondamenti **matematici/statistici**

Il modello **non necessariamente memorizza dati** o struttura la conoscenza ma **conserva i pesi e le polarizzazioni appresi durante l'addestramento**, consentendo previsioni basate sui dati di ingresso.



(3) Ignorano vincoli sui dati

Uno forse dei più grandi problemi è che si **basano su ciò che trovano** sul Web (e.g, scraping di siti Web) **senza** considerare eventuali **limitazioni di riutilizzo**

Non ancora chiarissimo come licenziare il risultato ottenuto



(4) Senza dati...

Non esistono!

Il risultato dipende
fortemente dal dato di input
fornito

Anche in termini di bias e
caratteristiche di qualità

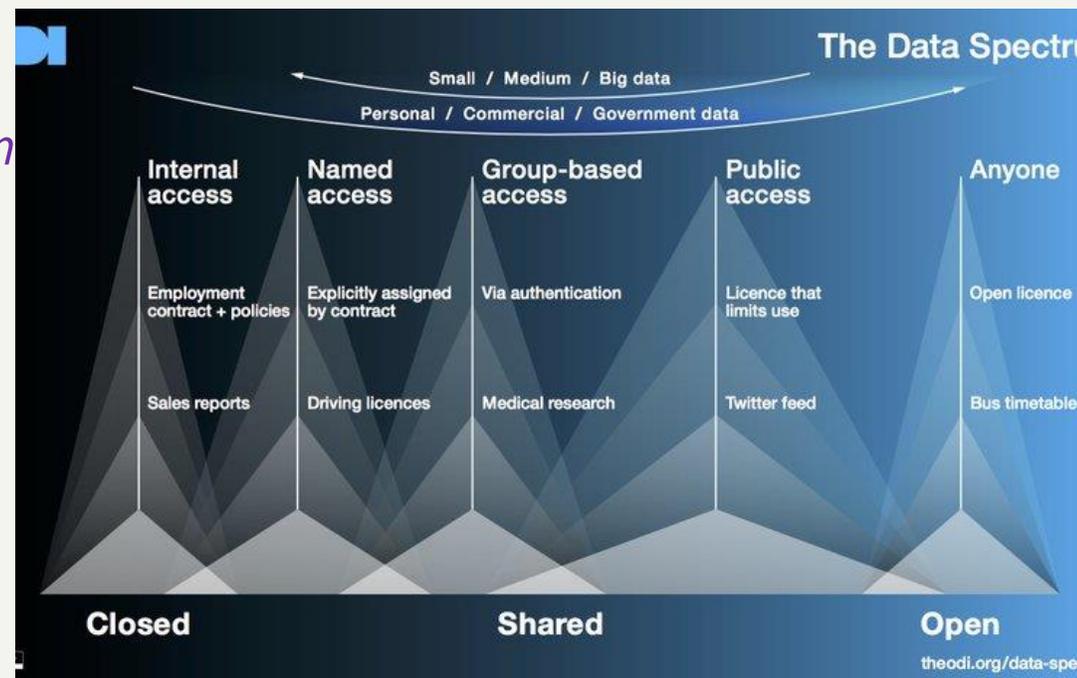


Ripartiamo da qui

Nell'era dell'Intelligenza artificiale abbiamo bisogno di...

Principle 1: *"a strong data infrastructure" as "the foundation for building an open, trustworthy data ecosystem on a global scale and that this can help address our most pressing challenges"*

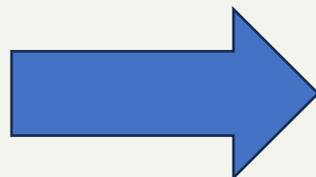
Principle 2: *"Strong data infrastructure includes data across the spectrum, from open to shared to closed. **But the best possible foundation is open data, supported and sustained as data infrastructure"***



Come migliorare i limiti delle attuali AI

(1)

Le allucinazioni



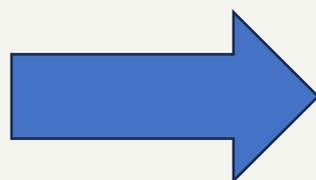
Si possono mitigare attraverso l'uso di tecniche come **RAG - Retrieval Augmented Generation**

Attraverso l'uso di dati strutturati disponibili per il riutilizzo

Come migliorare i limiti delle attuali AI

(2)

**Sono scatole
nere**



Non fanno ragionamenti, non capiscono il contesto

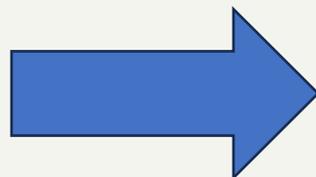
I **dati aperti** potrebbero essere utilizzati per **spiegare** il risultato ottenuto, soprattutto i linked open data (**explainability**)

I **dati aperti usati in input per l'allenamento possono essere conoscibili da chiunque** (principio della **trasparenza** - AI Act)

Come migliorare i limiti delle attuali AI

(3)

**Ignorano vincoli
sui dati**



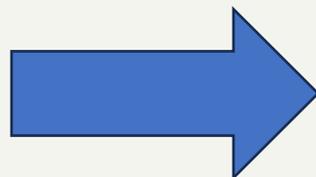
I dati aperti sono dati riutilizzabili da chiunque con pochi o zero limiti sul riutilizzo

I **dati aperti** diventano così fonte essenziale **per rispettare norme sul copyright**

Come migliorare i limiti delle attuali AI

(4)

**Senza dati non
esistono**



I dati aperti dovrebbero essere **accessibili** anche in maniera più diffusa grazie al **formato aperto**

Infrastruttura ideale per l'allenamento dell'AI

«*What **AI-ready data** means in practice depends on many factors, but established practices like **FAIR-ness** (findable, accessible, interoperable, reusable) and **linked data** are a strong starting point*»

Nel creare dati aperti, avere in mente le potenzialità per i software di AI

«*The **government** must **guarantee that there are simply no 'gaps to be filled' by secondary sources and hallucinations.** To do so, they should ensure that they are the **first and most important data provider for foundational models in all subject matters related to governmental affairs***»

ODI -

https://theodi.cdn.ngo/media/documents/UK_government_as_a_data_provider_for_AI.pdf

MA....



Garbage In - Garbage Out (GIGA law)



Se il dato **non è curato (dato di qualità)** anche sfruttando la **conoscenza di dominio degli esperti**, quello che accadrà è che anche i più avanzati modelli di intelligenza artificiale **risulteranno poco affidabili**

Prima indagine di AGID sull'uso dell'AI nelle pubbliche amministrazioni centrali

I principali risultati

L'indagine ha permesso di indagare numerosi aspetti legati a tecnologie, finanziamenti, modalità di procurement, stakeholder, impatti, criticità e sfide.

Il 42% dei progetti di IA nelle PA mira a migliorare l'efficienza operativa, il 24% a potenziare la gestione dei dati e il 18% a ottimizzare l'accesso ai servizi.

Circa il 75% ha un'estensione nazionale, ma non mancano iniziative sovranazionali. Le tecnologie più usate sono il Machine Learning tradizionale e, in crescita, l'IA generativa per testi e linguaggio naturale. Oltre il 60% dei progetti include chatbot e assistenti virtuali.

I dati per l'addestramento provengono soprattutto da banche dati interne, talvolta includendo dati personali o sintetici. Si rileva scarsa attenzione alla qualità dei dati, con possibili impatti negativi sull'affidabilità.

Le modalità di procurement sono varie, con prevalenza di Accordi Quadro e strumenti Consip.

Le competenze interne sono presenti ma limitate, con forte dipendenza da consulenti esterni.

Solo il 20% dei progetti ha KPI definiti, sollevando dubbi sulla capacità strategica delle amministrazioni.



Abbiamo bisogno della data quality by design!

Abbiamo bisogno di lavorare e agire in maniera diversa

- 1) Per il **pregresso** rivedere i processi per **inserire elementi di validazione nel ciclo di vita del dato**
- 2) Per **nuovi processi, pensare alla qualità dei dati** in termini di principi FAIR e caratteristiche **ISO fin dalle prime fasi di raccolta dei dati**
 - a. porsi le **giuste domande** conta
 - b. **raccogliere dati** secondo **principi di qualità**



Per concludere

1. La pubblica amministrazione può avere un ruolo cruciale nella costruzione dell'infrastruttura dei dati per l'AI (**data as a service**) in particolare con gli **open data**
2. Gli **open data che abbiamo ora non vanno benissimo**: bisogna puntare più in alto!
3. Non si può più lavorare con i dati di ogni natura come si lavorava anche solo 5 anni fa (**cultura del dato**)
4. Necessario **revisionare i processi di gestione dei dati** anche in ottica di apertura per alimentare **l'equità di accesso a quei dati utilizzati poi nell'allenamento** di software di intelligenza artificiale

Datathon Progetto SMERALD
13 Giugno 2025 - Palermo

Grazie!

Giorgia Lodi
giorgia.lodi@cnr.it

