DOI: 10.26314/GARR-Conf17-proceedings-30

# SemplicePA: SEMantic instruments for PubLIc administrators and CitizEns

Martina Miliani[1], Anna Gabbolini[1], Lucia C. Passaro[2], Francesco Sandrelli[1], Alessandro Lenci[2] , Roberto Battistelli[1]

[1]Eti3 s.r.l., [2]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica Università di Pisa

**Abstract.** La trasformazione digitale italiana sta procedendo a rilento rispetto a quella Europea, con un digital divide che penalizza soprattutto i comuni più piccoli e gli open data che faticano ad essere pienamente valorizzati, con il risultato che il Foia è ancora ben lungi dall'essere applicato come dovrebbe. Eppure non mancano le iniziative civiche, alcune stimolate dalle stesse amministrazioni. Così come non mancano le tecnologie di eccellenza, sviluppate all'interno di start-up che collaborano con università statali e centri di ricerca. SemplicePA nasce in questo contesto con lo scopo di fornire uno strumento utile alla cittadinanza e alle amministrazioni a partire dall'analisi semantica computazionale di un archivio spesso sconosciuto, l'Albo Pretorio.

**Keywords.** Keywords. Conservazione e condivisione dei dati, Natural Language Processing, Machine Learning, Big Data.

## Introduzione

Otto paesi su dieci, in Europa, hanno attivato una regolamentazione sugli open data. L'Italia, che si trova sotto la media europea, è tra i "follower" delle buone pratiche, con un "Mezzogiorno notevolmente indietro" [1]. Tra le cause, il grande divario tra le piccole e le grandi amministrazioni. In base all'osservatorio dell'Istat, le province autonome e l'85,5% dei comuni sopra i 60.000 abitanti possiedono un ufficio dedicato all'Information & Communication Technology (ICT), ovvero poco più dell'1% del totale dei comuni [2]. Anche i risultati del primo monitoraggio sull'applicazione del Freedom of Information Act (Foia) sono tutt'altro che positivi: il 73% degli utenti non ha ricevuto risposta e un diniego su tre era invece illegittimo [3]. Anche per questo, accanto all'Agenzia per l'Italia Digitale (AgID), sta lavorando il Team per la Trasformazione Digitale che ha una diversa concezione delle informazioni possedute dalla PA: "I dati sono nostri e li gestiamo insieme" [4]. Eppure in Italia si registrano già alcune iniziative sull'uso degli open data. Talvolta sono gli stessi enti pubblici a indire contest per premiare il migliore utilizzo dei dati aperti: a vincere l'hackathon sul tema della disabilità indetto dal Comune di Lecce, è stato il censimento delle barriere architettoniche in città [5]. Inoltre, sono tantissime anche le iniziative civiche, come il progetto di crowdfunding Ricostruzione Trasparente [6], il cui obiettivo è quello di tenere traccia di tutti gli atti pubblici che consentano di esercitare un controllo diffuso sugli attori della ricostruzione in seguito al terremoto del 2016 avvenuto nel Centro Italia. Si tratta di dati rilasciati dalle pubbliche amministrazioni, che sono stati poi rielaborati

e resi disponibili ai cittadini in modo del tutto nuovo. Purtroppo, sono ancora tante le risorse non adeguatamente valorizzate, come ad esempio, l'Albo Pretorio, l'archivio degli atti di ciascun comune.

La prima legge che sancisce la trasformazione digitale dell'Albo Pretorio risale al gennaio 2009 e giunge a pieno regime nel 2013 [7]. Nonostante la trasformazione sia avvenuta, cercare un atto all'interno dell'Albo sembra possibile soltanto conoscendo in anticipo il documento stesso: gli atti sono ricercabili solo in un arco di tempo ristretto, in genere 15 giorni, e nei siti di molti comuni per recuperare un provvedimento è necessario conoscerne la data, l'organo che lo ha emanato, l'oggetto o il suo numero identificativo. A mancare sono soprattutto le relazioni tra i singoli documenti, non solo tra quelli di uno stesso comune, ma anche, e soprattutto, tra comuni differenti.

## 1. Un motore di ricerca semantico

SemplicePA [8] è nato nel 2015 con l'obiettivo di valorizzare i contenuti degli atti registrati nell'Albo Pretorio di tutti i comuni italiani, di rendere navigabili le informazioni e soprattutto le relazioni che tra esse intercorrono. Il fine è quello di creare un Albo Pretorio Nazionale che consenta la navigazione dei contenuti attraverso un motore di ricerca semantico, in grado di estrarre informazioni significative, quali nomi di aziende, organizzazioni, persone e luoghi e mostrare le loro relazioni attraverso strumenti di visual analytics.

In Italia, Cogito [9] si basa sull'analisi semantica dei testi sfruttando un'ampia banca dati che vede raggruppate più ontologie differenti, costruite anche in diverse lingue. È nato invece all'Università di Pavia lo strumento Facility Live, che mostra il suo valore nei domini più ristretti: l'ontologia dietro a motori di ricerca come questo è molto più "specializzata" e per questo anche precisa e puntuale nel recupero delle informazioni richieste dall'utente [10]. Légilocal è un motore di ricerca semantico che in Francia si occupa della gestione degli atti, dedicando anche un framework apposito per la loro stesura, in modo che siano facilmente leggibili ed elaborabili dal motore di ricerca (Amardeilh, 2013). Sugli enti locali in Italia si è specializzato il motore di ricerca Sophia Semantic Search, che riconosce le entità elencate all'interno dei documenti e li classifica per similarità [11]. SemplicePA è un ambiente dotato di vari componenti in grado di arricchire i documenti amministrativi con diversi tipi di informazioni semantiche che ne consentono, oltre che l'indicizzazione, anche la codifica, la ricerca e la navigabilità, in una prospettiva del tutto nuova e in linea con il paradigma di trasformazione digitale sostenuto da AgID. I componenti principali del sistema sono descritti di seguito.

## 2. Estrazione delle entità

All'interno di ogni documento sono individuate diverse entità: persone, luoghi, aziende, organizzazioni, importi, date e indirizzi email ma anche elementi più specifici dei provvedimenti amministrativi come riferimenti legislativi e ad altri atti, partite iva, codici identificativi di gara e codici fiscali.

Al fine di estrarre questi nuclei informativi, il sistema sfrutta un modulo di analisi linguistica del testo (Dell'Orletta et al. 2014) e l'estrazione della terminologia (Passaro e

Lenci, 2016). L'estrazione vera e propria delle entità avviene integrando due approcci diversi, uno basato su regole e un altro su modelli di "machine learning". L'approccio a regole è basato su algoritmi che contengono precise espressioni regolari sull'estrazione. Ad esempio una porzione di testo sarà estratta e classificata come partita iva se costituita da un codice di undici cifre che rispetta precisi parametri. L'altro approccio, descritto in dettaglio in Passaro et al., 2017, è stato sviluppato da Eti3 in collaborazione con il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa. In questo caso, la probabilità che i termini estratti siano delle entità è dedotta dalla distribuzione delle parole all'interno di un corpus di training annotato manualmente con le entità rilevanti per il dominio della pubblica amministrazione. Infine, un modulo di "normalizzazione" si occupa di riportare le varie entità a una forma univoca standard per astrarre rispetto alle forme grafiche in cui una stessa entità viene citata all'interno dei vari documenti.

## 3. Ontologia

L'ontologia su cui si basa SemplicePA è costruita con un metodo bottom-up e top-down, da un lato attraverso l'individuazione automatica dei termini di dominio (Passaro e Lenci, 2016) e la loro espansione sfruttando metodi di semantica distribuzionale (Baroni e Lenci, 2010), e dall'altro attraverso la loro classificazione da parte di esperti di quel dominio. Inoltre, i documenti sono stati organizzati per aree tematiche sfruttando algoritmi di Topic Modeling basati su Latent Dirichlet Allocation (Blei, 2003; Blei 2012), un modello generativo bayesiano in grado di individuare gli "argomenti latenti" dei documenti, che sono rappresentati da una distribuzione di probabilità delle parole e dei documenti. I testi, quindi, vengono classificati sia in base alla presenza dei termini dell'ontologia, sia in base agli argomenti estratti automaticamente sfruttando LDA, che permette di cogliere le diverse aree tematiche degli atti amministrativi.

## 4. Network Analysis

Le relazioni tra le entità presenti nei documenti sono calcolate dalla piattaforma sulla base della compresenza all'interno degli atti. Una sezione è appositamente dedicata alla visualizzazione di reti in cui i nodi sono le entità estratte, le relazioni sono gli archi che le collegano e il peso degli archi è dato dal numero di documenti in cui le entità collegate sono presenti contestualmente. Si può partire da una persona, un'organizzazione o un'azienda

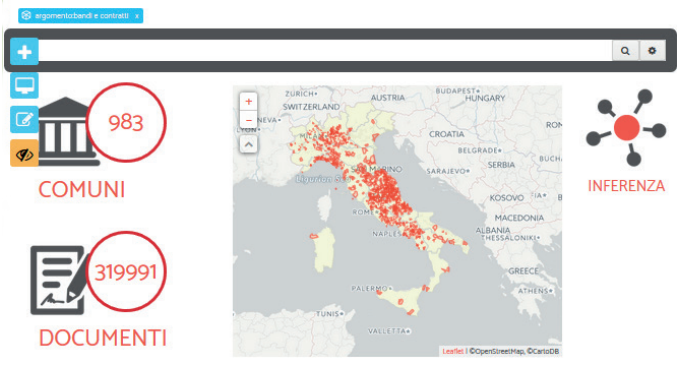



Fig. 1
La homepage di SemplicePA: un esempio di ricerca sui documenti processati associati all'argomento "bandi e contratti".

e decidere di visualizzare in una rete diversi tipi entità ad essa "collegate" (ancora aziende, persone, organizzazioni o atti stessi). In alternativa, è possibile visualizzare le relazioni tra gli elementi a partire da un gruppo di documenti selezionati.

## 5. Altri Strumenti

Tra gli altri strumenti offerti dalla piattaforma, una mappa, caricata automaticamente da OpenStreetMap [12] all'invio di ciascuna query, segnala i comuni e i luoghi citati all'interno dei documenti (Figura 1); le entità che appaiono all'interno dei documenti restituiti all'utente sono ordinate per frequenza, in modo da fornire una panoramica generale dei contenuti. Inoltre, per ciascun documento, sono rappresentati graficamente in una rete i riferimenti agli altri atti, allo scopo di ricostruire l'iter di pubblicazione dei documenti che si rifanno ad un unico procedimento amministrativo. Per una maggiore navigabilità, in fondo alla pagina di consultazione dell'atto sono presentati i documenti simili e una sezione di visual analytics mostra i trend delle pubblicazioni degli atti nel tempo, in base ai vari argomenti. Uno strumento aggiuntivo, infine, mette in contatto gli utenti connessi alla piattaforma attraverso una chat.

## 6. Conclusioni

SemplicePA nasce per valorizzare gli atti amministrativi di tutta Italia, grazie all'applicazione delle tecnologie del linguaggio più innovative che rendono le informazioni contenute nei documenti strutturate e navigabili. SemplicePA consente di consultare con facilità gli atti pubblicati anche ai comuni cittadini, mentre agli addetti ai lavori fornisce strumenti di analisi e di visualizzazione di dati che consentono una maggiore comprensione della realtà amministrativa. Gli strumenti di SemplicePA migliorano l'efficienza e l'organizzazione dell'ente in maniera del tutto automatica, e forniscono concretezza a obiettivi fondamentali quali trasparenza, Foia e anticorruzione. Un modo nuovo di interpretare la digitalizzazione della PA, che genera nuova conoscenza e la mette a disposizione di cittadini, amministratori e funzionari, perché siano più partecipi e più efficienti. L'innovazione tecnologica di SemplicePA contribuisce a più importanti innovazioni di carattere culturale, politico e sociale che dal territorio possono determinare un virtuoso miglioramento della PA.

## Riferimenti bibliografici

Amardeilh F., Bourcier D., Cherfi H., Dubai, C.H., Garnier A., Guillemin-Lanne S., Mimouni N., Nazarenko A., Paul È., Salotti S., Seizou M., (2013), The Légilocal project: the local law simply shared, JURIX, PP 11-14.

Baroni M., & Lenci A. (2010). Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4), PP 673-721.

Blei, D. M., Ng A. Y., Jordan I. M., (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, (3), PP 993-1022.

Blei, D. M. (2012). Probabilistic Topic Models, Communications of the ACM, (55:4), PP 77-84.

Dell'Orletta F., Venturi G., Cimino A., & Montemagni S. (2014). T2k2: a system for automatically extracting and organizing knowledge from texts. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014).

Passaro L. C., Lenci A. (2016), Extracting Terms with EXTra, Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives, Geneva, Editions Tradulex, PP 188-196.

Passaro L. C., Gabbolini A., Lenci A. (2017) INFORMed PA: A NER for the Italian Public Administration Domain. In Proceedings of the 4th Italian Conference on Computational Linguistics (CLiC-it 2017). Forthcoming.

1. Luca Tremolada, "L'Europa dei dati. Otto paesi su dieci hanno regole sugli open data, Il Sole 24 Ore, 5 Aprile 2017 (http://goo.gl/ZqVAfG) [Accesso: 09/11/2017].
2. Istat, "Le tecnologie dell'informazione e della comunicazione nella pubblica amministrazione locale", 2015 (http://goo.gl/7q4Loq) [Accesso: 09/11/2017].
3. Diritto di Sapere, "Ignoranza di Stato" (http://goo.gl/K4Cgx1) [Accesso: 09/11/2017].
4. Raffaele Lillo, "Data & Analytics Framework", Medium, 13 Febbraio 2017 (http://goo.gl/CtQXNy) [Accesso: 09/11/2017].
5. Piersoft, "Lecce, Luoghi accessibili per disabilità varie e di interesse", Umap, 30 Aprile 2016 (http://goo.gl/NJq6g7) [Accesso: 09/11/2017].
6. Ricostruzione Trasparente (http://ricostruzionetrasparente.it) [Accesso: 09/11/2017].
7. Qualità PA, "Albo Pretorio Online" (http://goo.gl/LxNLV7) [Accesso: 09/11/2017].
8. SemplicePA (http://www.semplicepa.it/) [Accesso: 09/11/2017].
9. Cogito, Expert System (http://goo.gl/CWJE2o) [Accesso: 09/11/2017].
10. Luca Piana, "Facility LIve, start-up italiana che sfida Google", L'Espresso, 14 Dicembre 2015 (http://goo.gl/sjxblF) [Accesso: 09/11/2017].
11. Celi, Language Technology, (http://goo.gl/2XxCqU) [Accesso: 09/11/2017].
12. OpenStreetMap (https://goo.gl/V96Vsw) [Accesso: 09/11/2017].

## Autori

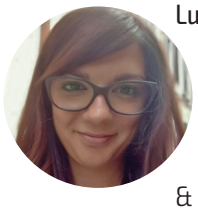
**Martina Miliani** martina.miliani@semplicepa.it

Giornalista pubblicista, per quattro anni ha vissuto a Palermo dove ha lavorato come cronista. Laureanda del corso Magistrale di Informatica Umanistica dell'Università di Pisa, collabora con ETI3 occupandosi prevalentemente di user experience nell'ambito del progetto SemplicePA.

**Anna Gabbolini** anna.gabbolini@eti3.it

Avvocato, con esperienza nel dominio della Pubblica Amministrazione e nell'ambito della gestione e rendicontazione di progetti R&S. In ETI3 si occupa di coordinamento UX, di analisi dei procedimenti automatizzabili, di elaborazione di tassonomie di base per implementazione delle ontologie, della definizione dei requisiti d'uso per applicativi di analisi

semantica e linguistica e della loro verifica.



**Lucia C. Passaro** lucia.passaro@for.unipi.it

È assegnista di ricerca presso il presso il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa, e membro del CoLing Lab. I suoi interessi di ricerca vanno dall'Affective computing, all'estrazione di informazione da corpora. Altri ambiti di interesse sono il text mining, la semantica distribuzionale, l'information retrieval, la Business & Competitive Intelligence.

**Francesco Sandrelli** francesco.sandrelli@eti3.it

Si occupa della ricerca e dello sviluppo delle tecnologie open source per varie aziende IT. Si occupa anche della definizione dell'architettura dei sistemi e della definizione delle tecnologie di riferimento. Partendo da una formazione scientifica e da un dottorato in Fisica, ha trasformato la passione per l'informatica in un'esperienza di oltre 10 anni come sviluppatore e Project manager.

**Alessandro Lenci** alessandro.lenci@unipi.it

Professore associato di linguistica presso il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa, dove dirige il Laboratorio di Linguistica Computazionale (CoLing Lab). Ha come principali aree di ricerca la semantica distribuzionale, lo sviluppo di strumenti e risorse per il trattamento automatico della lingua ed estrazione delle informazioni da testi.

**Roberto Battistelli** roberto.battistelli@eti3.it

Dalla sua esperienza come amministratore comunale è nata l'idea degli strumenti volti ad implementare la knowledge awareness nella Pubblica Amministrazione e, quindi, nell'ambito dei soggetti privati. In ETI3 si occupa di analisi dei requisiti e delle criticità, di progettazione e di sviluppo, della verifica dei risultati e dei test, del coordinamento e della collaborazione con i soggetti istituzionali.