



Agenzia nazionale per le nuove tecnologie,
l'energia e lo sviluppo economico sostenibile

Nuovi strumenti per data mining e indicizzazione

Dai dati alla conoscenza

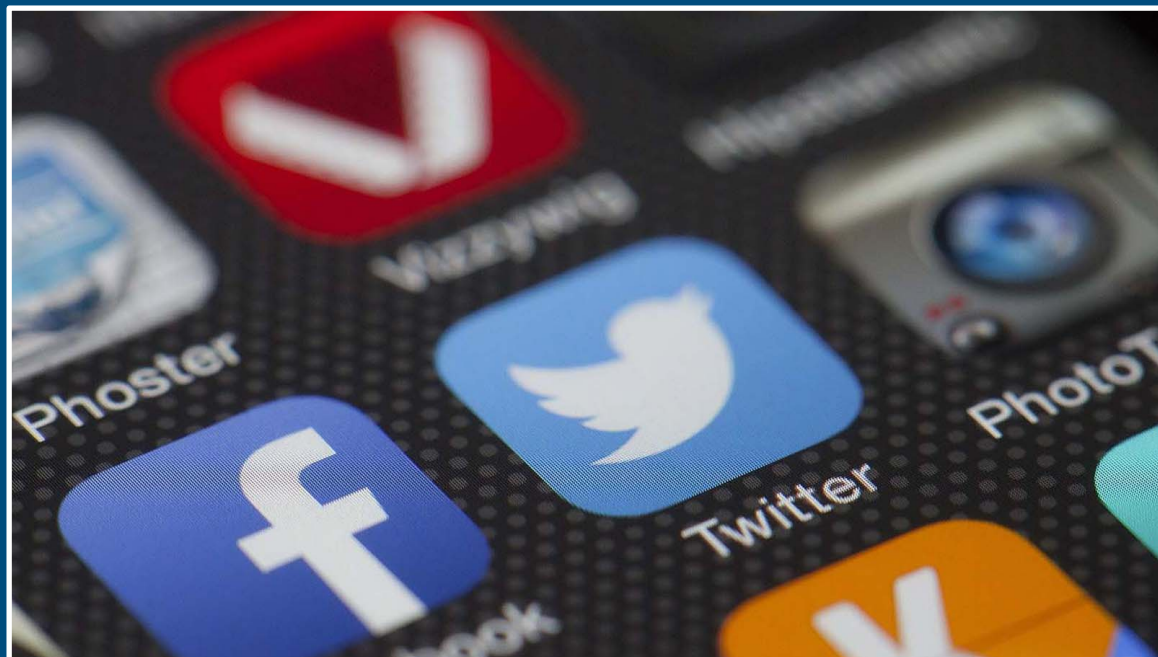
Palermo 18 gennaio 2018

Dr. Tommaso Nicosia *DTE-BBC* Ing. Giuseppe Alonge *SSPT-PROTER-OAC*



1101 0110 1100
0101 0010 1101
0001 0110 1110
1101 0010 1101
1111 1010 0000





Chi sono

Per iniziare

“Una base di conoscenza è una tecnologia usata per immagazzinare informazioni usate da sistemi computerizzati.” (wikipedia)

Una particolare base di dati.

Ma nascono le seguenti domande

- A quali domande riesce a rispondere la base di conoscenza ?.
- A chi deve fornire queste risposte ?.
- Per quale scopo fornirle ?.
- Come ottenerle ?.

Stato attuale

Come sono trattate le informazioni

- Dati su ambienti locali organizzato in cartelle e file
- Accessibili da terminale o desktop remoti
- Di non semplice trattamento per dimensioni e contenuto
- Informazioni non immediatamente disponibili per ricerche avanzate

Un esempio di dove arrivare

MPA 2/2018

 Consiglio Nazionale delle Ricerche
Ufficio Stampa

L'intelligenza artificiale esplora il clima e trova conferme e novità

In una ricerca dell'Iia-Cnr, pubblicata sulla rivista Scientific Reports del gruppo Nature, reti di neuroni artificiali che apprendono il funzionamento del sistema climatico dai dati osservati nel passato – confermano le azioni umane come causa principale del riscaldamento globale recente e conducono a nuove scoperte sui cambiamenti climatici dell'ultimo secolo

Sistemi automatici per analisi.

Using deep learning to listen for whales

January 10, 2014 | categories: Python, Biology, Programming, Bioacoustics, Machine Learning

stemming from the fact that CIFAR-10 uses RGB images as input, while my spectrogram uses one real number value per *pixel*, not unlike gray-scale images.

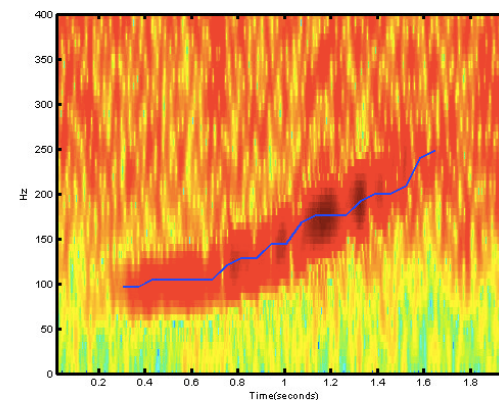
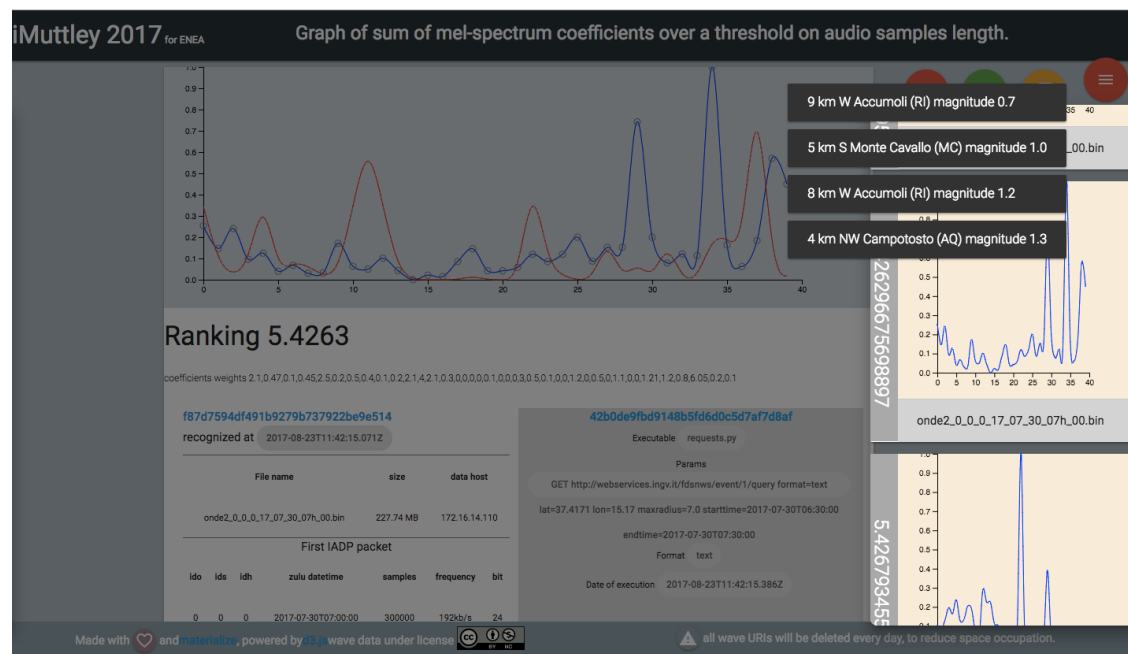


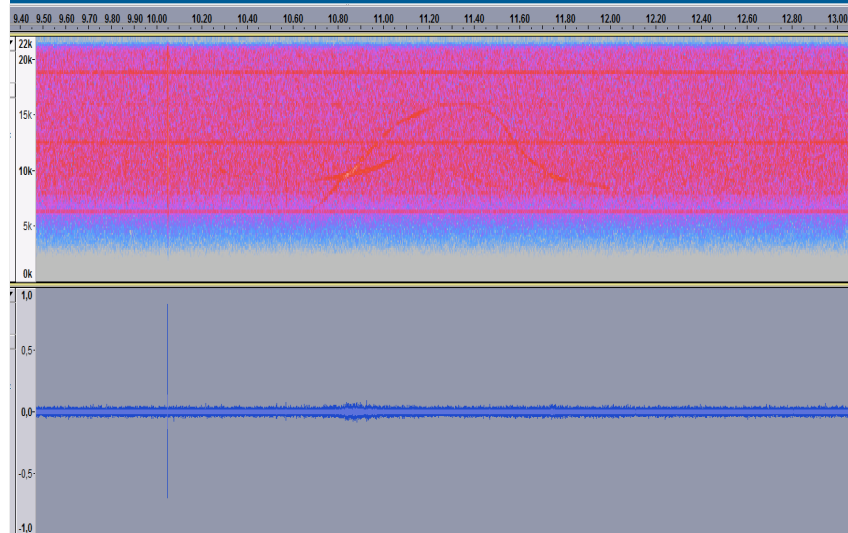
Figure 3: Spectrogram containing a right whale *up-call*.

Un portale per ricerca indicizzata

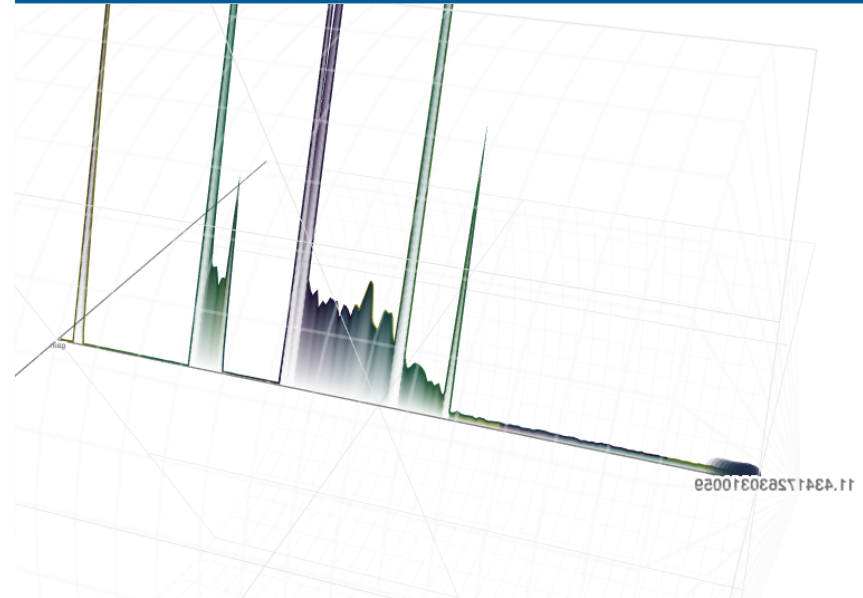


Un display in Virtual Reality per esplorare segnali audio

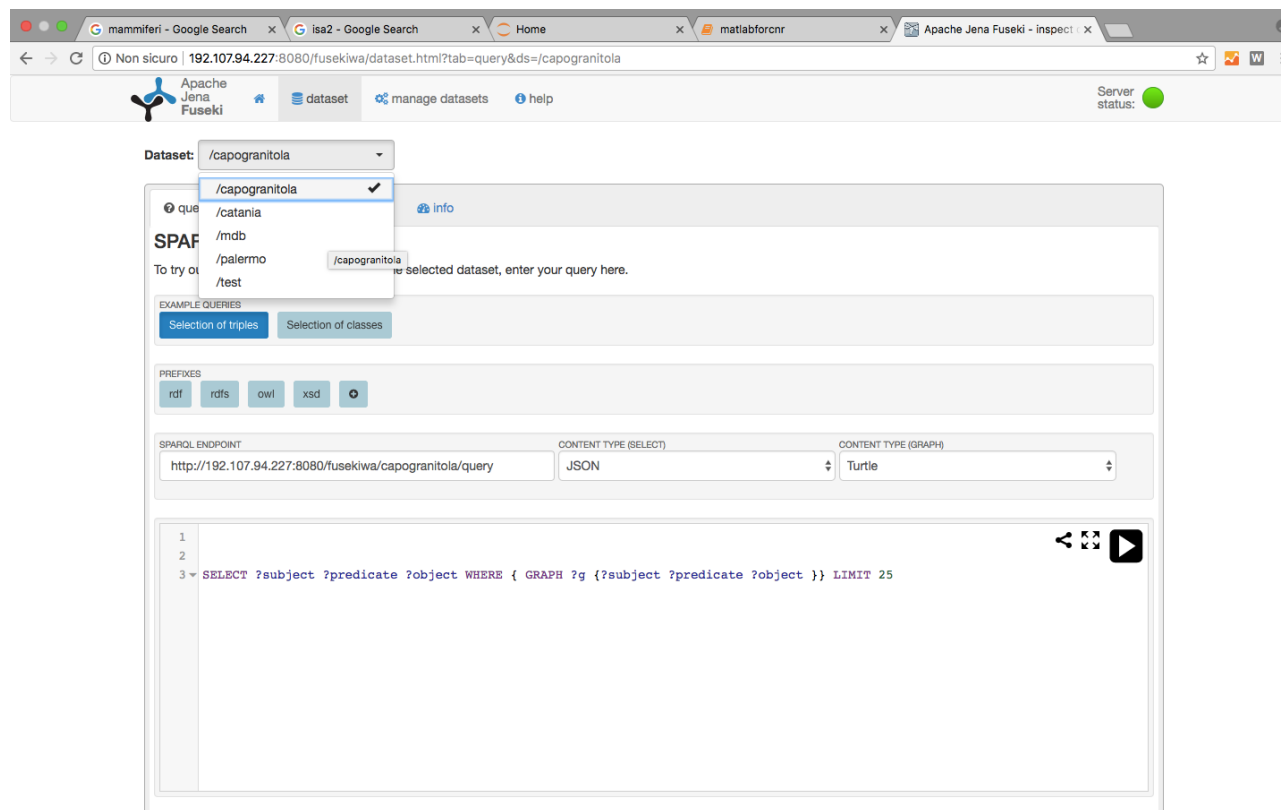
Da spettrogramma e segnale



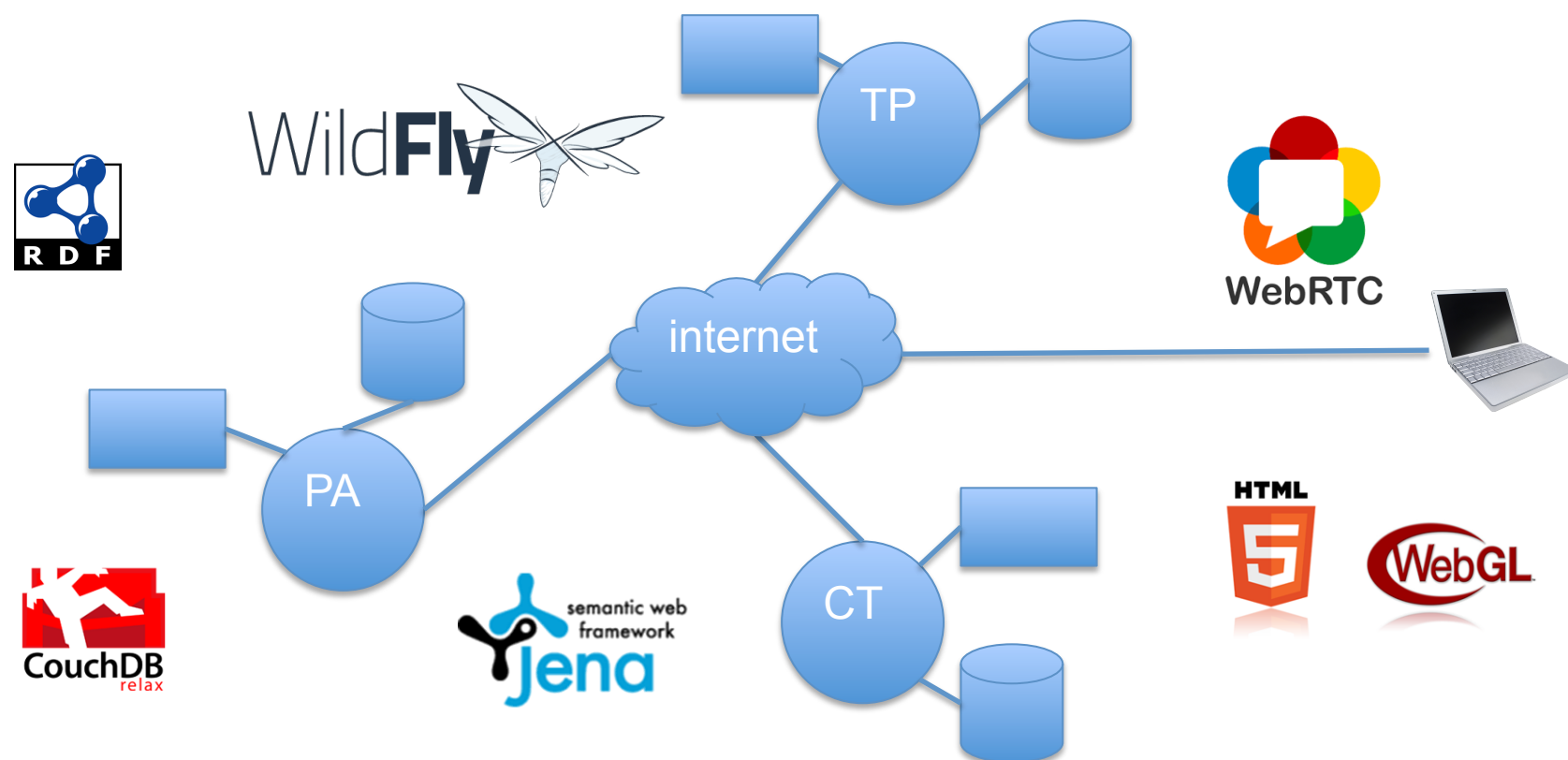
A spazio tridimensionale



Una base di conoscenza



Distribuzione dell'infrastruttura e tecnologie abilitanti



Che struttura si utilizza ?

Le informazioni estratte dal dato grezzo, sono smembrate e organizzate per nodi di un grafo. Ogni soggetto rappresenta un nodo in un grafo, e un predicato l'arco che lega due nodi, formando una struttura a grafo orientato.

Si consideri il seguente 'cammino' del grafo composto dal file, osservazione biologica e mammifero marino.



Il cammino direttamente espresso dai predicati del grafo, risponde alla domanda:
Quale specie osservo sul file x.bin ?

Convertita in query:

```
SELECT ?specie WHERE { ?s a file; nome x.bin; contiene_osservazione ?o. ?o specie ?specie }
```

Inferenza

Il percorso inverso ottenuto dalle leggi che regolano i predicati



Di risposta alla domanda:

Il delfino, in quale file è stato osservato ?

Risorsa

O 'soggetto', rappresenta il dato, ne descrive il contenuto nei suoi aspetti generali.

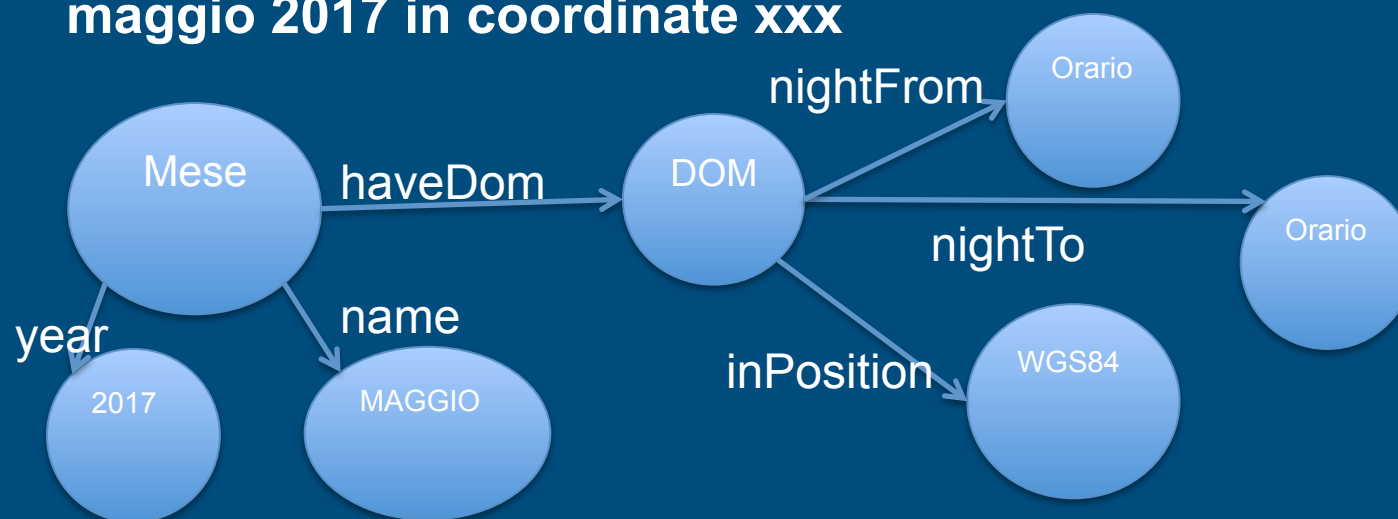
E' composto da un' insieme di attributi detti 'predicati' che referenziano altre risorse o 'oggetti'.

Esempio di istanza Risorsa

- Ha nome "Onde2_0_0_0_17_08_30_13h_00.bin"
- È "una sorgente di campioni audio"
- Ha formato "proprietario IADP"
- È stato inserito in catalogo in data "12-09-2017"
- **Contiene campionamenti di giorno 30**
- **Contiene campionamenti di mese "MAGGIO"**
- **Contiene campionamenti di anno "2017"**
- Autore "INFN"
- Distribuito con diritti "CC-BY-SA-NC"
- Si trova in IPV4 '172.16.14.110'
- Ha dimensione "238800000" Bytes sul filesystem
- Contiene "300000" pacchetti
- Di parametro IDO "0"
- Di parametro IDS "0"
- Di parametro IDH "0"
- Di parametro nbit "24"
- Di parametro fs "192"
- Portale o servizio di accesso al dato "onde2_0_0_..."
- Contiene misure "A" (istanze di algoritmi specializzati)
- **Contiene osservazioni "B" (istanze di genere biologico)**

Una domanda complessa

Elencami i delfini che sono presenti durante la notte nel mese di maggio 2017 in coordinate xxx



```
SELECT ?delfini WHERE {  
  GRAPH ?env { ?data a MESE ; name "MAGGIO"; year "2017"; haveDom ?dom; inPosition xx. ?dom nightFrom ?nf; nightTo ?nt }.  
  GRAPH ?infn{ ?o a OsservazioniBiologiche; specie ?delfini; data FILTER( Range(?data ?dom ?nf, ?data ?dom ?nt))}.
```



Il problema di base

Cosa stiamo cercando, nel senso di quali sono i domini in cui la nostra base di conoscenza si vuole estendere, ovvero quali sono i dati a nostra disposizione.

Risposte

A quali domande riesce a rispondere la base di conoscenza ?.

- A domande sui domini del dato prodotto (aggregabili con altri domini).

A chi deve fornire queste risposte ?

- Al ricercatore e al pubblico che debba ottenere risposte nel minor tempo possibile.

Per quale scopo fornirle ?

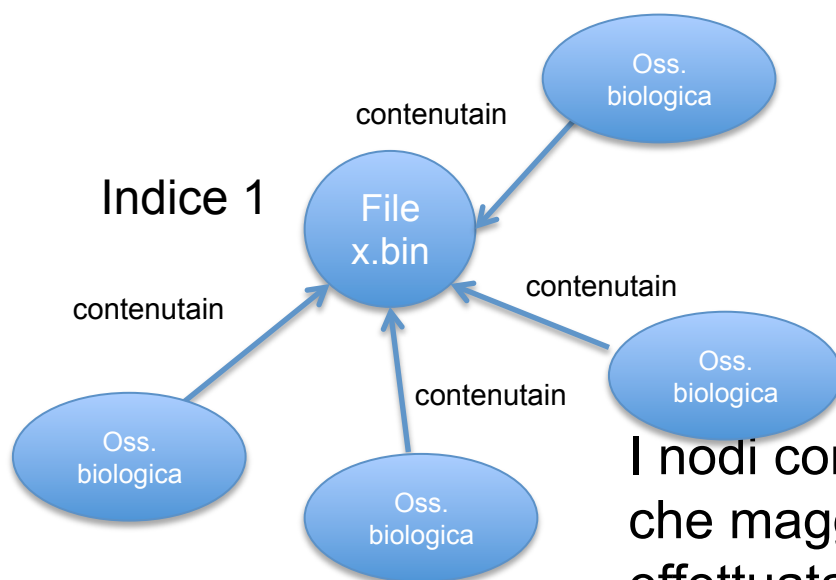
- Per accrescere la conoscenza

Come ottenerle ?

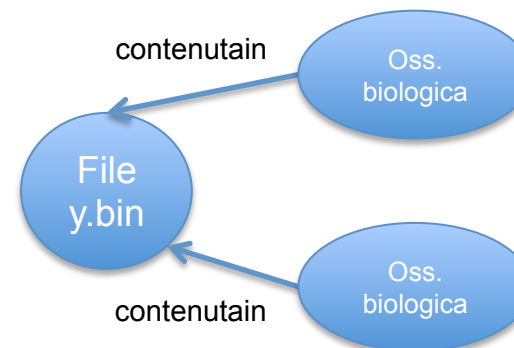
- Con ricerche per indicizzazione delle informazioni contenute nel dato (meglio ancora per linguaggio naturale).

Per indicizzare una ricerca ?

Si ordinano i nodi per il numero di archi che lo puntano.



Indice 2



I nodi con maggior numero di archi, sono quelli che maggiormente interessano alla ricerca effettuata.

Per concludere

I dati, se intesi come risorsa e sapientemente trattati, possono fornire un vasto spunto di applicazioni, in special modo nell'ambito della ricerca scientifica.

Quello che si propone è una infrastruttura tale da consentire la semplice e rapida ricerca delle informazioni necessarie, fornire le estensioni per le interazioni uomo-macchina, e i moduli per interrogazioni su ambienti automatizzati.

Dr. Tommaso Nicosia
Ing. Giuseppe Alonge



Nuovi strumenti per data mining e indicizzazione - Palermo – 18 gennaio 2018

GRAZIE PER L'ATTENZIONE.

