GlobalStat

Web Semantico - Primo elaborato

Boschi Francesco - 0000939879

26 marzo 2021

Indice

Capitolo 1

Introduzione

Nel 2010 l'Istituto Universitario Europeo [1] ha dato il via al progetto Global Governance Programme [2], il quale ha come obbiettivo principale la nascita di una comunità di professori, studenti e ricercatori di alto livello che contribuiscano e siano di supporto alle future generazioni in ambito politico e non solo.

GlobalStat [3] nasce, grazie ad una prima collaborazione con l'istituto Fundação Francisco Manuel dos Santos [4] che è stato di fatto il primo finanziatore, come sottosezione del Global Governance Programme.
Dal momento della sua nascita ad oggi, GlobalStat ha stabilito e intensificato nuovi programmi di partnership, in quanto riconosciuto come risorsa fondamentale in ambito pubblico e accademico. Dal 2016 GlobalStat è infatti partner con l'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCED) [5] con l'obbiettivo di produrre nuove tecniche di gestione e visualizzazione dei dati necessari ad analizzare e favorire lo sviluppo economico.
Anche lo stesso Servizio ricerca del Parlamento europeo [6] ha deciso di allacciare rapporti con GlobalStat per ottenere statistiche ed inforgrafiche utili allo staff del Parlamento e a tutti i paesi dell'Unione Europea uniti ai loro partner globali.

Attualmente GlobalStat coopera con oltre 80 istituzioni internazionali, come Eurostat, l'Organizzazione per il Cibo e l'Agricoltura e l'Organizzazione Internazionale del Lavoro, con le quali vi è un continuo scambio di dati bidirezionale; le organizzazioni, oltre a sfrtuttare gli strumenti messi a disposizione da GlobalStats, forniscono dati e statistiche utili a mantenere l'ecosistema aggiornato.

Capitolo 2

Obbiettivi

Come afferma Gaby Umbach, direttore della fondazione GlobalStats [3], la statistica ricopre un ruolo fondamentale in numerosi ambiti della nostra vita, da quello politico a quello sociale.
Tale importanza si traduce in un aumento della domanda di dati statistici, che siano affidabili e pubblicamente disponibili, fattore spesso assente e fondamentale, specialmente nell'era della globalizzazione nella quale la velocità alla quale i dati vengono prodotti rende impossibile la loro gestione e analisi in maniera manuale.
Le fitte interconnessioni derivanti dalla globalizzazione portano ad avere forte impatto in ambito sociale, personale, culturale, politico, economico e ambientale, rendendo i dati che GlobalStat raccoglie vitali per poter gestire al meglio le risorse e le possibilità che ciascun paese e ciascun individuo possiedono.

GlobalStat si pone quindi come obbiettivo quello di soddisfare questa esigenza di informazioni trasparenti e disponibili pubblicamente, così che possano essere reperite e visualizzate in maniera semplice e intuitiva da chiunque ne abbia necessità.
Considerando l'obbiettivo alla base del progetto e il carattere multidimensionale della globalizzazione, GlobalStat presenta dati su una vasta gamma di argomenti, ad esempio:

- Dati demografici
- Dati relativi all'economia
- Dati relativi all'energia e alle risorse naturali
- Dati relativi all'ambiente e all'inquinamento
- Dati relativi alla produzione nei vari settori

- Dati relativi alla libertà, ai conflitti e ai pericoli
- Dati governativi
- Dati relativi alla salute e alle condizioni di vita
- Dati relativi a fattori etici e morali, come crimini, giustizie, diritti, educazione, uguaglianza e condizioni lavorative
- Dati relativi alle migrazioni
- Dati relativi allo sviluppo tecnologico
- Dati relativi a trend e mode

Emerge quindi come GlobalStat non miri solo e migliorare il lato strettamente economico di un paese, ma anche a informare sul modo in cui gli uomini vivono, le libertà di cui godono e le restrizioni che devono affrontare quotidianamente.

2.1 Presentazione del sito WEB

Il sito web si presenta con un'interfaccia tanto semplice quanto chiara e intuitiva, che vede sulla parte sinistra il menù suddiviso in categorie e sottocategorie con tutti gli indicatori disponibili, mentre in quella destra voci per ottenere informazioni sul progetto.

Una volta selezionato l'indicatore, di default verranno mostrate le statistiche in modalità mappa, dalla quale è possibile selezionare lo stato di interesse per visualizzare le informazioni specifiche, oltre che applicare i filtri messi a disposizione.
La visuale ad istogramma rappresenta gli stessi dati sotto forma differente.

La modalità di visualizzazione trend permette di avere una visione chiara dell'andamento dell'indicatore con il passare degli anni, a livello globale o nel paese selezionato tramite gli appositi filtri.

L'ultima modalità di visualizzazione è certamente quella più riutilizzabile per sviluppare progetti esterni a GlobalStat, in quanto permette non solo di consultare, ma anche di scaricare i dati relativi all'indicatore selezionato dettagliati per ciascun paese e anno.
Il download viene fatto in formato xlsx, quindi sotto forma di foglio di calcolo Excell. Proprio per questo motivo, in questa sezione non sono sfruttate particolari tecnologie relative al web semantico.

Ciascuna sezione analizzata fino ad ora presenta inoltre un la classica icona delle informazioni: alla sua pressione un popup permette all'utente di ottenere informazioni dettagliate sul significato dell'indicatore, note aggiuntive e soprattutto la fonte dalla quale è stato reperito, la quale spesso, come analizzeremo nel capitolo 4, fa ampio uso delle tecnologie del web semantico.

Figura 2.1: Homepage del sito GlobalStat.

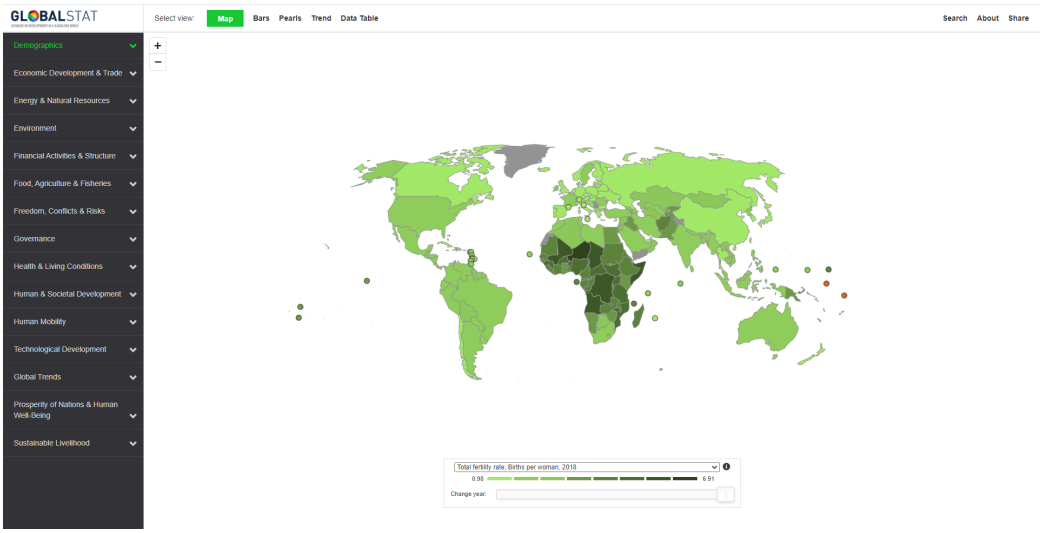

Figura 2.2: Mappa del sito GlobalStat.

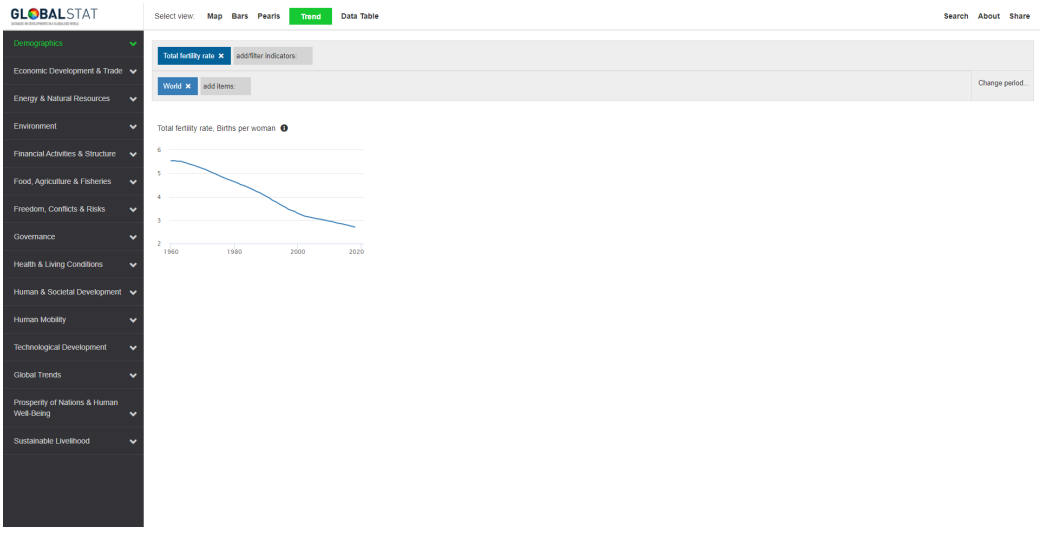

Figura 2.3: Trend del sito GlobalStat.

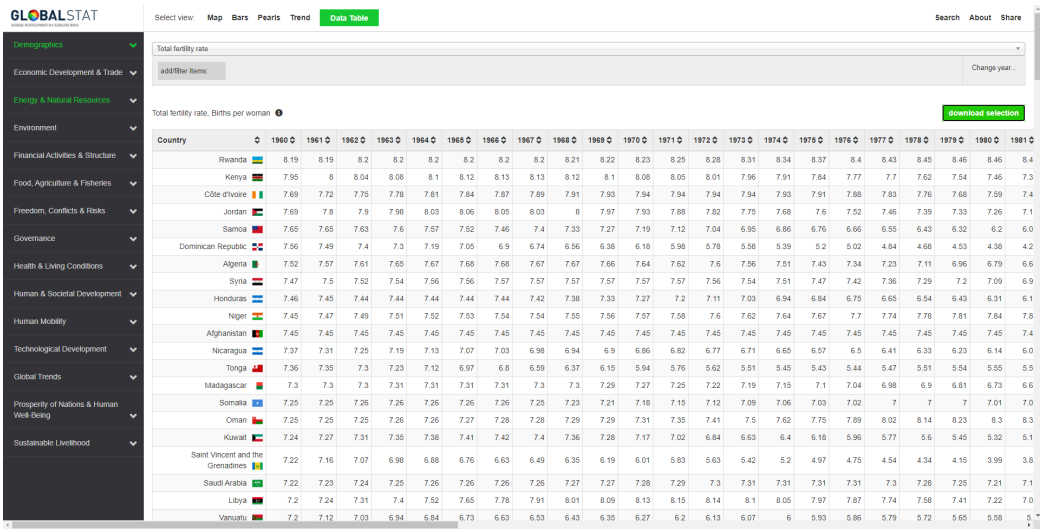

Figura 2.4: Tabella del sito GlobalStat.

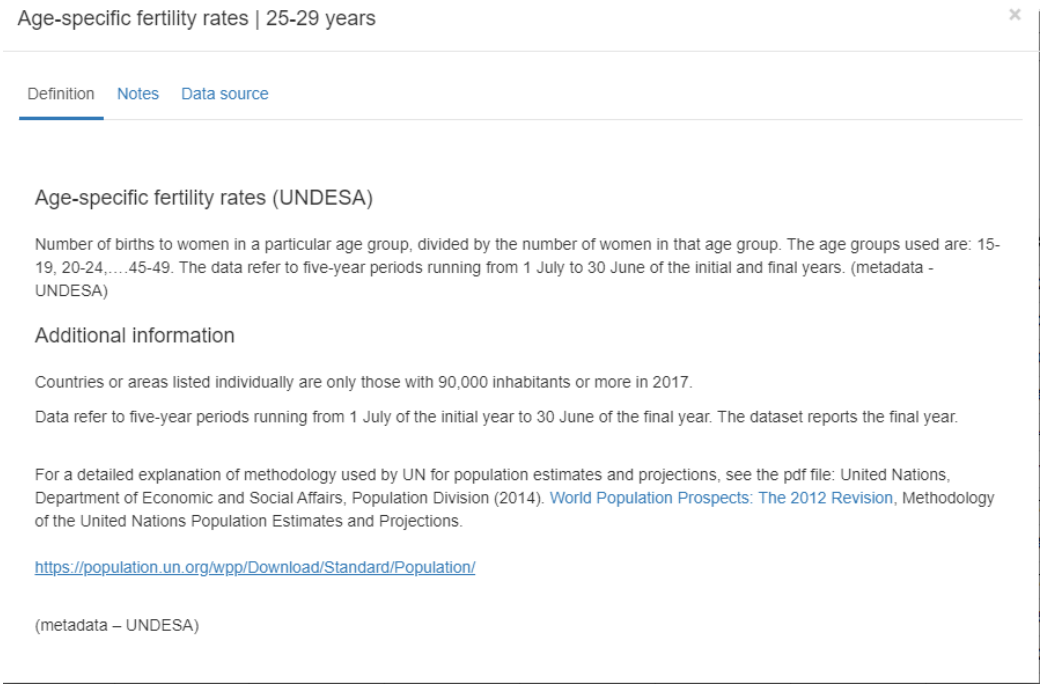

Figura 2.5: Info del sito GlobalStat.

Capitolo 3

Metodologie e sorgenti dati

I dati raccolti sono sensibili al carattere di vasta portata del progetto; per questo motivo è stato necessario provvedere apposite metodologie atte a produrre statistiche effettivamente utili e soprattutto il più possibile affidabili.
GlobalStats raccoglie dati sui 193 stati nazionali sovrani riconosciuti membri delle Nazioni Unite (ONU) [7] e, sulla base di questi, fornisce una panoramica sulle prestazioni dei singoli stati, dei continenti, di undici comunità di cooperazione e integrazione regionale e di organizzazioni internazionali.
Questa forte cooperazione con le comunità è il pilastro principale dell'aggregazione dei dati effettuata da GlobalStat, in quanto tramite esse gli stati si impegnano in una collaborazione regionale per sostenersi a vicenda e migliorare lo sviluppo e le condizioni di vita.

Sebbene tendenzialmente i confini tra stati siano statici nel tempo, per tenere traccia di eventuali separazioni o formazione di nuovi territori, GlobalStat calcola in maniera dinamica i dati aggregati adattandosi al formato con cui vengono resi disponibili. Esempio lampante in cui questa metodologia ha permesso di avere dati coerenti nonostante le vicende politiche è la dissoluzione della Cecoslovacchia avvenuta nel 1993.
Anche in fase di visualizzazione, quindi, i dati rispecchieranno di anno in anno la composizione degli stati facenti parte dell'ONU, rendendo più complesso il calcolo di medie e la possibilità di effettuare comparazioni nel tempo.

Seguendo la prassi di statistica internazionale, GlobalStat calcola e rende disponibile dati aggregati se sono soddisfatte due condizioni:

1. L'indicatore in questione è disponibile per più della metà dei membri del gruppo

2. Se la prima condizione è soddisfatta, il valore aggregato viene calcolato solo se la popolazione totale dei paesi per i quali sono disponibili i dati rappresenta almeno 2/3 della popolazione totale del gruppo

Indipendentemente da tali condizioni, seguendo un approccio prudenziale, i dati per Asia e Europa non vengono calcolati fino al 1992, anno dello scioglimento dell URSS.
Questa scelta è dovuta al fatto che i confini geografici dell'URSS erano condivisi tra due continenti, per i quali non è possibile dedurre il contributo dell'Unione in maniera isolata.

Se le condizioni 1 e 2 sono soddisfatte, GlobalStat applica metodologie di calcolo degli aggregati differenti dipendentemente dalla tipologia di dati:

- Gli aggregati dei dati dei singoli paesi, se espressi come valori assoluti, sono calcolati semplicemente tramite la somma ignorando i valori mancanti

- Gli aggregati dei dati dei singoli paesi, se espressi come valori rapporti (tassi/proporzioni/percentuali), sono calcolati come medie ponderate sulla base del conteggio della popolazione sul totale, ignorando i valori mancanti

GlobalStat si affida a ben 113 fonti differenti di dati dipendentemente dagli indicatori necessari e dai paesi di riferimento.
Pertanto, unendo questo fattore alle metodologie sopra elencate, quando un utente consulta GlobalStat, dovrebbe essere consapevole di questa peculiarità multi-fonte ed essere quindi cauto nel confrontare dati e indicatori differenti, poiché potrebbero potenzialmente variare e presentare incoerenze dovute all'intervallo temporale, alle pratiche di raccolta e dalla loro specificità.

Capitolo 4

Architettura e tecnologie utilizzate

GlobalStat, come analizzato in precedenza, funge principalmente da aggregatore di numerosissimi dati provenienti da sorgenti differenti, le quali possono essere a loro volta aggregatori piuttosto che semplici banche dati.
Nel primo caso GlobalStat produce le proprie statistiche a partire dai file riassuntivi tabulari, resi disponibili per esempio in formato .xlsx, facendo uso ridotto delle tecnologie del web semantico in maniera diretta essendo precedentemente sfruttate dall'aggregatore di riferimento.
In tutti gli altri casi, invece, nei quali le stime devono essere effettuate fondendo dati provenienti da fonti differenti, è fondamentale l'utilizzo di tecnologie per disambiguare il significato delle risorse e interconnetterle all'interno del web.

4.1 RDF, RDFS

RDF è la tecnologia alla base della sezione open data del progetto, fondamentale per descrivere e mettere in relazione tra loro le risorse all'interno del web tramite l'utilizzo di triple.
Tendenzialmente, in aggiunta ad RDF, viene utilizzato RDFS, che fornisce un ulteriore livello di conoscenza connettendo i concetti forniti tramite RDF alle primitive di corrispondenza (Class, Subclass, Property ecc.), permettendo così di fare inferenza e interrogare il dominio di appartenenza. Le seguenti analisi prenderanno come riferimento una risorsa messa a disposizione da World Bank [8], un'istituzione finanziaria internazionale che fornice dati a paesi di tutto il mondo ed è quindi una delle principali sorgenti dati sfruttate

da GlobalStat.

```
<?xml version="1.0" encoding="UTF-8"?>
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:content="http://purl.org/rss/1.0/modules/content/"
    xmlns:dc="http://purl.org/dc/terms/"
    xmlns:dcat="http://www.w3.org/ns/dcat#"
    xmlns:sioc="http://rdfs.org/sioc/ns#">

    <rdf:Description rdf:about="https://datacatalog.worldbank.org/
        ↪ dataset/global-economic-prospects">
      <rdf:type rdf:resource="http://rdfs.org/sioc/ns#Item"/>
      <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Document"/>
      <content:encoded>&lt;p&gt;The Global Economic Prospects</
          ↪ content:encoded>
      <dc:creator></dc:creator>
      <dc:Frequency>3</dc:Frequency>
      <dcat:granularity></dcat:granularity>
      <dc:license>notspecified</dc:license>
      <dc:accessRights></dc:accessRights>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:Distribution rdf:resource="https://datacatalog.wb.org/
          ↪ dataset/global-economic/resource/ID"/>
      <dcat:theme rdf:resource="https://datacatalog.wb.org/
          ↪ taxonomy_term/2031"/>
      <dc:temporal>2018 - 2022</dc:temporal>
      <dc:title>Global Economic Prospects</dc:title>
      <dc:date rdf:datatype="http://www.w3.org/2001/XMLSchema#
          ↪ dateTime">2017-01-31T00:28:40-05:00</dc:date>
      <dc:created rdf:datatype="http://www.w3.org/2001/XMLSchema#
```

```
        ↪ dateTime">2017-01-31T00:28:40-05:00</dc:created>
    <dc:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#
        ↪ dateTime">2021-02-09T00:14:01-05:00</dc:modified>
    <sioc:num_replies rdf:datatype="http://www.w3.org/2001/
        ↪ XMLSchema#integer">0</sioc:num_replies>
  </rdf:Description>
</rdf:RDF>
```

Il file XML in questione fa riferimento ad una specifica risorsa disponibile nel catalogo di WordBank, raggiungibile tramite l'apposito URI presente in **rdf:about**.
A partire dal tag **rdf:type** sono elencate tutte le proprietà della risorsa. Queste ultime rappresentano delle specifiche caratteristiche che, per essere riconosciute in maniera standard e inconfutabile, sono connesse ad uno specifico vocabolario che ne determina l'ontologia. Nel caso specifico vengono sfruttati tre dizionari differenti, ciascuno relativo ad un ambito ed in grado di disambiguare al meglio il significato della proprietà.

Il sistema di metadati Dublin Core [9] (identificabile dalla proprietà **dc:**) definisce i quindici elementi core per rappresentare le risorse, vale a dire "Creator", "Date", "Description", "Format" ecc.
Il sistema Data Catalog Vocabulary [10] (identificato da **dcat:**) è invece utilizzato per descrivere i dataset e i servizi all'interno di un catalogo utilizzando un modello e un vocabolario standard, così da facilitare l'aggregazione (fondamentale all'interno di GlobalStat) e la consultazione.
L'ultima specifica di ontologie utilizzata è la Semantically-Interliked Online Communities [11] (**sioc:**), che fornisce i concetti e le proprietà principali per descrivere informazioni relative a comunità online.

Questo esempio era solo un caso specifico selezionato tra le centinaia di sorgenti alle quali GlobalStat fa riferimento e alla miriade di risorse in esse contenute.
XML non è quindi l'unico formato disponibile; ciascuna sorgente sfrutterà i formati che ritiene più consoni, WordBank stesso permette di reperire i dati anche sotto forma di JSON, venendo così il più possibile incontro alle esigenze e alle preferenze dell'utilizzatore.
In maniera analoga anche la scelta dei vocabolari per le ontologie varia e viene fatta dipendentemente dagli elementi che devono essere rappresentati all'interno della risorsa e al grado di specificità richiesto dall'ente.

4.2 OWL

Sia DCAT che SIOC sono definiti sfruttando **OWL**, attualmente disponibile alle versione 2.
OWL 2 Web Ontology Language [12], abbreviato OWL 2, è un linguaggio per le ontologie utilizzato nel web semantico. Le ontologie forniscono classi, proprietà e valori che sono memorizzati sotto forma di documenti del web semantico.
Le ontologie fornite da OWL possono essere combinate tra di loro oltre che essere utilizzate per arricchire informazioni contenuto in documenti RDF.
OWL è un linguaggio computazionale basato sulla logica: la conoscenza espressa tramite esso può essere dedotta da software i quali potranno inoltre verificare la consistenza e fare inferenza di nuove informazioni.

Per quanto concerne DC, invece, essendo uno dei primi progetti nati in ambito ontologie, non è definito mediante OWL.
La differenza nel linguaggio rende più difficoltosa l'integrazione di DC con i dizionari più recenti; inoltre con il passare degli anni sono emerse numerose limitazioni come l'utilizzo di "LastName" e "FirstName" per identificare il creatore di una risorsa.
Possono infatti essere presenti numerosi casi di omonimia all'interno di un singolo database, oltre ad avere spesso difficoltà a disambiguare il nome dal cognome.
Per questi motivi, negli ultimi anni, stanno nascendo numerosi progetti atti a costruire versioni di DC basate su OWL, facilitandone così l'integrazione con ontologie definite ad hoc e disambiguando ulteriormente i termini [13].

Nonostante le numerose limitazioni analizzate, vista l'ampia diffusione dovuta alla longevità di Dublin Core, è tutt'ora uno dei vocabolari più diffusi per identificare i creatori di risorse.

Capitolo 5

Commenti finali

GlobalStat è sicuramente un progetto ambizioso, viste le difficoltà e i problemi che costantemente deve affrontare per portare statistiche affidabili, oltre che ammirevole in quanto si pone come obbiettivo principale quello di migliorare nel complesso la vita delle persone.
Quest'ultima, tra tutte le motivazioni, è stata quella che mi ha maggiormente spinto ad approfondire questo lavoro.
Attualmente GlobalStat funge principalmente da aggregatore: grazie alla sua folta rete di collaborazioni ottiene dati su scala globale, spesso a loro volta già precedentemente processati, e produce tramite essi statistiche che mette gratuitamente a disposizione degli utenti, principalmente in modalità visuale.

Non vi è quindi la possibilità di reperire i dati direttamente in maniera programmata, ma il sito stesso effettua un reindirizzamento alla sorgente originale, la quale adotterà poi delle specifiche politiche e tecnologie per l'accesso ai dati.
Sebbene per gli enti governativi e le organizzazioni internazionali i dati siano già accessibili in maniera automatica tramite un agente software, la speranza è che in futuro tale possibilità venga estesa a chiunque, permettendo così una diffusione di queste importanti informazioni su scala ancora più ampia.

Inoltre, come tendenzialmente accade in statistica, l'importanza e il valore cresce insieme alla quantità di dati disponibili: proprio per questo, negli anni a venire, i vantaggi che GlobalStat sarà in grado offrire emergeranno maggiormente ed è quindi fondamentale continuino ad siano accessibili in maniera semplice e automatizzata.

Bibliografia

[1] European university institute. Online; consultato il 16/03/2021.

[2] Global governance programme. Online; consultato il 16/03/2021.

[3] Globalstat. Online; consultato il 16/03/2021.

[4] Fundação francisco manuel dos santo. Online; consultato il 16/03/2021.

[5] Organizzazione per la cooperazione e lo sviluppo economico. Online; consultato il 16/03/2021.

[6] Servizio ricerca del parlamento europeo. Online; consultato il 19/03/2021.

[7] Stati membri delle nazioni unite. Online; consultato il 17/03/2021.

[8] Datacatalog worldbank. Online; consultato il 25/03/2021.

[9] Dublin core (dc). Online; consultato il 26/03/2021.

[10] Data catalog vocabulary (dcat). Online; consultato il 26/03/2021.

[11] Semantically-interlinked online communities (sioc). Online; consultato il 26/03/2021.

[12] Web ontology language (owl). Online; consultato il 26/03/2021.

[13] Dublin core and owl. Online; consultato il 26/03/2021.