31/5/2023

Report delle modalità di acquisizione e del formato dati

D.2.2.1 Fighting Cybercrime with OSINT

M. Gnaldi, L. Grilli, A. Milani, A. Navarra, M.C. Pinotti, V. Poggioni, F. Santini, C.Taticchi,
UNIVERSITÀ DEGLI STUDI DI PERUGIA

Sommario

Breve introduzione al deliverable

L'obiettivo di questo deliverable è quello di riassumere le principali fonti da cui poter attingere informazione OSINT da poter utilizzare per gli scopi del progetto Fighting Cybercrime with OSINT (FICO). In base alle fonti possono essere individuati anche i metodi di acquisizione dei dati ed il formato, in modo da individuare, in un seguente deliverable, i sistemi di memorizzazione (Database Management Systems, DBMS) in grado di memorizzarli in modo più appropriato. Ci occuperemo prima delle fonti aperte collegate al terrorismo internazionale (in particolare Jihadista), ed in seguito dei dati provenienti da fonti aperte riguardanti la corruzione.

Fonti collegate al terrorismo internazionale Jihadista

Una importante fonte è rappresentata da materiale propagandistico pubblicato online (magazine online) da gruppi terroristici. Per esempio, Dābiq (in arabo: دابق) è stata una rivista online pubblicata dallo Stato Islamico a scopo di propaganda. La rivista è stata pubblicata per la prima volta nel luglio 2014 in diverse lingue. Il primo numero riportava la data "Ramadan 1435" del calendario islamico. Data la sua natura, la rivista era disponibile, come in molti altri casi del genere, solamente utilizzando browser che consentono l'accesso e la navigazione nel Deep Web.[1]
Nel settembre del 2016, Dābiq ha cessato le sue pubblicazioni venendo sostituita con un'altra rivista online, Rumiyah (in arabo: Roma), pubblicata in inglese e in altre lingue. Gli analisti hanno ipotizzato che ciò fosse dovuto al fatto che l'ISIS ha perso ed è stato cacciato dal territorio, dopo l'offensiva a Dabiq da parte dell'Esercito siriano libero e dall'esercito Turco. Il titolo della testata che sostituisce Dābiq, si riferisce ad una profezia che riguarda una nuova caduta di Roma.[2]

La rivista è stata rilasciato in diverse lingue, tra cui inglese, francese, tedesco, russo, indonesiano e uiguro; ha soppiantato diverse altre riviste precedenti scritte in lingue differenti (come Dar al-Islam, Konstantiniyye e Istok, riportate nel seguito).

[1] Da Inspire a Dabiq, Ecco Come Nascono i Magazine Jihadisti su smartweek.it, 18 novembre 2015. https://web.archive.org/web/20160221151548/http://www.smartweek.it/da-inspire-a-dabiq-ecco-come-nascono-i-magazine-jihadisti/.
[2] After the Islamic State, su newyorker.com, 12 dicembre 2016, https://www.newyorker.com/magazine/2016/12/12/after-the-islamic-state.

In Tabella 1 sono riportati tutti i quindici numeri di Dābiq ed il link al loro pdf archiviato, compresa la data di pubblicazione. Il formato dati in questo caso è un documento pdf che contiene testo ed immagini mischiate insieme.

Tabella 1: I 16 numeri di Dābiq ed i link ai loro pdf archiviati. La tabella fa riferimento a link su Wikipedia.

Numero	Titolo originale	Periodo di pubblicazione
1	The Return of Khilafah[3]	5 luglio 2014
2	The Flood[4]	27 luglio 2014
3	A Call to Hijrah[5]	10 settembre 2014
4	The Failed Crusade[6]	11 ottobre 2014
5	Remaining and Expanding[7]	21 novembre 2014
6	Al Qa'idah of Waziristan: A Testimony from Within[8]	29 dicembre 2014
7	From Hypocricy to Apostasy: The Extinction of the Grayzone[9]	12 febbraio 2015
8	Shari'ah Alone Will Rule Africa[10]	30 marzo 2015
9	They Plot and Allah Plots[11]	21 maggio 2015
10	The Law of Allah or the Laws of Men[12]	13 luglio 2015
11	From the Battles of Al-Ahzāb to the War of Coalitions[13]	9 agosto 2015
12	Just Terror[14]	18 novembre 2015
13	The Rafidah from Ibn Saba' to the Dajjal[15]	19 gennaio 2016
14	The Murtadd Brotherhood[16]	13 aprile 2016
15	Break The Cross[17]	31 luglio 2016

Rumiyah è composto invece da 13 numeri, a partire dal 5 Settembre 2016 fino al 9 Settembre 2017, anno della cessazione della sua pubblicazione. Sia Dābiq che Rumiyah sono scritti in lingua inglese.

Sono presenti però altri esempi di magazine online. Dar al-Islam (arabo: دار الإسلام, romanizzato: Dār al-ʾIslām, lett. 'Casa dell'Islam') è il titolo di una rivista online in lingua francese prodotta dallo Stato

islamico (IS) tra il 2014 e il 2016. I dieci numeri della rivista sono stati rilasciati in totale e il progetto jihadology.net ha versioni inalterate che sono disponibili online.[18]
Altri esempi di rivista redatte per conto dello Stato Islamico consistono per esempio in Istok (lingua Russa) e Konstantiniyye (lingua Turca).

Un'altra fonte OSINT è rappresentata da **Youtube**, dove è possibile trovare video e canali di propaganda che possono essere monitorati al fine di interpretare messaggi, monitorare trend e anomalie durante successive pubblicazioni. Nella maggior parte dei casi, visto il medium in questione, i video consistono in propaganda sottoforma di musica (e testo/preghiera ovviamente). Alcuni video di esempio sono:

- https://www.youtube.com/watch?v=pMfTKK-OFjg&list=PLxNQ5etDv3wIiLBPqdI4SXYmxqapS0Ji_
- https://www.youtube.com/watch?v=qpzEk7WyYG4
- https://www.youtube.com/watch?v=GfNHXAkzWSs&list=PL-iPtbDIsIBpzdSCtJa0nv8NCbMlRtCMT
- https://www.youtube.com/watch?v=xn2JrjI2qT8
- https://www.youtube.com/watch?v=MPqQkc47kfU

Un'altra fonte consiste in articoli scritti nei quotidiani o riviste online ufficiali, in lingua italiana (ad esempio, Fatto Quotidiano, Repubblica, etc) o meno (ad esempio, New York Times). Queste riviste ovviamente non contengono messaggi dei gruppi terroristici ma li riportano solamente. La loro analisi può essere comunque importante se aggregata attraverso il tempo ed editori differenti in modo da capire anche in questo caso i trend.

In generale, le possibili fonti consistono nei social network come *Facebook*, *Twitter*, *Instagram*, *Reddit*, *Youtube*, *Tik Tok*, in relazione a commenti di reazione rispetto ad un determinato evento di propaganda o terrorismo pubblicato su un dato social, o in base anche a post con uno specifico tag o topic predeterminato.

Discussione

Per quanto riguarda i magazine online, l'acquisizione automatica al momento dell'uscita del numero risulta complicata a causa delle tecniche di anonimizzazione spesso utilizzate, come ad esempio reti Tor per l'accesso al Deep Web. Inoltre, l'automatizzazione risulta di scarso interesse visto il modesto numero di edizioni rilasciate (tra le 10 e le 20 in molti casi), caratteristica che rende meno interessante l'uso di crawler e scraper Web (o Deep Web).

Risulta invece interessante l'utilizzo di strumenti informatici per il trattamento automatico del pdf e la sua suddivisione automatica in testo e immagini, e conseguente loro analisi e classificazione con metodi di AI; ad esempio, Machine Learning, Natural Language Processing, Topic Analysis, Emotion Analysis, Sentiment Analysis, Opinion Mining, Argumentation, Analisi Statistica del testo ed in generale Algoritmi di analisi. In questo caso ci si concentra soprattutto su analisi di testo e immagini.

Queste informazioni possono quindi essere anche automaticamente messe in comparazione tra numeri differenti dello stesso magazine, in modo da misurarne trend e divergenze tra rilascio e rilascio.

[18] jihadology.net: https://jihadology.net/2016/08/20/new-issue-of-the-islamic-states-magazine-dar-al-islam-10/.

Uno strumento software commerciale utilizzato per l'analisi è rappresentato SPSS, un software di statistica sviluppato da IBM per la gestione dei dati, l'analisi avanzata, l'analisi multivariata, la business intelligence e le indagini penali. In precedenza era prodotto dalla SPSS Inc., che è stata acquisita da IBM nel 2009. Un tempo la sigla era acronimo di Statistical Package for Social Science perché nato nell'ambito delle scienze sociali. Col tempo sono state sviluppate numerose funzioni di statistica, tale da renderlo utile per qualsiasi ambito, non solo le scienze sociali, ma anche quelle mediche/epidemiologiche, economiche, demografiche, agrarie, di marketing ecc. Contiene anche strumenti di Machine Learning come Decision Tree e Neural Network.

Analisi automatizzate possono essere invece realizzate sui social network: per esempio, per quanto riguarda il magazine Rumiyah (9 numeri della rivista), possiamo menzionare uno studio dell'Europol del 2018.[19]
In questo studio, durante il periodo 1/11/2016-31/10/2017, sono stati raccolti tutti i post di Twitter che: (1) menzionavano il termine "Rumiyah"; (2) sono stati pubblicati entro 21 giorni dall'uscita di un nuovo numero; e (3) sono stati pubblicati da un account che utilizzava l'interfaccia in lingua inglese (Stati Uniti o Regno Unito). Ci si è concentrati infatti sugli utenti che hanno postato sulla versione in lingua inglese di Rumiyah (che è anche pubblicata in molte altre lingue). Sono stati raccolti i dati degli utenti pubblicamente disponibili, i dettagli del primo post di ciascun utente dopo l'uscita di un nuovo numero, la distribuzione dei messaggi successivi e lo stato dell'account (al termine del periodo di raccolta dei dati). È stato rilevato un totale di 9968 utenti distinti che hanno postato circa uno (o più) dei nove numeri di Rumiyah che sono stati esaminati. Si tratta di una media di 1108 utenti per numero. Un simile studio è stato fatto in [1] per quanto riguarda Dābiq. Questi studi sono stati eseguiti utilizzando Sentinel [2].

Per quanto riguarda i social network, i dati di alcuni di essi sono di difficile acquisizione proprio per come sono strutturati: quelli più vicino ad un sistema di messaggistica, come WeChat, WhatsApp e Telegram, necessitano l'accesso al gruppo, che in certi casi può essere riservato. Alcune varianti di questi social possono essere più liberamente accedute, come i canali di Telegram. In generale comunque, visti i numerosi scandali (ad esempio quello riguardante Cambridge Analytica e Facebook) e recenti cambi di direzione commerciale (dovuto all'acquisizione di Twitter da parte di Elon Musk), lo scraping di dati dai social network è al giorno d'oggi molto difficile se non avendo a disposizione le API messe a disposizione a pagamento dagli stessi social.

In ogni caso, sembra in diversi casi potrebbe essere necessario operare con lingue differenti, come per esempio, Inglese, Francese, Arabo, etc, oppure concentrarsi si alcuni tra questi linguaggi in modo più specifico. Ci potrebbe quindi essere la necessità di lavorare con modelli e librerie multi-lingua, oppure prevederla in un secondo momento.

Risorse software e documentazione

Simili considerazioni possono essere effettuate per i quotidiani e riviste online, che spesso (ma non sempre) sono (completamente o in parte) accessibili solamente attraverso un sistema di Paywall. Alcuni giornali online, come ad esempi il New York Times, offrono anche API per poter accedere ai metadati e contenuti degli articoli da programma.[20]

[19] Who disseminates Rumiyah? Examining the relative influence of sympathiser and non-sympathiser Twitter users: https://www.europol.europa.eu/cms/sites/default/files/documents/dgrinnell_smacdonald_dmair_nlorenzodus_who_disseminates_rumiyah_0.pdf.
[20] API del New York Times: https://developer.nytimes.com/apis.

In questo caso gli strumenti da utilizzare sarebbero Web scraper in grado di collezionare gli articoli giorno per giorno in modo automatico, in base al contenuto ed alcune keyword di interesse.
Per quanto riguarda le risorse disponibili per l'implementazione un modulo di Web scraping, possiamo individuare interi libri [3,4,5,6,7,8] e librerie dedicate, come ad esempio *Scrapy*[21] e *BeautifulSoup*[22] utilizzato insieme a *Requests*[23] per Python. Altre librerie, meno popolari ma sempre in Python, sono *Requests-HTML*[24], *Selectolax*[25], *pyspider*[26], e *AutoScraper*[27], mentre considerando altri linguaggi di programmazione come NodeJS, per lo stesso scopo esistono *Selenium*[28], *Puppeteer*[29] e *Playwriright*[30], anche se spesso sono previsti wrapper Python per queste librerie.

Youtube invece, e Google in generale, mette a disposizione i propri dati attraverso delle API interrogabili da programmi scritti in Go, Java, JavaScript, .NET, PHP, Python, e Ruby. Le API sono chiamate YouTube Data API (v3).
[31] Ciascuna chiamata consuma una certa quota a disposizione, che ha un limite corrispondente a 10000 unità di costo al giorno. Per progetti che necessitano un limite più alto, è necessario chiedere l'autorizzazione compilando una richiesta e descrivendo il progetto in questione.

L'uso delle API di Reddit[32] risulta essere ancora libero, anche se con un limite di 30 richieste al minuto. Dopo il recente cambio di Twitter, ci sono voci che parlano di un prossimo futuro cambiamento anche per Reddit, che potrebbe chiudere le proprie API e farne pagare l'utilizzo. In Python esistono delle librerie che implementano dei wrapper alle chiamate delle API di Reddit: le più famose sono *PSAW* (più idonea per accesso a dati storici) e *PRAW* (più idonea per operazione online, compreso postare).

Anche Twitter mette a disposizione le sue Twitter API v2. Per 100 dollari al mese è possibile recuperare 10000 tweet al mese. Gratuitamente invece non è possibile leggere nessun tweet, mentre si può scrivere via software fino ad un massimo di 1500 tweet al mese (sono possibili solamente operazioni di send/delete).

Meta fornisce invece API per Facebook, Instagram e WhatsApp. Per quanto riguarda Facebook, esistono vari servizi di scraping a pagamento, da 39 a 550 dollari al mese.[33] È possibile utilizzare librerie di scraping come facebook_scraper[34] che riescono a recuperare post e commenti ai post (ed altra informazione come numero di like e condivisione per esempio), in base al nome di una pagina (che ovviamente deve essere aperta al pubblico). Questo tipo di scraping è però soggetto a ban temporanei dell'IP della macchina che effettua lo scraping. Esiste infatti un team dedicato a queste rilevazioni ("external data misuse (EDM) team"). Lo scraping di pagine Facebook in questo modo

[21] Scrapy: https://scrapy.org.
[22] BeautifulSoup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
[23] Requests: https://pypi.org/project/requests/.
[24] Requests-HTML: https://github.com/psf/requests-html.
[25] Selectolax: https://github.com/rushter/selectolax.
[26] Pyspider: https://github.com/binux/pyspider.
[27] AutoScraper: https://github.com/alirezamika/autoscraper.
[28] Selenium: https://github.com/SeleniumHQ/selenium.
[29] Puppeteer: https://github.com/puppeteer/puppeteer.
[30] Playwriright: https://playwright.dev
[31] Youtube Data API: https://developers.google.com/youtube/v3/getting-started.
[32] API Reddit: https://www.reddit.com/dev/api/.
[33] Scraping di Facebook: https://research.aimultiple.com/facebook-scraping/.
[34] Facebook_scraper: https://pypi.org/project/facebook-scraper/.

viola i termini di servizio e potrebbe avere conseguenza legali. Le politiche API di Facebook non consentono lo scraping di pagine pubbliche senza un adeguato processo di revisione dell'app e un token di accesso. Pertanto, l'utilizzo dell'API Graph di Facebook sarebbe il modo più conforme per ottenere informazioni pubbliche. Infine è possibile utilizzare *Facebook Graph API* (che sono basate su HTTP): esse sono gratis, anche se sono limitate in vario modo nel rateo temporale di utilizzo. Come detto in precedenza però, la app deve essere verificata passando la fase di revisione del codice effettuata da un team di Facebook. Per recuperare dati pubblici la app deve avere *Page Public Content Access* (*PPCA*), che non è facile ottenere e viene spesso viene rifiutato (i dati devono comunque essere aggregati e anonimizzati). Le *Instagram Graph API*[35] e *Instagram Basic Display API*[36], che sono offerte anch'esse da Meta, funzionano con principi sostanzialmente differenti.
Le prime API fanno riferimento ad app costruite per conto di account "Businness" e "Creator", ed hanno delle limitazioni: per esempio, si possono eseguire query su un massimo di 30 hashtag univoci per conto di un account Instagram Business o Creator in un periodo continuo di 7 giorni. Esse sono pensate per creare delle app che gestiscono, analizzano e aggiungono contenuti per profili di questo tipo. In generale comunque, nessuna API di Instagram consente di recuperare informazioni di profilo, post, commenti, o contenuti di uno specifico utente. Anche in questo caso esistono dei servizi a pagamenti che lo permettono (come *Apify*).[37] Per costruire uno scraper (ovviamente con il rischio di ricevere un ban dell'IP da cui si effettuano le richieste) esistono alcune librerie, come per esempio *Instascrape*.[38]
Tutte le API di Meta Le API ritornano come risultato un documento in formato JSON.

Una soluzione potrebbe essere quella di stipulare degli accordi di utilizzo delle API con un account accademico. Le informazioni necessarie per applicare all'utilizzo accademico di Twitter prevendono la sottomissione di link di identificazione (ad esempio, un link al proprio Google Scholar e alla propria pagina presso Università e Dipartimento).[39] La richiesta può essere effettuata da studenti di Laurea (Master Thesis), di dottorato, oppure Ricercatori. In questo modo è possibile acquisire fino a 10 milioni di Tweet al mese, compreso tutto lo storico di Twitter. Al momento questa opzione sembra ancora essere resa disponibile da Twitter in questi termini, nonostante il cambiamento recente (primi mesi del 2023) sull'accesso alle API voluto da Elon Musk dopo l'acquisto di Twitter.
Anche Facebook, dopo i passati scandali, ha dedicato delle *Fort Researcher API* (dove *FORT* sta per *Facebook Open Research and Transparency*) al mondo accademico.[40] È possibile chiedere l'accesso alla Research Platform tramite form online.[41]

In generale parliamo comunque di informazione non-strutturata (video) oppure semi-strutturata (per esempio HTML, JSON), come semplice testo, metadati (spesso definiti da chiave-valore) oppure file (video/audio) rappresentati in diversi formati, che richiedono quindi database No-SQL (come per esempio, MongoDB, in cui la memorizzazione di documenti JSON per i metadati è diretta), le cui caratteristiche saranno oggetto di futuri deliverable di questo progetto (in particolare, il D.2.3.1).

[35] L'API non può accedere agli account consumer di Instagram (ovvero account Instagram non Business o non Creator).
[36] L'API Instagram Basic Display consente agli utenti della tua app di ottenere informazioni di base sul profilo, foto e video nei loro account Instagram. L'API può essere utilizzata per accedere a qualsiasi tipo di account Instagram, ma fornisce solo accesso in lettura ai dati di base.
[37] https://research.aimultiple.com/instagram-scraping/.
[38] Insta-scrape: https://pypi.org/project/insta-scrape/.
[39] Twitter per progetti accademici: https://developer.twitter.com/en/products/twitter-api/academic-research/application-info.
[40] Fort Researcher API: https://fort.fb.com/researcher-apis.
[41] Accesso piattaforma di ricerca di Facebook: https://fort.fb.com/intake.

Infine, facciamo presente che è possibile estrarre automaticamente immagini e testo da documenti pdf, come le sopracitate riviste di propaganda. Esistono diverse librerie Python in grado di svolgere il compito; ovviamente il miglior pacchetto deve essere valutato con esperimenti sul campo, per esempio direttamente su tutti i numeri di una di queste riviste (ad esempio Dābiq). In alcuni casi, immagini e testo sembrano essere abbastanza sovrapposti o vicini, considerando anche lo stile di impaginazione della rivista. Questo potrebbe portare ad alcuni errori di estrazione con questi pacchetti.

Fonti collegate alla corruzione

La Banca Dati Nazionale dei Contratti Pubblici (BDNCP) detenuta dall'Autorità Nazionale Anticorruzione (Anac) è la banca dati open più rilevante per la misurazione del rischio di corruzione. Si tratta di un patrimonio di grande valore per quantità dei dati contenuti, tale da permettere il calcolo di indicatori di rischio di corruzione con un estremo grado di dettaglio territoriale, settoriale e temporale, difficilmente replicabile altrove. La BDNCP contiene informazioni su ogni singolo contratto pubblico gestito da stazioni appaltanti italiane. Per tutti i contratti (indipendentemente dall'importo a base d'asta), la banca dati include le informazioni contenute nel bando di gara (codice identificativo CIG, tipo di procedura, importo a base d'asta, numero lotti, dati stazione appaltante, ecc.). Per i contratti sopra la soglia che ne impedisce l'affidamento diretto, la banca dati consente inoltre di tenere traccia di ogni singolo contratto lungo tutto il suo ciclo di vita, ovvero dalla pubblicazione del bando, passando per l'aggiudicazione e per le fasi di avvio, possibili varianti e fine contratto, arrivando fino alle fasi finali relative al collaudo dell'opera. La banca dati è disponibile per la consultazione (tramite cruscotto grafico), e inoltre permette di scaricare i dati in formato csv o JSON.

L'istituzione della Banca Dati Nazionale dei Contratti Pubblici (BDNCP)[42] è avvenuta presso l'ex-AVCP (dal 2016 ANAC), tramite il D. Lgs. n. 235 del 30 dicembre 2010, che all'art. 62-bis modificava e integrava il D. Lgs. n. 82/2005, noto come "Codice dell'Amministrazione Digitale" (CAD). Inserita ufficialmente tra le banche dati di interesse strategico nazionale, il suo compito è «favorire la riduzione degli oneri amministrativi derivanti dagli obblighi informativi ed assicurare l'efficacia, la trasparenza e il controllo in tempo reale dell'azione amministrativa per l'allocazione della spesa pubblica in lavori, servizi e forniture, anche al fine del rispetto della legalità e del corretto agire della pubblica amministrazione e prevenire fenomeni di corruzione»[43]. La "Legge Anticorruzione" n. 190 del 2012 ha imposto l'obbligo per le Pubbliche Amministrazioni di pubblicare nei propri siti web istituzionali i dati concernenti gli appalti pubblici, dati che devono essere trasmessi in conformità anche all'ANAC in formato digitale, la quale, a sua volta, deve provvedere alla loro pubblicazione su un sito web liberamente consultabile da tutti i cittadini. Con Delibera dell'ANAC n. 264 del 2018 contenente il "Regolamento concernente l'accessibilità dei dati raccolti nella BDNCP", sono state stabilite le modalità di accesso ai dati raccolti, ossia la totale libertà di accesso per chiunque nel rispetto della normativa in materia del trattamento dei dati personali, e il loro riutilizzo secondo le modalità di cui all'art. 7 del D. Lgs. n. 33/2013. Restano escluse dall'accesso libero, invece, le «annotazioni riservate inserite nel casellario informatico delle imprese di cui alla lettera g) dell'art. 3, il cui accesso resta regolamentato dalle specifiche disposizioni di settore»[44]. Durante lo stesso

[42] Banca Dati Nazionale dei Contratti Pubblici: https://dati.anticorruzione.it/superset/dashboard/appalti/.
[43] D. Lgs. n. 82/2005, art. 62-*bis*.
[44] Delibera ANAC n. 264/2018, art. 4.

anno, alla BDNCP è stato assegnato il primo posto nella competizione indetta dalla Commissione europea "Better Governance through Procurement Digitalization" 2018 – categoria "National Contract Register", essendosi distinta per completezza, integrità dei dati, interoperabilità, disponibilità di funzioni di accesso e analisi delle informazioni, governance e sostenibilità.

La BDNCP presenta la struttura architettonica di un data lake, nel quale confluiscono i dati trasmessi dalle stazioni appaltanti relativi alle diverse fasi del ciclo di vita di qualsiasi tipo di procedimento di appalto per contratti pubblici. Questi dati vengono poi incrociati e integrati con informazioni provenienti da un universo frammentato di fonti, tra cui: il casellario giudiziario, l'anagrafe tributaria, la Banca Dati Antimafia, l'INPS, le Camere di commercio, ecc.

Ad oggi, si conferma la possibilità di consultare liberamente la banca dati aggiornata attraverso il portale dei dati aperti dell'ANAC. Esso permette di scaricare il contenuto della banca dati e di visualizzare diverse infografiche nella sezione Analytics.

Nel dettaglio, la BDNCP è organizzata per schede, ovvero tabelle riferite a una specifica fase del procedimento di appalto. La chiave di collegamento tra esse è il "Codice Identificativo di Gara" (CIG), codice alfanumerico e unico, assegnato a ogni gara d'appalto. A seguire l'elenco delle schede (si veda a questo riguardo la pagina "Portale dati aperti dell'ANAC")[45].

- **Aggiudicatari**
  Descrizione: Gli aggiudicatari sono gli operatori economici che hanno vinto la gara, ai quali, a valle di alcune verifiche, verrà affidato il contratto. Se per uno stesso cig si hanno diversi aggiudicatari o gruppi di aggiudicatari questi sono distinti dall'identificativo di aggiudicazione.
- **Aggiudicazioni**
  Descrizione: Informazioni sull'aggiudicazione: gli aggiudicatari sono gli operatori economici che hanno vinto la gara, ai quali, a valle di alcune verifiche, verrà affidato il contratto.
- **Attestazioni SOA**
  Descrizione: L'Attestazione SOA è un documento, rilasciato da una Società Organismo di Attestazione a seguito di un'istruttoria in cui viene vagliato il possesso dei requisiti sulla base dei lavori svolti nel periodo precedente. L'attestazione serve all'impresa a comprovare, in sede di gara, la capacità di eseguire lavori appartenenti ad una certa categoria d'opera (Categoria) e fino ad un certo valore (Classifica).
- **Avvio Contratto**
  Descrizione: Informazioni comunicate dalla Stazione Appaltante riguardanti la fase iniziale del contratto.
- **Bando CIG**
  Descrizione: Il bando è un documento attraverso il quale la stazione appaltante rende pubbliche le informazioni su una procedura di selezione del contraente, determinando gli elementi dell'appalto o della procedura di gara ed invitando le imprese a presentare un'offerta entro un termine prefissato. I dataset BandiCIG contengono i dati essenziali sulle gare di valore superiore a 40.000E pubblicate nel periodo di riferimento.
- **Bando SMARTCIG**
  Descrizione: Il bando è un documento attraverso il quale la stazione appaltante rende pubbliche le informazioni su una procedura di selezione del contraente, determinando gli elementi

[45] Portale dati aperti dell'ANAC: https://dati.anticorruzione.it/opendata/#SHOW1.

dell'appalto o della procedura di gara ed invitando le imprese a presentare un'offerta entro un termine prefissato. I dataset SmartCIG contengono i dati essenziali sui contratti modico valore (<40k€) oppure non sottoposti agli obblighi di comunicazione rendicontati dalle stazioni appaltanti nel periodo di riferimento.

- **Categorie DPCM Aggregazione**
  Descrizione: Informazioni sulle categorie merceologiche, per i controlli ai sensi dell'articolo 9, comma 3, del decreto-legge 24 aprile 2014, n. 66, convertito, con modificazioni, dalla legge 23 luglio 2014, n. 89 (aggregazione della spesa pubblica).
- **Categorie Opera**
  Descrizione: Informazioni sulla tipologia di opera, per appalti di Lavori.
- **Centro Di Costo**
  Descrizione: Unità organizzativa della stazione appaltante. In linea di principio un Centro di Costo è l'unità organizzativa alla quale il contratto è destinato.
- **Collaudo**
  Descrizione: Informazioni comunicate in seguito al collaudo finale dell'opera o del prodotto/servizio.
- **CUP**
  Descrizione: Il Codice Unico di Progetto (CUP), assegnato dal Dipartimento Programmazione Economica della Presidenza del Consiglio è il codice che identifica un progetto d'investimento pubblico. La sua richiesta è obbligatoria per tutta la "spesa per lo sviluppo", inclusi i progetti realizzati utilizzando risorse provenienti da bilanci di enti pubblici o di società partecipate, direttamente o indirettamente, da capitale pubblico e quelli realizzati con operazioni di finanza di progetto, "pura" o "assistita", o comunque che coinvolgono il patrimonio pubblico, anche se realizzati con risorse private.
- **Fine Contratto**
  Descrizione: Informazioni comunicate dalla Stazione Appaltante dopo il termine del contratto.
- **Fonti di Finanziamento**
  Descrizione: Informazioni sulle fonti di finanziamento dell'appalto, ovvero sulla provenienza delle risorse economiche con le quali verranno pagati i fornitori.
- **Lavorazioni**
  Descrizione: Informazioni sulle lavorazioni comprese nel contratto.
- **Modalità Realizzazione**
  Descrizione: codelist delle modalità di realizzazione del contratto.
- **Pubblicazioni**
  Descrizione: I bandi di gara devono essere pubblicati sul sito web dell'amministrazione (il cosiddetto Profilo del Committente) ed a livello locale, regionale, statale o europeo a seconda del loro valore e della Amministrazione che espleta la gara. Nel dataset pubblicazioni vengono riportate le date in cui il bando è stato pubblicato.
- **Quadro Economico**
  Descrizione: Nel quadro economico confluiscono tutte le voci di costo di un'opera pubblica o di un appalto.
- **Sospensioni**
  Descrizione: Informazioni sulle eventuali sospensioni durante l'esecuzione dei lavori o del contratto.
- **Stati di Avanzamento**
  Descrizione: Informazioni sulle fasi intermedie del contratto, inviate dalla stazione appaltante all'ANAC con cadenza periodica per contratti di entità rilevante.
- **Stazione Appaltante**

Descrizione: Set di informazioni sulla Stazione Appaltante. La Stazione Appaltante è una persona giuridica, che può essere una Amministrazione o un Ente Pubblico o un soggetto che riceve un finanziamento pubblico e deve espletare una gara, per soddisfare i propri fabbisogni o quelli di altre amministrazioni.

- **Subappalti**
  Descrizione: Il subappalto è il contratto con cui l'appaltatore affida ad un altro soggetto (subappaltatore) una parte dell'appalto che gli è stato affidato.
- **Tipo Fattispecie Contrattuale**
  Descrizione: codelist del Tipo Fattispecie Contrattuale.
- **Tipologia Scelta del Contraente**
  Descrizione: codelist delle Tipologia Scelta del Contraente.
- **Varianti**
  Descrizione: Informazioni su eventuali variazioni autorizzate durante l'esecuzione rispetto al contratto originario, in seguito a circostanze impreviste.

Discussione

Anche in questo caso, il formato dei dati utilizzato dalla banca dati è JSON (oltre che csv), per il quale possiamo prevedere di utilizzare un DBMS come MongoDB. Tale valutazione sarà comunque oggetto di deliverable successivi a questo (in particolare, D.2.3.1).

Bibliografia

[1] Grinnell, D., Macdonald, S. & Mair, D. (2017). The Response of, and on, Twitter, to the Release of Dabiq Issue 15. Paper presented at the 1st European Counter Terrorism Centre conference on online terrorist propaganda, 10-11 April 2017, The Hague. https://www.europol.europa.eu/publications- documents/response-of-and-twitter-to-release-of-dabiq-issue-15

[2] Preece, A., Spasic, I., Evans, K., Rogers, D., Webberley, W., Roberts, C., & Innes, M. (2018). 'Sentinel: A Codesigned Platform for Semantic Enrichment of Social Media Streams'. IEEE Transactions on Computational Social Systems, 5(1), 118–131. https://doi.org/10.1109/TCSS.2017.2763684

[3] Mitchell, Ryan. Web scraping with Python: Collecting more data from the modern web. " O'Reilly Media, Inc.", 2018.

[4] Heydt, Michael. Python Web Scraping Cookbook: Over 90 proven recipes to get you scraping with Python, microservices, Docker, and AWS. Packt Publishing Ltd, 2018.

[5] Kouzis-Loukas, D. (2016). Learning scrapy. Livery Place: Packt Publishing.

[6] Chapagain, A. (2019). Hands-On Web Scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others. Packt Publishing Ltd.

[7] Smith, V. (2019). Go Web Scraping Quick Start Guide: Implement the power of Go to scrape and crawl data from the web. Packt Publishing Ltd.

[8] Mitchell, R. (2013). Instant web scraping with Java. Birmingham, AL: Packt Publishing.