UNIONE EUROPEA
Fondo Sociale Europeo
Investiamo nel tuo futuro

pon SPAO
SISTEMI DI POLITICHE ATTIVE PER L'OCCUPAZIONE

ANPAL
Agenzia Nazionale Politiche Attive del Lavoro

RAPPORTO TECNICO

PIATTAFORME DI ANALITICA VISUALE PER LA VALORIZZAZIONE DEL PATRIMONIO INFORMATIVO E DOCUMENTALE INAPP ANALISI PRELIMINARI E IPOTESI DI LAVORO

Nicola Lettieri

Dicembre 2021

Sommario

Il rapporto analizza motivazioni e contenuti di una ricerca incentrata sullo sviluppo di strumenti integrati per l'applicazione di tecniche di *data integration*, *data mining* e *visual analytics* al patrimonio informativo e documentale INAPP in funzione anche delle attività previste dal PTA 2021-2023. Muovendo in questa direzione, il lavoro prende in esame due temi tra loro collegati: i) metodi innovativi per l'accesso e la fruizione dei dati e documenti a vario titolo prodotti, trattati o gestiti dall'Istituto (risultati di indagini, cataloghi bibliografici, *dataset* normativi etc.); ii) possibili forme di integrazione di euristiche computazionali (*machine learning*, *natural language processing*, *network analysis* etc.), per supportare in modi nuovi la valutazione dell'impatto delle politiche pubbliche. Il documento è strutturato come segue. Il paragrafo 1 tratteggia il *framework* scientifico, metodologico e applicativo in cui si colloca la riflessione. Il paragrafo 2 descrive il *concept* di una piattaforma analitica online pensata per connettere, analizzare e visualizzare in maniera integrata dati normativi e microdati amministrativi. La presentazione del prototipo in termini di architettura, funzionalità e tecnologie diventa lo spunto per formulare delle ipotesi in merito a possibili future attività dell'istituto in questa direzione. I paragrafi successivi sono dedicati a una breve disamina delle prospettive di sviluppo di una ricerca su questi temi e a una prima rassegna della letteratura scientifica e delle fonti normative rilevanti in materia.

Outline

The report analyzes reasons and contents of a research on the creation of integrated tools for the application of data integration, data mining and visual analytics techniques to INAPP information and documentary assets according to activities foreseen in 2021-2023 PTA. The document examines two interrelated topics: i) innovative methods for accessing and using data and documents produced, processed, or managed by the Institute (surveys, bibliographic catalogues, regulatory datasets, etc.); ii) potential integration of computational heuristics (machine learning, Natural Language Processing, network analysis etc.), to support in new ways the evaluation of the impact of public policies. The document is structured as follows. Section 1 outlines the scientific, methodological, and applicative framework of the analysis. Section 2 describes the concept of an online analytical platform designed to connect, analyze, and visualize regulatory data and administrative microdata in an integrated manner. The presentation of the prototype in terms of architecture, functionality and technologies becomes the starting point for formulating hypotheses regarding possible future activities of the Institute in this direction. The following sections are devoted to a brief examination of the prospects for developing research on these issues and to a first review of the scientific literature and relevant regulatory sources on the subject matter.

Keyword

Valorizzazione patrimonio informativo pubblico, *data integration*, *visual analytics*, *data-driven science*, *interactive data exploration*, dati amministrativi, *open data*.

INDICE

1. Introduzione

L'elaborazione di strumenti innovativi per il trattamento e l'analisi dei dati costituisce una delle principali sfide poste dallo sviluppo delle ICT e dal diluvio di dati che caratterizza la società dell'informazione. La capacità di manipolare *dataset* via via più grandi e di estrarre da essi conoscenza riveste oggi un ruolo cruciale in tutti gli ambiti dell'agire umano, dalla produzione di beni e servizi al marketing per arrivare alla ricerca scientifica. Se si pone mente a quest'ultimo aspetto, un'analisi della letteratura scientifica degli ultimi dieci anni evidenzia come la transizione verso metodi di indagine *computation* e *data-driven* sia ormai realtà anche in ampi settori delle scienze sociali (Lazer et al. 2009; Cioffi Revilla 2014; Conte et al 2014), compresi ambiti disciplinari - si pensi all'economia (Einav, 2014) e al diritto (Ashley, 2017; Lettieri & Faro, 2014) - direttamente coinvolti nei processi di analisi e *design* delle politiche pubbliche di specifico interesse per il nostro istituto.

In questo scenario, ricco di opportunità scientifiche ed applicative, emergono esigenze e problematiche di ordine diverso in vario modo connesse al trattamento e alla valorizzazione dei patrimoni informativi.

a) Fruizione, esplorazione e comprensione dei dati

La quantità e l'eterogeneità di informazioni oggi accessibili supera sempre più spesso i limiti delle nostre capacità cognitive. Di fatto, abbiamo a disposizione più dati di quelli che possiamo gestire con gli strumenti tradizionali di *information retrieval* con rischi di confusione, sottoutilizzo quando non di paralisi decisionale. È sempre più necessario immaginare nuovi modi di rappresentare i dati semplificandone la manipolazione e la comprensione da parte tanto di esperti di dominio quanto di utenti generici.

b) Analisi ed estrazione di conoscenza

Nei *dataset* di grandi dimensioni sono nascoste conoscenze spesso inaccessibili alle sole tecniche di analisi statistica. La valorizzazione dei Big data a fini esplicativi, predittivi e di supporto alle decisioni è oggi legata anche all'utilizzo, spesso combinato, di un ventaglio sempre più ampio di euristiche computazionali che va dal *machine learning* alla *network analysis* passando per il *natural language processing* e le ontologie.

La ricerca di risposte alle esigenze appena accennate suggerisce una proiezione verso ambiti di ricerca diversi ma spesso complementari. Il miglioramento delle modalità di interazione con i dati può senza dubbio trarre vantaggio delle acquisizioni tecniche e metodologiche maturate nei settori dell'*Information Visualization* (Spence 2001; Mazza, 2009) e della *Visual Analytics* (Keim 2008; Zhang, 2010; Dill, 2012). Gli sforzi compiuti in queste aree di ricerca hanno condotto alla elaborazione di nuove modalità di rappresentazione dei dati capaci di facilitare notevolmente la manipolazione e comprensione dell'informazione (v., con qualche dettaglio in più, par. 2.2). Dal canto suo, invece, l'estrazione da grandi *dataset* di conoscenza utile a fini scientifici o operativi proietta verso ambiti di ricerca che, pur avendo denominazioni diverse - *data science*, *computational science, machine science, e-science* - sono accomunati dall'impiego di raffinate euristiche computazionali capaci di contribuire ad una comprensione sempre più profonda della realtà (Hey et al. 2009; Hine 2006; Boulton et al. 2012; Evans & Rzhetsky, 2010; Dhar 2013; Reed et al. 2005; Humphreys 2004).

Come emerge dal numero crescente di sperimentazioni in corso (v. Thorp 2011; McCandles 2010; Jordan 2008; Rosling 2006), l'integrazione tra visualizzazione e *data mining* rappresenta oggi una frontiera della ricerca, una frontiera alla quale è bene probabilmente che rivolga la propria attenzione anche il nostro istituto.

Inapp non solo annovera tra le sue finalità l'utilizzazione *dei dati* come *base informativa per la valutazione delle politiche pubbliche* ma, in ragione dell'appartenenza al SISTAN e, più in generale, di un quadro normativo europeo e nazionale sulla valorizzazione dei patrimoni informativi pubblici (v. par. 2.1 e rassegna normativa in par. 7.1 e 7.2), è chiamato anche a condividere il suo patrimonio informativo con una platea di *stakeholder* caratterizzati da esigenze e capacità di analisi molto diverse: ricercatori, amministratori e *policy-maker*, organizzazioni della società civile, giornalisti e, ovviamente, cittadini. Servono, pertanto, soluzioni di data analysis capaci, da un lato, di produrre nuovi *insight* utili in sede di *policy analysis* e, dall'altro, di facilitare l'accesso alle tante risorse informative interne. I dati derivanti dalle *indagini periodiche* (*PLUS*, *RIL*, *QdL*) e dalle *indagini occasionali* (*Dottori di ricerca*, *PIAAC*) offrono, in quest'ottica, materia prima utilizzabile per sperimentazioni potenzialmente molto interessanti e innovative.

In linea con quanto detto finora, l'analisi proposta in questo documento si muove in due direzioni. Il primo e più immediato è quello di sperimentare nuovi modi per valorizzare il patrimonio informativo dell'Istituto attraverso tecniche di *information visualization* capaci di agevolare la manipolazione, comprensione e fruizione dei dati in nostro possesso. Il secondo obiettivo - di più lungo periodo - è esplorare come l'integrazione in uno stesso strumento di euristiche computazionali diverse - in particolare il *machine learning* e la *network analysis* - possa essere sfruttata per estrarre dai dati statistici ed amministrativi nuove conoscenze utili alla valutazione dell'impatto delle politiche pubbliche

2. Framework applicativo e metodologico

L'analisi trae ispirazione da sviluppi tecnologici, metodologici e scientifici maturati in contesti applicativi e di ricerca diversi tra loro. Di seguito una sintetica panoramica - ancora in corso di scrittura - volta ad illustrare le premesse e le motivazioni alla base del lavoro.

2.1 Valorizzazione dei patrimoni informativi pubblici: quadro normativo di riferimento

L'interesse verso strumenti innovativi per la condivisione e l'analisi dei dati in possesso dell'istituto trova ragion d'essere, tra le altre cose, negli indirizzi ricavabili dal quadro normativo vigente. Nel contesto europeo, l'accesso alle informazioni generate dalle PA è sin dai primi anni 2000 oggetto di una serie di norme ed atti di indirizzo volti a definire scopi e forme di una valorizzazione dei patrimoni informativi pubblici intesa come parte integrante della strategia per lo sviluppo economico, sociale e scientifico della società dell'informazione.

La Direttiva 2003/98/CE[1] sulla condivisione e il riuso della *Public Sector Information* (PSI) rappresenta in quest'ottica il punto di avvio un processo regolativo[2] ancora in evoluzione ma

[1] Direttiva 2003/98/CE del Parlamento europeo e del Consiglio, del 17 novembre 2003, recante disposizioni in tema di "*Riutilizzo dell'informazione del settore pubblico"*.

[2] Ne fa parte una lunga serie di interventi tra i quali si segnalano: i) Direttiva 2013/37/CE del Parlamento europeo e del Consiglio di modifica la direttiva 2003/98/CE relativa al riutilizzo dell'informazione del settore

che ha già condotto a scelte gravide di ricadute sul piano degli ordinamenti nazionali quali l'opzione in favore del paradigma *open data* o la formulazione del principio del *libero utilizzo e riutilizzo delle informazioni pubbliche*. Come emerge da una serie di Comunicazioni della Commissione UE[3], i dati pubblici sono chiamati a svolgere ruoli diversi e di grande rilievo, in alcuni casi (v. punti *c* ed *e*) molto vicini alle finalità istituzionali dell'Inapp:

a. stimolare la crescita economica e l'innovazione
b. contribuire ad affrontare le sfide della società con lo sviluppo di soluzioni innovative
c. *favorire l'elaborazione di politiche evidence-based aumentando l'efficienza nelle pubbliche amministrazioni;*
d. offrire materia prima per lo sviluppo e l'applicazione di nuove tecnologie che, come l'intelligenza artificiale, richiedono grandi quantità di dati di alta qualità;
e. *favorire la partecipazione dei cittadini alla vita politica e sociale aumentando la trasparenza del governo e delle attività amministrative*.

Gli indirizzi generali definiti dalla Commissione in tema di *Public Sector Information* (di recente arricchiti da un riferimento specifico al tema degli *open data* con la Direttiva 2019/1024 del Parlamento europeo e del Consiglio del 20 giugno 2019 relativa all'apertura dei dati e al riutilizzo dell'informazione del settore pubblico[4]) sono stati affiancati da serie di atti regolativi di livello nazionale che hanno completato, con disposizioni di tenore e rango diverso, il quadro normativo di riferimento in materia[5].

All'interno di questo scenario, la cui analisi trascende gli scopi di questo rapporto, meritano di essere segnalate, per quanto di nostro interesse, alcune disposizioni del *Codice dell'Amministrazione Digitale* (CAD)[6], testo del 2005 più volte modificato nel corso degli anni da

pubblico; ii) la Strategia per il Mercato Unico Digitale; iii) l'atto di indirizzo *Building a European data economy*; iv) il Regolamento 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati. Per un elenco dettagliato dei provvedimenti v par. 7.1

[3] Si v., tra le altre, la Comunicazione delle Commissione UE "*Dati aperti. Un motore per l'innovazione, la crescita e una governance trasparente*" COM(2011)882 e la Comunicazione della Commissione UE "*Digital Single Market Strategy"*COM(2015)192

[4] La direttiva in parola ribadisce esplicitamente il legame esistente tra l'accesso alle informazioni e l'esercizio dei diritti fondamentali riconosciuti ai cittadini dell'Unione Europea: "*l'accesso all'informazione è un diritto fondamentale. La Carta dei diritti fondamentali dell'Unione europea (la «Carta») stabilisce che ogni persona ha diritto alla libertà di espressione che include la libertà di opinione e la libertà di ricevere o di comunicare informazioni o idee senza che vi possa essere ingerenza da parte delle autorità pubbliche e senza limiti di frontiera*". Di particolare interesse, per l'Istituto , quanto previsto dall'art. 10 che si concentra in particolar modo sui dati della ricerca precisando: "*gli Stati membri promuovono la disponibilità dei dati della ricerca adottando politiche nazionali e azioni pertinenti per rendere i dati della ricerca finanziata con fondi pubblici apertamente disponibili («politiche di accesso aperto») secondo il principio dell'apertura per impostazione predefinita e compatibili con i principi FAIR*[…] *in conformità del principio «il più aperto possibile, chiuso il tanto necessario».*"

[5] Per una lista dei principali interventi normativi/atti di indirizzo di livello nazionale in materia v. par. 7.1

[6] Istituito con d.lgs. 7 marzo 2005, n. 82, il Codice dell'Amministrazione Digitale (CAD) è un testo unico che riunisce e organizza le norme riguardanti l'informatizzazione della Pubblica Amministrazione nei rapporti con i cittadini e le imprese. Il decreto istitutivo è stato più volte modificato e integrato, da ultimo con il d.lgs.

cui emerge in maniera chiara la rilevanza assegnata dal legislatore nazionale alla valorizzazione del patrimonio informativo pubblico. Muovendo in questa direzione, l'art. 2 del CAD affida alle amministrazioni pubbliche il compito di assicurare "*la disponibilità* [...] *l'accesso, la trasmissione* [...] *e la fruibilità dell'informazione in modalità digitale"* utilizzando le tecnologie dell'informazione e della comunicazione "*nel modo più adeguato al soddisfacimento degli interessi degli utenti*".

Sulla base di queste premesse, il CAD articola una serie di norme[7] che includono la condivisione dei dati tra le finalità istituzionali "*di ogni amministrazione*". Il compito di garantire, all'interno come all'esterno della PA, un utilizzo avanzato dei dati pubblici attraverso ricorso alle più aggiornate soluzioni di condivisione e *data analysis* assume una rilevanza specifica. Non a caso, ai sensi dell'art. 52 CAD, rubricato "*Accesso telematico e riutilizzo dei dati*", le attività volte a garantire l'accesso telematico e il riutilizzo dei dati delle pubbliche amministrazioni "*rientrano tra i parametri di valutazione della performance dirigenziale*".

Accanto al CAD si collocano, per quanto di nostro interesse, diversi atti di programmazione e di indirizzo, che istituiscono un legame esplicito tra valorizzazione del patrimonio informativo pubblico, miglioramento delle condizioni di accesso e fruizione di dati, e utilizzo di tecniche di visualizzazione. Vanno menzionate, in quest'ottica, le *Linee Guida Nazionali per la Valorizzazione del Patrimonio Informativo Pubblico* predisposte dall'Agenzia Italiana per il Digitale[8] (AgID). Nel definire i criteri per la pubblicazione delle informazioni sui siti istituzionali[9], le Linee guida invitano espressamente a "*fornire, ove possibile, strumenti di visualizzazione e navigazione, anche georiferita, dei dati, che possano facilitare la lettura degli stessi*"[10]. Indirizzi dello stesso tenore sono contenuti nel *Piano triennale per l'informatica nella pubblica amministrazione 2017-2019*[11]. La sezione 4.1.2 del Piano, dedicata alle strategie da adottare per la diffusione degli

179/2016 e poi con il d.lgs. 217/2017 per promuovere e rendere effettivi i diritti di cittadinanza digitale.

[7] Le norme da citare sarebbero molte. Sul piano dei principi ci limitiamo a richiamare, l'art. 3 (*Diritto all'uso delle tecnologie*), e l'art. 7 (Diritto a servizi on-line semplici e integrati), che danno sostanza al diritto - assistito anche dalla possibilità di agire in giudizio ai sensi del d.lgs 198/2009- dei cittadini ad accedere ad informazione pubblica di qualità in termini di *fruibilità*, *accessibilità* e *tempestività*. Di uguale interesse, nella nostra prospettiva, l'art. 50 (*Disponibilità dei dati delle pubbliche amministrazioni*) che contiene, tra le altre cose, il riferimento alla necessità che i dati delle pubbliche amministrazioni siano "*formati, raccolti, conservati, resi disponibili e accessibili*" in modi *"che ne consentano la fruizione e riutilizzazione, da parte delle altre pubbliche amministrazioni e dai privat*i".

[8] L'art.71 CAD assegna all'AGID il compito di definire in termini operativi tutte le regole tecniche necessarie alla realizzazione delle finalità del Codice.

[9] Le *Linee Guida Nazionali per la Valorizzazione del Patrimonio Informativo Pubblico* sono il documento di riferimento per le pubbliche amministrazioni italiane a vario titolo tenute a rendere disponibili i propri dati in formato aperto. Il documento propone una serie di azioni volte a supportare la fruibilità e il rilascio del patrimonio informativo esposto dalla PA italiana. Le Linee sono definite e aggiornate dall'AGID che, in linea con quanto previsto dall'articolo 52 del CAD, ha il compito di coordinare e promuovere tutte le politiche nazionali di *open data*.

[10] La visualizzazione viene richiamata anche in altri punti delle Linee guida. Nella sezione dedicata agli *Aspetti organizzativi* della strategia in materia di open data, il ricorso ad "*infografiche interattive*" e alla *data visualization* viene presentato tanto come un ausilio innovativo per il trattamento dei dati *mashup* (par. 1.9.1.2.2 *Linea 2: Dati Mashup*) quanto come strumento per favorire l'*engagement* degli utenti e degli stakeholder (par. 1.9.1.2.4 *Linea 4: Coinvolgimento/Engagement*)

[11] Previsto dalla legge 28 dicembre 2015, n. 208, il *Piano Triennale per l'informatica nella Pubblica*

Open data, suggerisce la realizzazione di strumenti che consentano la generazione e la diffusione standardizzata di informazioni "*attraverso strumenti di data visualization e dashboard tematici*"[12].

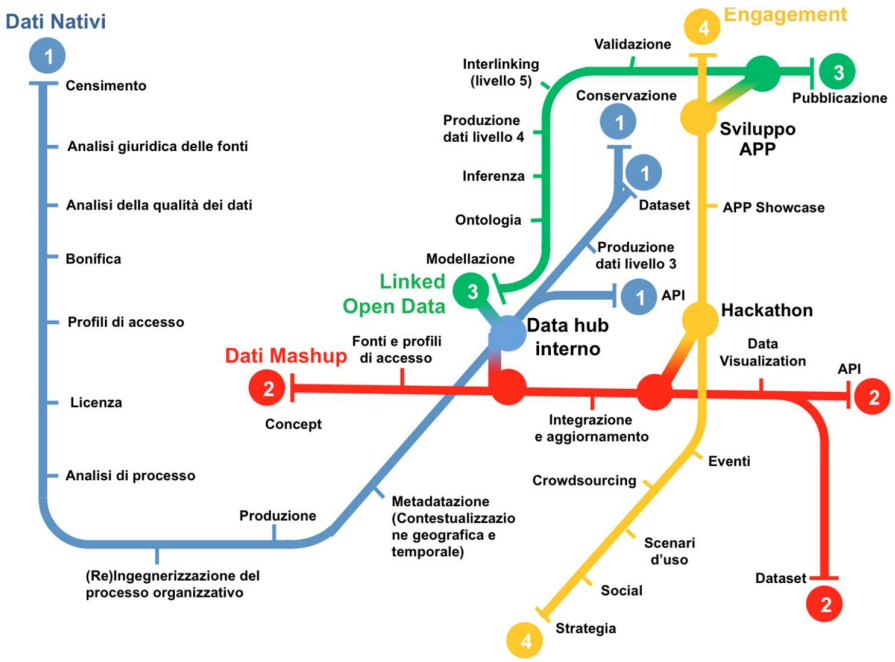

Fig. 1: Modello operativo: produzione e pubblicazione di dati aperti
(fonte: Linee Guida AgID Valorizzazione dei patrimoni informativi pubblici)

2.2 Microdati amministrativi su larga scala e *data-driven science*

Gli ultimi anni sono stati caratterizzati da un crescente sforzo delle amministrazioni pubbliche per avviare la raccolta su larga scala dei dati generati all'interno delle loro attività istituzionali. Grazie anche all'affermarsi del paradigma dell'open-government (Lathrop & Ruma, 2010; Carloni, 2014), i dati amministrativi - specie quando raccolti in grandi quantità e con elevata risoluzione temporale - sono sempre più considerati una risorsa preziosa per l'analisi e la pianificazione delle policy. Questo per qualità intrinseche ben sintetizzate di recente in (Einav & Levin, 2014): *"Administrative data is a powerful resource for a number of reasons. First, it*

amministrazione è il documento di indirizzo strategico ed economico destinato a guidare operativamente la trasformazione digitale del Paese e diventa riferimento per le amministrazioni centrali e locali nello sviluppo dei propri sistemi informativi.

[12] Mette conto osservare come una analoga attenzione alle modalità pratiche di messa a disposizione delle informazioni sia presente anche nella legislazione europea. A titolo meramente esemplificativo si può richiamare la Direttiva 2019/1024 sopra citata. L'art. 9 del provvedimento stabilisce che "*gli Stati membri adottano modalità pratiche per facilitare la ricerca dei documenti disponibili per il riutilizzo, come elenchi dei documenti più importanti, insieme ai rispettivi metadati, ove possibile e opportuno accessibili online e in formati leggibili meccanicamente, e dei portali collegati agli elenchi di contenuti*".

typically covers individuals or entities over time, creating a panel structure, and data quality is high. Moreover, because the coverage is "universal", administrative datasets can be linked to other, potentially more selective, data".

I benefici derivanti da queste caratteristiche si collocano su piani diversi. I dati amministrativi permettono di integrare con informazioni semanticamente ricche i dati generati dalle nostre interazioni online migliorando le nostre capacità di comprensione e previsione di processi socio-economici (Lazer et al., 2017). Dall'altro, la loro analisi offre alle pubbliche amministrazioni nuove opportunità per migliorare la valutazione delle politiche e la definizione dell'agenda. In tutti i casi in cui i governi hanno consentito l'accesso a grandi dataset amministrativi, sono stati conseguiti risultati interessanti tanto per la ricerca quanto per le discussioni politiche. Come evidenziato in (Einav & Levin, 2014), quando combinati con tecniche analitiche avanzate tipiche della scienza data-driven, "l*arge scale administrative data allow researchers to think "not just in terms of estimating average effects, but in terms of estimating mappings from measurable heterogeneity into treatment effects and optimal policies*".

In questo scenario, è comunque ancora difficile sfruttare a pieno il potenziale dei dati anche a causa della mancanza di strumenti che offrano analisi personalizzate e soluzioni avanzate di progettazione dell'interazione in formato open source. Le sfide da affrontare da parte dei ricercatori che desiderano trarre vantaggio da ampi set di dati sono diverse: ottenere accesso ai dati, sviluppare le capacità di gestione e programmazione necessarie per lavorare con dataset su larga scala e infine pensare ad approcci creativi per riassumere, descrivere e analizzare le informazioni.

2.3 Dalla information visualization alla visual analytics

Come sottolineato più volte sopra, la visualizzazione rappresenta uno dei principali strumenti per facilitare la manipolazione e la comprensione dei dati. Appaiono dunque evidenti le ragioni di interesse per le aree di ricerca che operano in questa direzione. Una fra queste è la Visual Analytics, espressione che identifica un campo di ricerca innovativo che . l'espressione *Visual Analytics* (VA) si identifica un campo di ricerca innovativo che mira a fornire alle persone modi nuovi per trasformare dataset di grandi dimensioni in conoscenza, consentendo loro di manipolare l'informazione in tempo reale sulla base delle loro scoperte.

Il termine "*visual analytics*" è apparso intorno al 2000 nel settore della *computer science* (Keim et al., 2008; Wong and Thomas, 2004) per poi spostarsi gradualmente in altri contesti, dando vita a un'area di ricerca multidisciplinare che combina visualizzazione, interazione uomo-computer, *data mining*, *data management*, tecniche di elaborazione dati geo-spaziali e temporali, statistica e metodi di supporto alle decisioni basate sull'elaborazione di dati georeferenziati (Keim et al., 2010). L'approccio ha ampliato notevolmente gli orizzonti della *Information Visualization* e del *Data mining*, producendo nuove tecniche rilevanti sia dal punto di vista scientifico che da quello dell'*information retrieval* (R. Baeza-Yates et alii, 1999; M. Kobayashi, K. Takeda, 2000; G. Chowdhury, 2010).

Stando alla definizione offerta da (Kohlhammer et al., 2011), incentrata sugli obiettivi di questa emergente area di ricerca, la VA è orientata alla creazione di strumenti e tecniche che consentano alle persone di:

- sintetizzare le informazioni e ricavarne di nuove a partire da dati massivi, dinamici, ambigui e spesso contrastanti;
- rilevare l'atteso e scoprire l'inaspettato;
- supportare valutazioni tempestive, ragionevoli e comprensibili;
- comunicare le valutazioni in maniera efficace rispetto alle azioni da intraprendere.

Nella prospettiva della VA, le abilità cognitive umane si integrano con la potenza computazionale alla base dell'elaborazione dei dati e con la visualizzazione diventando parte di un processo cooperativo: gli utenti guidano l'analisi mentre il sistema fornisce i mezzi di interazione per concentrarsi su task di *information retrieval*, trattamento e valutazione specifici. In un numero crescente di aree applicative, la VA facilita le interazioni all'interno dei processi decisionali che coinvolgono soggetti diversi e conducono dalla elaborazione dei dati alla decisione. Le rappresentazioni visive dei documenti, dei task e dei dati rilevanti permettono di tener traccia del percorso e di collaborare con maggiore facilità.

Nell'ultimo decennio, le tecniche VA hanno assunto rilievo in un numero crescente di contesti dalla fisica alla *business intelligence*, aree per lo più interessate all'analisi ampi *dataset* di informazioni. Lo studio delle dinamiche dei mercati finanziari (Ziegler et al., 2008) - solo per fare un esempio segnato da punti di contatto con l'analisi su larga scala di dati amministrativi di nostro interesse - richiede il trattamento e la manipolazione di enormi quantità di dati. L'esplorazione di questi dati si rivela, pertanto, una sfida cruciale per monitorare il mercato, comprendere le situazioni storiche, prevedere le tendenze o identificare schemi e gruppi di eventi ricorrenti.

2.4 Visualizzazione di dati economico/amministrativi: alcune esperienze

Nel corso dell'ultimo decennio le esperienze di studio basate sull'analisi e la visualizzazione dei dati sono cresciute in quasi tutte le aree di ricerca, con ricadute anche sul contesto economico e giuridico [4]. Alcuni progetti, strettamente collegati al trattamento di dati economici ed amministrativi, offrono un'idea della vivacità del settore, restituendo spunti interessanti per lo sviluppo di strumenti *ad hoc.* Se ne riportano alcuni esempi a scopo meramente indicativo.

a) US Census Bureau Data Visualization Gallery

Un esempio interessante di applicazione di tecniche di visualizzazione a dati amministrativi è offerto dall'United States Census Bureau, l'ufficio censimenti del Dipartimento del Commercio degli Stati Uniti. Supervisionato dall'*Economics and Statistics Administration* (ESA), il Bureau ha creato sul proprio sito una sezione "*Infographic and visualizations*" (fig. 2) che sfrutta diversi tipi di visualizzazione per rendere disponibili al pubblico dati relativi ad aspetti diversi dei cambiamenti nella crescita e nella ridistribuzione della popolazione degli Stati Uniti. Oltre i dati provenienti dal censimento decennale, il sito ospita altre visualizzazioni tematiche, dalle dinamiche familiari, alla migrazione e alla mobilità geografica, agli indicatori economici.

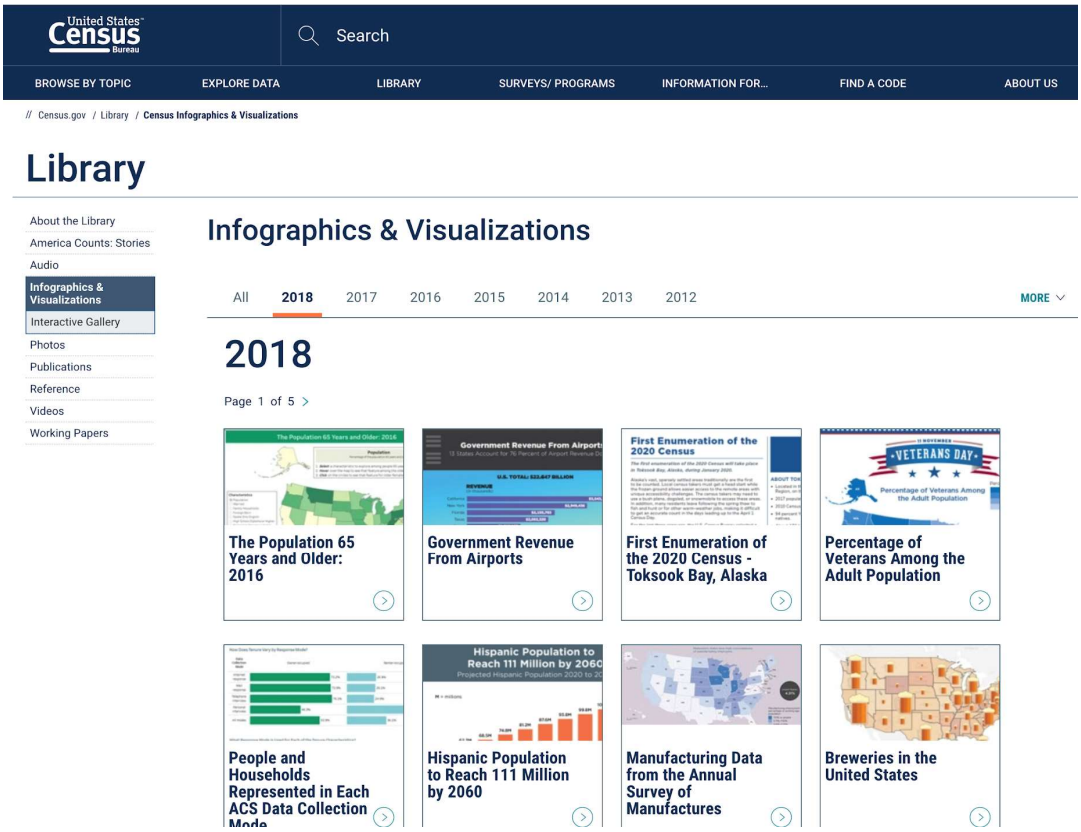

Figura 2 - *US Census Bureau*: pagina infografiche e visualizzazioni

Particolarmente interessanti si rivelano, per la tematica trattata, le visualizzazioni che esplorano il rapporto tra titolo di studio e tipologia di occupazione per lavoratori di area STEM (rappresentati da linee colorate) e non-STEM (rappresenti in grigio). Interagendo con il grafico (fig. 2) è possibile tracciare le differenze occupazionali fra le diverse categorie di laureati e individuare le aree di maggiore interesse ai fini dell'impiego.

Il modello grafico offre inoltre la possibilità di estendere l'analisi ad elementi diversi dal titolo di studio, quale il sesso, la razza o l'origine ispanica. Confrontando i livelli occupazionali di uomini e donne collegati all'area STEM, ad esempio, si scopre che gli uomini hanno maggiori probabilità di specializzarsi in ingegneria e di ottenere successivamente un impiego coerente con gli studi svolti.

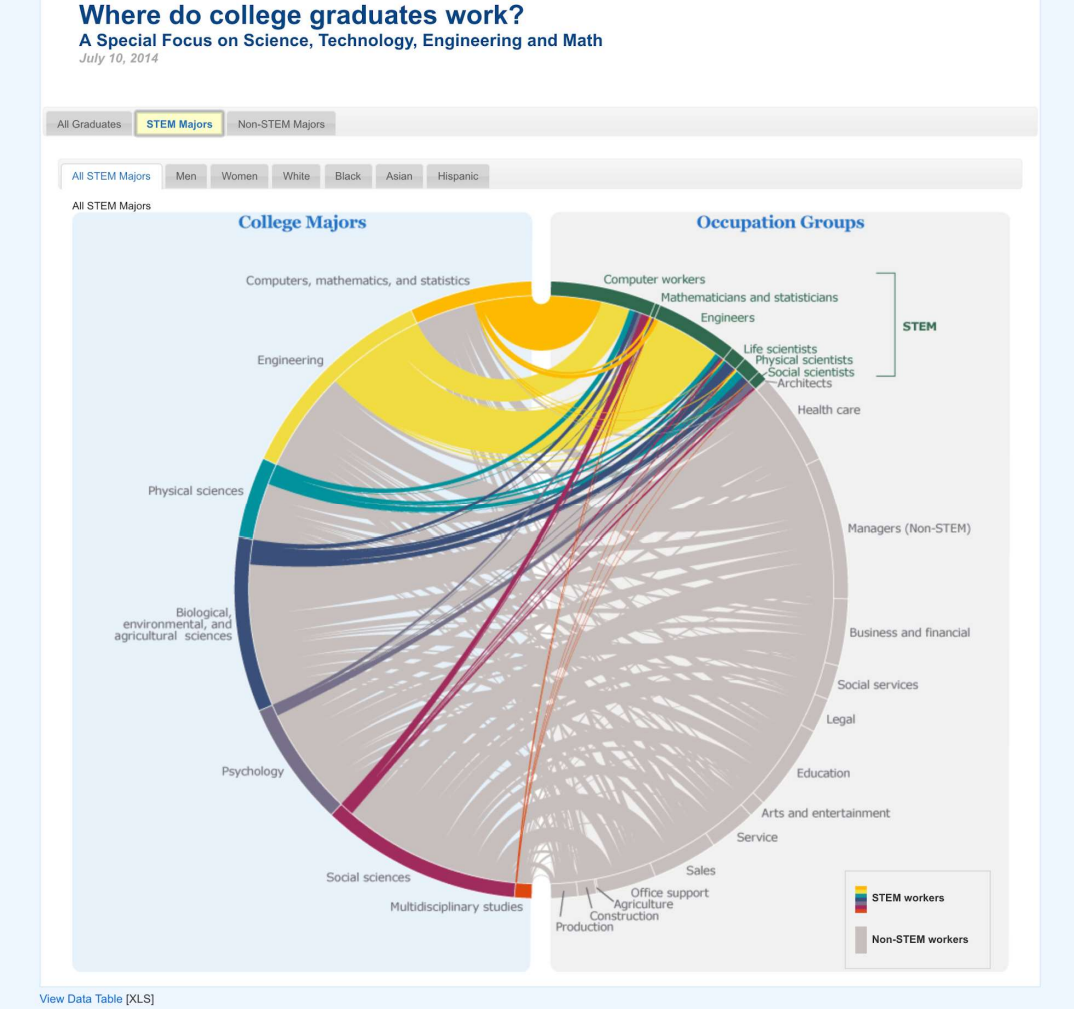

Figura 3 - *US Census Bureau*: visualizzazione relativa al rapporto occupazione/titolo di studio

b) GapMinder World TrendAnalyzer

GapMinder World Trendalyzer (Zambrano et al. 2008, Rosling 2011), è un servizio web sviluppato dalla *GapMinder Foundation*[13] e basato sulla tecnologia *Google Motion Charts*[14] per potenziare la sua grafica, che mostra serie temporali di statistiche di sviluppo per tutti i paesi del mondo e molte regioni subnazionali. La Gapminder Foundation ha anche prodotto una serie di altri

[13] *Gapminder Foundation* è un'impresa senza scopo di lucro che promuove lo sviluppo globale sostenibile e il conseguimento degli obiettivi di sviluppo del Millennio delle Nazioni Unite mediante un maggiore uso e comprensione delle statistiche e di altre informazioni sullo sviluppo sociale, economico e ambientale a livello locale, nazionale e a livelli globali.

[14] *Google Motion Charts* è un servizio web interattivo che crea grafici grafici dalle informazioni fornite dall'utente. L'utente fornisce dati e una specifica di formattazione espressa in JavaScript incorporato in una pagina Web; in risposta il servizio invia un'immagine del grafico.

progetti, tra cui: *World Income Distribution*, un display interattivo di statistiche sulla distribuzione del reddito delle famiglie per il Bangladesh, il Brasile, la Cina, l'India, l'Indonesia, il Giappone, la Nigeria, il Pakistan e gli Stati Uniti e il mondo nel suo insieme in ogni anno dal 1970 al 1998.

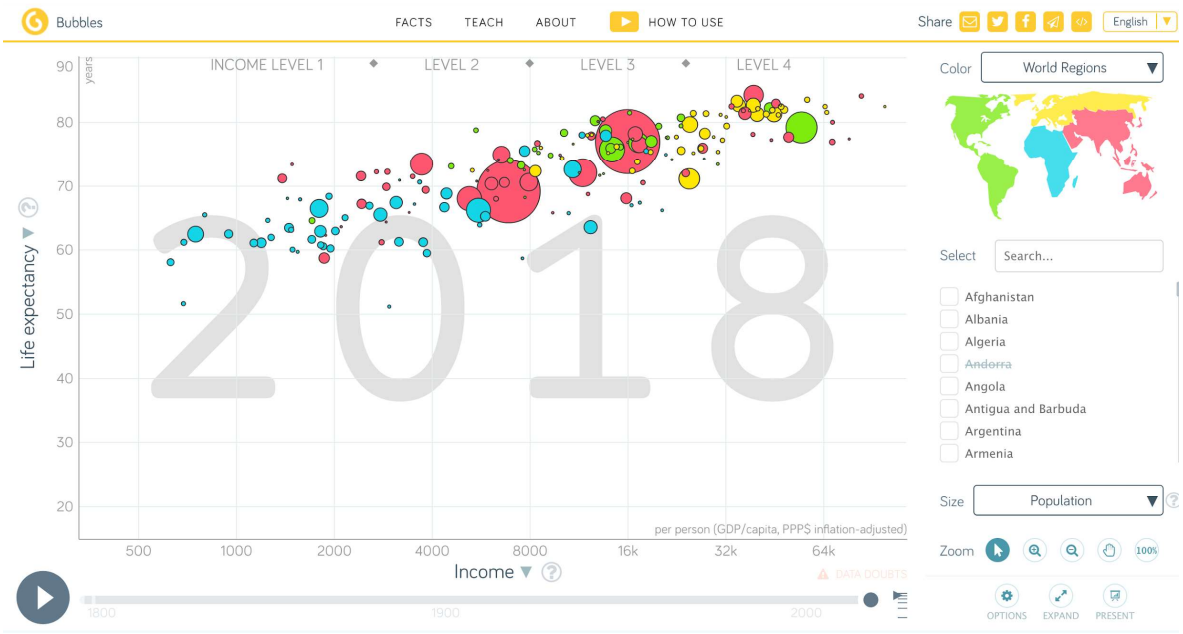

Figura 4: *Gapminder TrendAnalyzer - bubblechart* relazione reddito/aspettativa di vita

b) *MIT Observatory on Economic Complexity*

L'Osservatorio sulla complessità economica (OEC) è un ambiente *online* per la consultazione e l'analisi dei dati del commercio internazionale e degli indicatori di complessità economica resi disponibili attraverso visualizzazioni interattive. Il tool, basato sulla libreria grafica Java D3[15], è il risultato di un progetto condotto presso il laboratorio *Macro Connections* (ora *Collective Learning*) attivo presso i *Media Lab* del *Massachusetts Institute of Technology*. Tra gli aspetti più interessanti del progetto OEC vi è senza dubbio la possibilità di accedere a intuitive ed efficaci descrizioni visuali dei dati relativi alle condizioni economiche e alle politiche di scambio praticate da diversi Paesi nel mondo.

Tra gli aspetti più interessanti della sperimentazione condotta all'interno dell OEC, oltre alla quantità ed eterogeneità delle informazioni gestite[16], vi è il ricorso contestuale ad un ventaglio molto ampio di soluzioni di *visual analytics* diverse che includono treemap (fig. 5), *geomap* (fig. 6), *scatterplot* e *bubblechart* (fig. 7), grafi (fig. 8).

[15] D3.js è una libreria JavaScript per creare visualizzazioni dinamiche ed interattive partendo da dati organizzati, visibili attraverso un comune browser. Per fare ciò si serve largamente degli standard web: SVG, HTML5, e CSS. Per una panoramica delle visualizzazioni possibili con D3 v. https://d3js.org/

[16] OEC rende disponibili oltre 50 anni di dati relativi al commercio internazionale attraverso decine di milioni di visualizzazioni interattive modificabili dinamicamente dagli utenti.

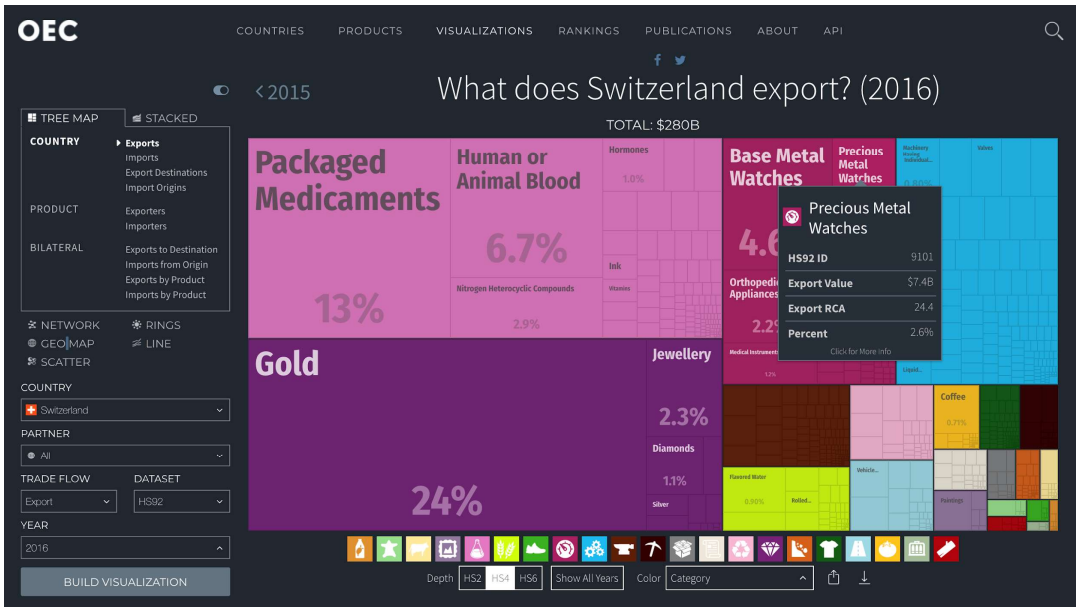

Figura 5:MIT Observatory on Economic Complexity
Tree Map - visualizzazione riassuntiva dei valori di esportazione nel mercato svizzero.

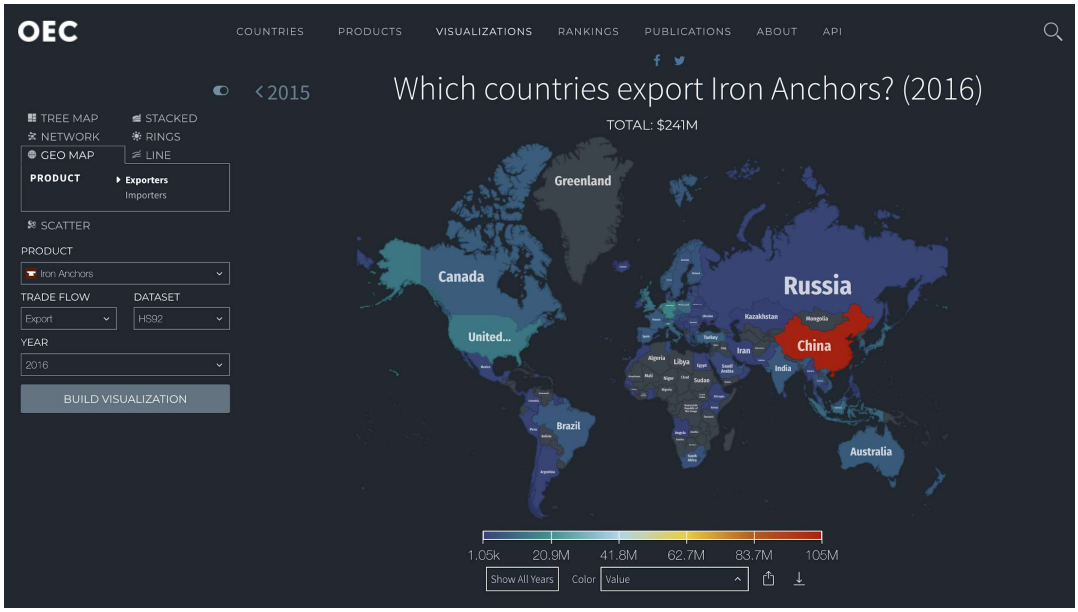

Figura 6: MIT Observatory on Economic Complexity
Geo Map - andamento del mercato delle ancore di acciaio per l'anno 2016.

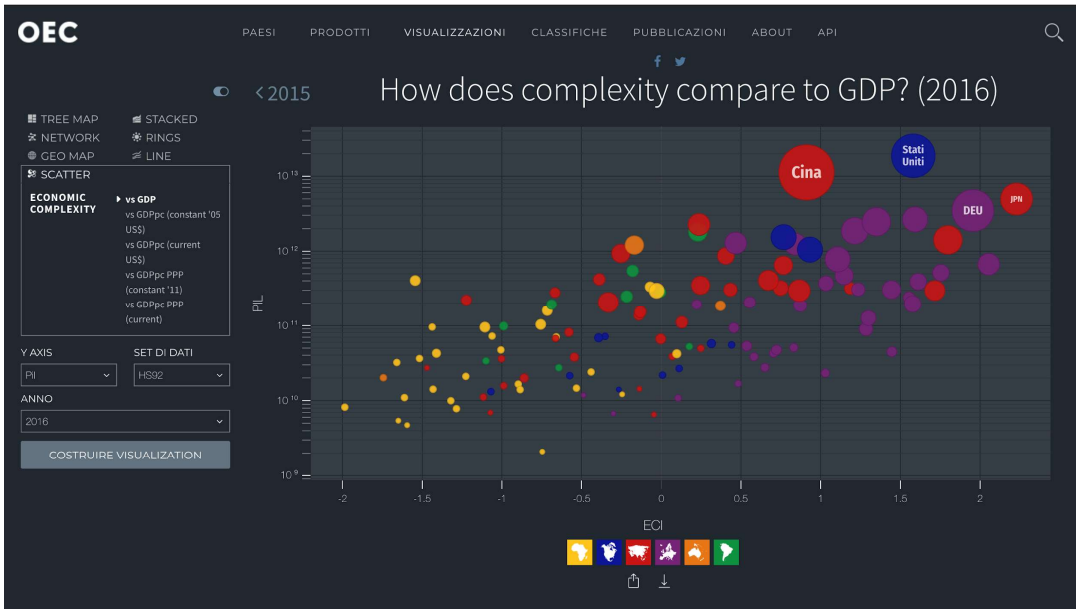

Figura 7: *MIT Observatory on Economic Complexity.* Scatter - visualizzazione descrittiva della complessità economica delle diverse aree del mondo in relazione al PIL

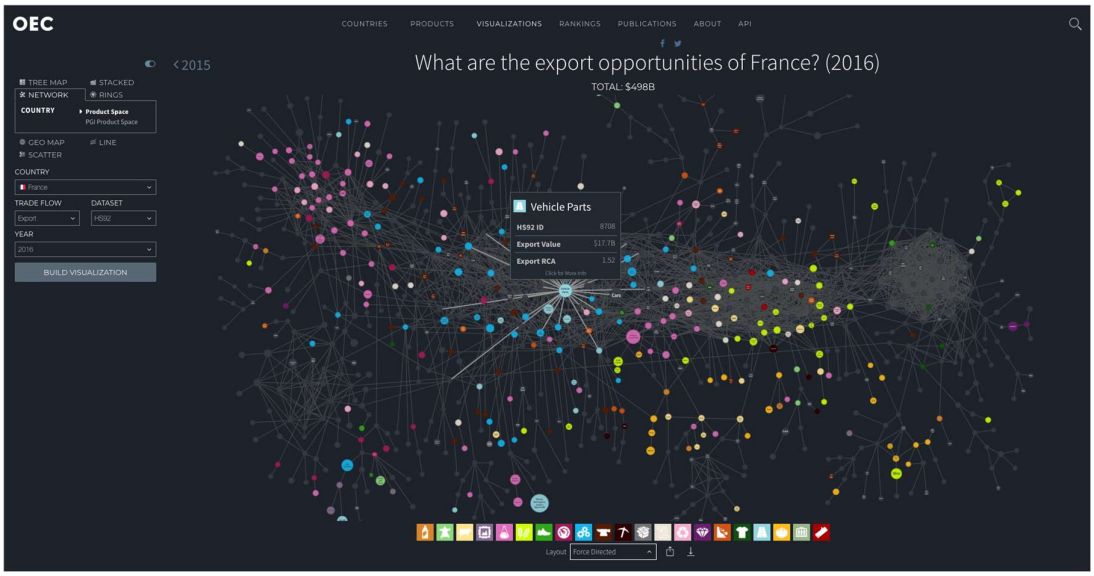

Figura 8: *MIT Observatory on Economic Complexity*
Network - visualizzazione relativa alla rete delle esportazioni francesi per l'anno 2016.

3. Visual analytics e valorizzazione del patrimonio informativo INAPP: un prototipo

Lo scenario sinora descritto rende evidenti le numerose ragioni di carattere normativo, scientifico e tecnologico che spingono verso la creazione di *tool* che supportino in modi nuovi l'accesso, la fruizione e l'estrazione di conoscenza dai dati pubblici. Muovendo in quest'ottica, si

presenta di seguito il *design* di una piattaforma analitica pensata per esplorare sperimentalmente l'applicazione di tecniche di *visualizzazione* e *data mining* a documenti e dati (risultati di indagini, cataloghi bibliografici, *dataset* normativi etc.) in vario modo nella disponibilità dell'Istituto.

L'obiettivo di fondo dell'esperimento, già tradottosi nello sviluppo di un prototipo[17] testato su dati di tipo normativo e amministrativo, è quello di mettere a fuoco le problematiche tecniche connesse alla implementazione di nuove forme di valorizzazione del patrimonio informativo Inapp considerato nel suo complesso, offrendo al contempo spunti in merito a possibili sviluppi futuri.

Come si avrà modo di evidenziare, l'architettura tecnologica immaginata è suscettibile di una agevole evoluzione verso l'implementazione di funzionalità più avanzate e, in ultima istanza, verso la soddisfazione di esigenze via via più complesse. In termini generali, i bisogni presi a riferimento nel design del prototipo sono i seguenti.

Condivisione dei dati Inapp con il pubblico

Un primo obiettivo del *tool* è quello di facilitare la fruizione dei dati a vario titolo disponibili in Istituto con infografiche interattive che rendano più agevoli ed intuitivi l'accesso, la comprensione e la rielaborazione delle informazioni.

Ricerca

Un secondo obiettivo è quello di sviluppare e sperimentare nuovi supporti - in particolare tecniche di Intelligenza Artificiale - per lo studio di fenomeni economici, sociali e giuridici di interesse per l'Istituto e, in particolare, per la comprensione dei diversi fattori che incidono sulle dinamiche del mercato del lavoro.

Supporto al data-driven policy making

Una terzo obiettivo, di più lungo periodo, potrebbe essere quello di raffinare il *tool*[18] incorporando al suo interno funzionalità più avanzate utili ai soggetti istituzionali coinvolti nell'elaborazione delle politiche pubbliche in linea, tra l'altro, con l'attenzione riservata dai più recenti atti di indirizzo di livello nazionale[19], alla realizzazione di "*strumenti analitici*" per supportare la definizione di "*policy data-driven*".

[17] Sviluppato grazie a una collaborazione con colleghi del Dipartimento di Informatica dell'Università di Salerno il *tool* è già stato oggetto di una prima pubblicazione in ambito computer science (cfr. A. Guarino, N. Lettieri *et alii*, 2019)

[18] La scelta si porrebbe in linea di continuità con il principio del riuso delle soluzioni informatiche all'interno della PA di cui all'art. 69 CAD.

[19] Si veda, in particolare, il *Piano Triennale per l'Informatica nella Pubblica amministrazione 2017–2019,* documento di indirizzo strategico ed economico elaborato da AgID e dal Team per la Trasformazione Digitale con cui si definiscono i modelli di riferimento per lo sviluppo dell'informatica pubblica italiana e le strategie operativa di trasformazione digitale del Paese.

3.1 Premessa: uno sguardo ai *tool* commerciali

Il punto di partenza della sperimentazione è stato rappresentato da una ricognizione delle tecnologie oggi utilizzabili per la generazione di visualizzazioni interattive a partire da grandi quantità di dati. L'indagine ha messo in evidenza l'esistenza di un numero ormai considerevole di soluzioni commerciali per la realizzazione di infografiche e visualizzazioni sia *stand alone* e *online*.

Tableau+ (Murray, 2013) - per citare solo uno dei più noti tra i *software* in commercio - permette di esplorare e analizzare (identificare trend nascosti, rappresentare le grandezze più importanti) in modo visuale volumi consistenti di informazioni a partire dalle più svariate tipologie di dati (*ASCII delimited*, *csv, XML*, *JSON e*d altri formati tabulari).

Nonostante queste caratteristiche, le soluzioni commerciali presentano dei limiti intrinseci spesso legati alle soluzioni *general purpose*: la loro versatilità, il loro essere adatti a molti impieghi e non specializzati per particolari esigenze presenta come effetto indesiderato la scarsa adattabilità ad esigenze specifiche. A questo si aggiunge il limite rappresentato dalla natura proprietaria e non *open source* della gran parte delle soluzioni.

La circostanza impedisce o comunque complica non solo la personalizzazione delle funzionalità, ma anche il controllo effettivo sugli algoritmi usati nell'elaborazione dei dati, cosa questa particolarmente problematica quando si ha a che fare con il trattamento di *informazioni pubbliche* per *scopi di interesse pubblico*[20]*.* Alla luce di queste considerazioni, lo sviluppo di un *tool ad hoc* mostra diversi vantaggi*:*

a. possibilità di implementare soluzioni customizzate aderenti in maniera specifica ai *dati trattati* (struttura, caratteristiche, eventuali errori) e agli *obiettivi conoscitivi* dei degli utenti siano essi cittadini, ricercatori o policy maker;
b. quantità di strumenti/euristiche/metodi integrabili nello stesso ambiente ed applicabili agli stessi dati (es. l'integrazione di funzionalità appartenenti a domini diversi come il machine learning e la *network analysis*);
c. soluzioni *open source* e riutilizzabili;
d. costi notevolmente ridotti rispetto alle soluzioni proprietarie.

3.2 Funzionalità: prime ipotesi

La piattaforma è pensata per integrare gradualmente, all'interno di una struttura modulare ancora in via di sviluppo e suscettibile di espansione, funzionalità diverse che possono essere ricondotte alla categoria della "*interactive data exploration*" (Dimitriadou et al 2014; Idreos et al 2015; Summa et al. 2016)*,* una "metafunzionalità" che sfrutta l'efficace integrazione di tecniche di *data mining*, *visual analytics* e *human-computer interaction* (HCI) per consentire agli utenti di esplorare dati processati con euristiche più o meno complesse.

L'architettura della piattaforma (fig. 9) è stata pensata avendo a mente la possibile aggiunta di moduli nuovi. Di seguito una breve descrizione delle caratteristiche dei moduli già previsti

[20] Non è un caso che il CAD contenga al proprio interno un dichiarazione esplicita del *favor* del legislatore nei confronti di soluzioni *open source* (si v., in particolare l'art. 69 rubricato "*Riuso delle soluzioni e standard aperti*"). Per una panoramica sul punto si v., con specifico riferimento all'attività amministrativa, l'analisi condotta in (Marzano, 2009).

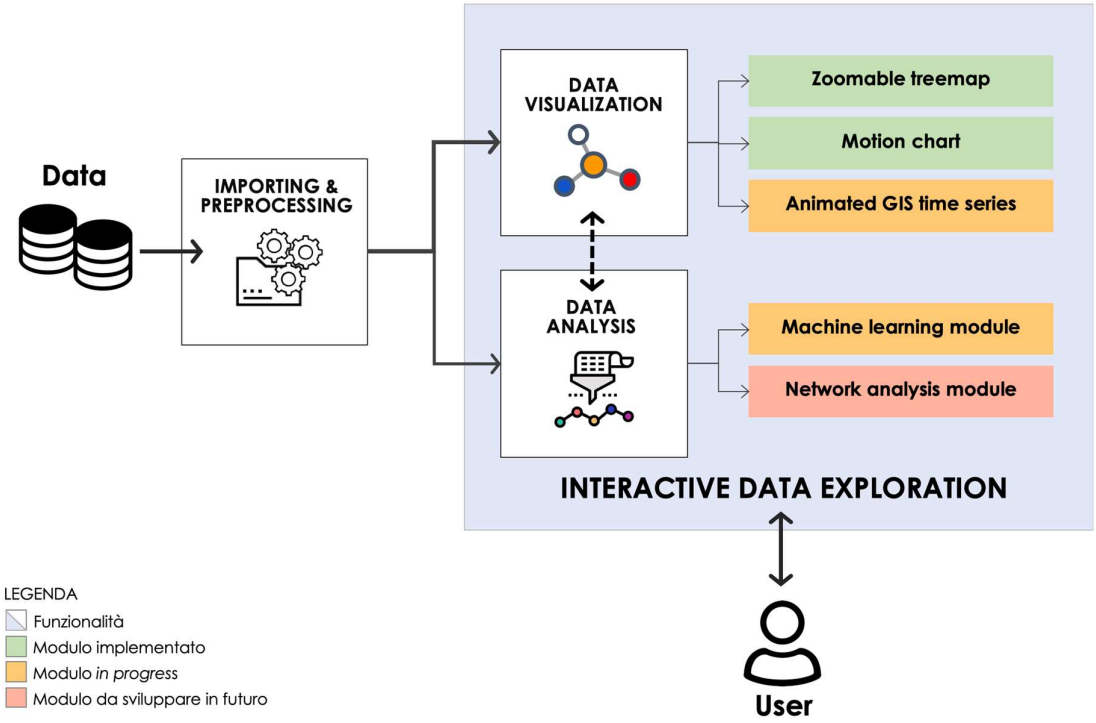

Figura 9: Architettura funzionale della piattaforma

Importing & preprocessing
La trasformazione delle informazioni in infografiche interattive presuppone che i dati siano processati per: i) correggere inconsistenze ed errori eventualmente presenti nei database di provenienza; ii) convertire le informazioni dal formato tabulare iniziale (xls; csv etc.) in formati utilizzabili dai layer di visualizzazione (es. *JSON*; *gfx*; *KML* etc.).

Data visualization
La funzionalità di visualizzazione permette di estrarre in *real time* rappresentazioni grafiche diverse a partire dallo stesso *dataset*, una volta ultimata la fase di *preprocessing*. I moduli su cui si fonda (alcuni già implementati completamente, altri in via di sviluppo) sono tre (per una prima analisi dei moduli si v. par 4):

- *Zoomable treemap*
- *Motion chart*
- *Animated GIS time series*

Data Analysis
La funzionalità di analisi dei dati, ancora in fase di sviluppo, è destinata ad estrarre dal dataset conoscenze utili agli utenti attraverso euristiche diverse. La letteratura sull'applicazione di euristiche *data-driven* allo studio dei dati amministrativi e delle dinamiche del mercato del lavoro ha suggerito l'implementazione di due diversi moduli (per una prima analisi si v. par 5):

- *Machine learning*
- *Social Network Analysis*

3.3 Architettura e tecnologie

Dal punto di vista tecnologico, la piattaforma oggetto di sviluppo si basa sulle tecnologie *web Java Spring* con una classica architettura *3-layer*[21] che segue lo schema riportato in figura 10.

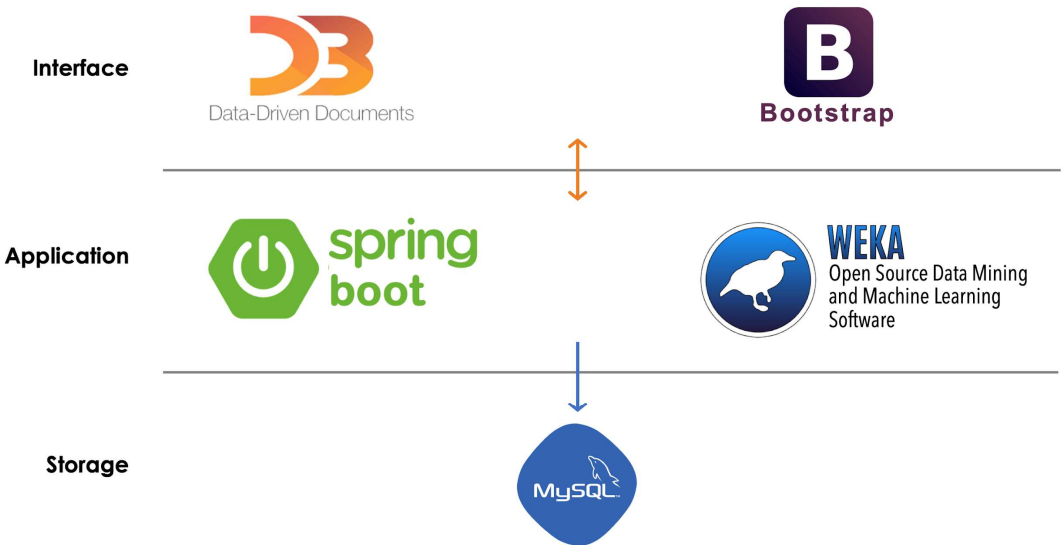

Figura 10: Architettura del sistema

[21] In ingegneria del software e nell'ambito dei sistemi informatici, l'espressione architettura three-tier ("a tre strati") indica una particolare architettura software per l'esecuzione di un'applicazione web che prevede la suddivisione dell'applicazione in tre diversi moduli o strati dedicati rispettivamente alla interfaccia utente, alla logica funzionale (*business logic*) e alla gestione dei dati persistenti. Tale architettura va tipicamente a mappare a livello fisico-infrastrutturale quella del sistema informatico ospitante l'applicazione da eseguire. I moduli sono pensati per interagire fra loro secondo le linee generali del paradigma *client-server* (l'interfaccia è cliente della *business logic*, e questa è cliente del modulo di gestione dei dati persistenti) e utilizzando interfacce ben definite. In questo modo, ciascuno dei tre moduli può essere modificato o sostituito indipendentemente dagli altri conferendo scalabilità e manutenibilità all'applicazione.

Interface
la piattaforma presenta una interfaccia intuitiva, *responsive*, implementata con uno scheletro in *HTML, Bootstrap* (per lo stile) ed attualmente utilizza la libreria *JavaScript D3.js* - che sfrutta dati in formato *JSON* - per la visualizzazione;

Application
il lato server dell'applicazione è implementato in Java utilizzando il *framework* Spring Boot. La piattaforma si avvale inoltre di WEKA, diffusa libreria per il machine learning. Questo layer si occupa di gestire tutte le richieste dell'utente e la comunicazione con il database *MySQL* al livello *Storage.* I dati ricevuti vengono elaborati e forniti al layer *interface* in formato JSON.

Storage
I dati sono ospitati in un database di tipo *MySQL* .Un driver *JDBC*[22] permette al *layer Storage* di comunicare con il layer *Application*.

4. Un esperimento con dati normativi e microdati amministrativi

Un primo esperimento - ancora in corso - ha rivolto la propria attenzione a dati amministrativi derivanti dalle comunicazioni obbligatorie di cui al d. lgs. 23 aprile 2000, n. 181 s.m.i.. Partendo dalla struttura XML dei *dataset* COB disponibile sul sito CLICKLAVORO[23] e presa a riferimento per una prima simulazione delle visualizzazioni implementabili, è stata elaborata e sottoposta a una serie di test preliminari una serie di soluzioni di *visual analytics* pensate per contribuire a ricavare *insight* relativi all'impatto delle riforme del mercato del lavoro[24].

[22] In informatica, JDBC (*Java DataBase Connectivity*), è un connettore (driver) per database che consente l'accesso e la gestione della persistenza dei dati sulle basi di dati da qualsiasi programma scritto con il linguaggio di programmazione Java, indipendentemente dal tipo di DBMS utilizzato.

[23] Dal sito *CLICKLAVORO* (https://www.cliclavoro.gov.it/Barometro-Del-Lavoro/Pagine/Microdati-per-la-ricerca.aspx) è possibile scaricare l'elenco dei metadati associati alle COB. Più in generale, Il Ministero del Lavoro e delle Politiche Sociali e l'Inps da aprile 2013 mettono a disposizione, per scopi di ricerca, due file di dati elementari per l'analisi e la valutazione dell'evoluzione del mercato del lavoro: un campione di lavoratori dipendenti ed autonomi desunti dalle banche dati Inps, che traccia le storie lavorative individuali fino al 2018 (*File LoSaI: Longitudinal Sample Inps*) e un campione di lavoratori dipendenti e parasubordinati estratto dal Sistema delle Comunicazioni Obbligatorie, integrato da eventi di lavoro autonomo desunti dagli archivi Inps (*File CICO: Campione Integrato delle Comunicazioni Obbligatorie*).

[24] Uno degli aspetti che rende interessante l'esperimento è la consapevolezza che utilizzando i dati delle COB si avrebbe a disposizione virtualmente (ovviamente se si eccettuano i rapporti in nero) un'immagine dell'intero universo di riferimento del fenomeno oggetto di indagine. Un vantaggio importante rispetto alle indagini tradizionali sui dati amministrativi costruite, di regola, su sondaggi campione che rappresentano una piccola percentuale della popolazione statistica (es. Sestito e Viviano, 2016). Va rilevato in ogni caso come spesso, anche nel campo dei Big data, si ponga il problema della rappresentatività dei dati su cui le analisi vengono svolte e della illusione di avere a disposizione tutte le informazioni rilevanti. La circostanza viene sottolineata bene in (Lazer, 2017): "*The core issue with any data is who and what get represented. With surveys, one might ask, for example, which respondents are accessible, and what they can accurately reveal. The scale of big data sets creates the illusion that they contain all relevant information on all relevant people. However, the difference between big and everything is still infinite, and the core issues of social science research around validity and generalizability still apply. Further, certain big data can be quite brittle,*

Come anticipato sopra, sono stati sinora progettati e parzialmente implementati *tre moduli* che consentono di esplorare interattivamente il *dataset* attraverso una pagina web che offre informazioni su come la struttura del mercato del lavoro sia cambiata negli ultimi anni.

4.1 Organizzazione gerarchica e analisi relazioni tra elementi del dataset

Obiettivo

Offrire una rappresentazione intuitiva e navigabile interattivamente delle proporzioni che legano gli elementi di insiemi scelti ed ordinati gerarchicamente dall'utente. A titolo esemplificativo, immaginiamo di voler avere un quadro sinottico e navigabile delle seguenti informazioni:

- Quanti dei contratti di lavoro stipulati dal 2008 al 2015 sono a tempo determinato/indeterminato.
- In che misura i livelli di istruzione (elevato, basso) sono associati a ciascuna delle due tipologie di rapporti di lavoro
- In quali proporzioni i titoli di studio appartenenti al *livello elevato* (laurea magistrale, triennale etc.) e *basso* (scuola secondaria di 1° e 2° grado etc.) sono associati ai contratti di lavoro TI e TD.

Soluzione

La scelta per questo tipo di contenuti è caduta sulla visualizzazione *Treemap,* rappresentazione quantitativa di dati gerarchici basata su insiemi rettangoli di dimensioni che variano in funzione della consistenza numerica della categoria di dati presi in considerazione.

Ogni rettangolo contiene, nascosti al proprio interno, altri rettangoli più piccoli, elementi di livello gerarchico inferiore connessi dal livello superiore visualizzabili con un clic. La rappresentazione attraverso rettangoli incapsulati facilita la navigazione permettendo di passare dalla visione generale del rettangolo principale che rappresenta tutto l'insieme preso in considerazione, a zoom successivi che rendono visibili rettangoli riferiti a insiemi di dati più piccoli e di livello inferiore. Ad ogni rettangolo si possono associare informazioni diverse (etichette, *tooltip* e finestre *pop-up*) da attivare con un clic o con il passaggio sull'area.

Abbiamo usato l'algoritmo *Treemap* (Shneiderman, 1992) per generare, distribuendole su più livelli, le aree che corrispondono alle entità selezionate dall'utente all'interno del *dataset*. In figura 11 uno schema del processo implementato.

---

vulnerable to changes in the data generation process and to attacks motivated by the fact that they are materially consequential."

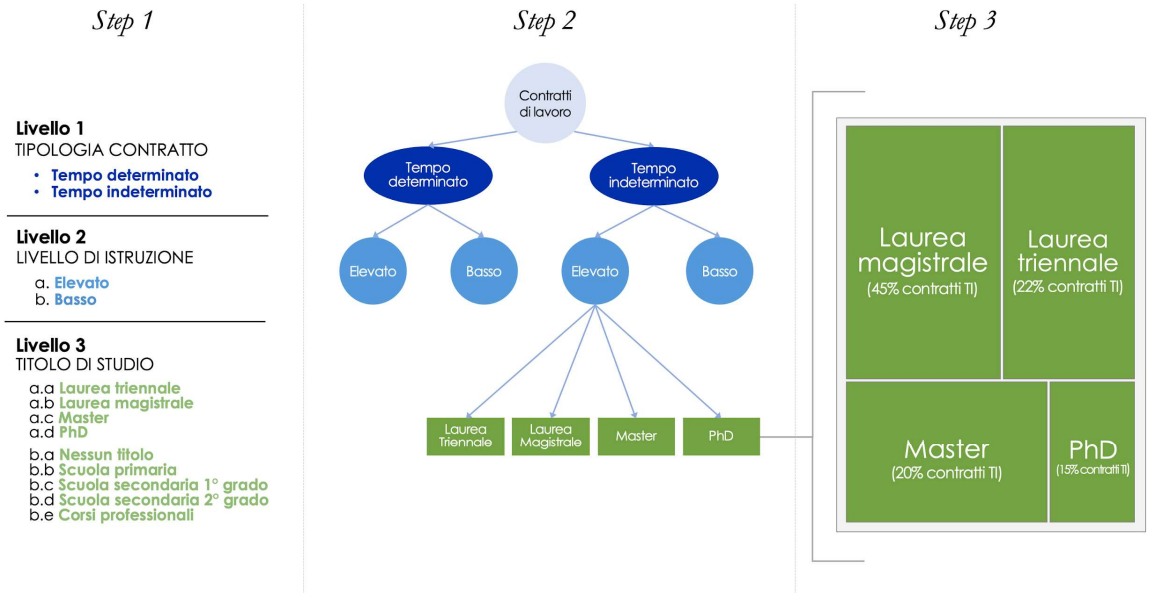

Figura 11: Schema del processo di trasformazione della struttura dati risultante dalla *query* dell'utente nell'albero gerarchico alla base della *treemap*

Step 1: l'utente seleziona e distribuisce secondo l'ordine gerarchico desiderato le entità del *dataset* che intende visualizzare.

Step 2: estrazione dei dati in base alla *query* utente e generazione dell'albero gerarchico delle entità da inserire nella treemap

Step 3: calcolo della dimensione delle aree appartenenti ai diversi livelli e generazione della *treemap*

Nell'esperimento in corso, utilizzando una *select box* nella barra laterale sinistra della piattaforma, l'utente (fig. 12a) personalizza la sua esplorazione dei dati selezionando fino a tre entità/quantità che riempiranno i livelli della *TreeMap* zoomabile.

Nella fig. 12b è possibile visualizzare (leggermente sovrapposti) i tre livelli della treemap generata nel caso in cui l'utente scelga di visualizzare le relazioni di tipo quantitativo che legano, nell'ordine, le seguenti entità: i) Contratti di lavoro (ii) Contratti TI e contratti TD; iii) Livello di istruzione[25].

[25] Ogni rettangolo nell'ultimo livello rappresenta il numero di persone con un'istruzione di alta qualità (ad esempio un dottorato) che hanno un contratto a tempo indeterminato.

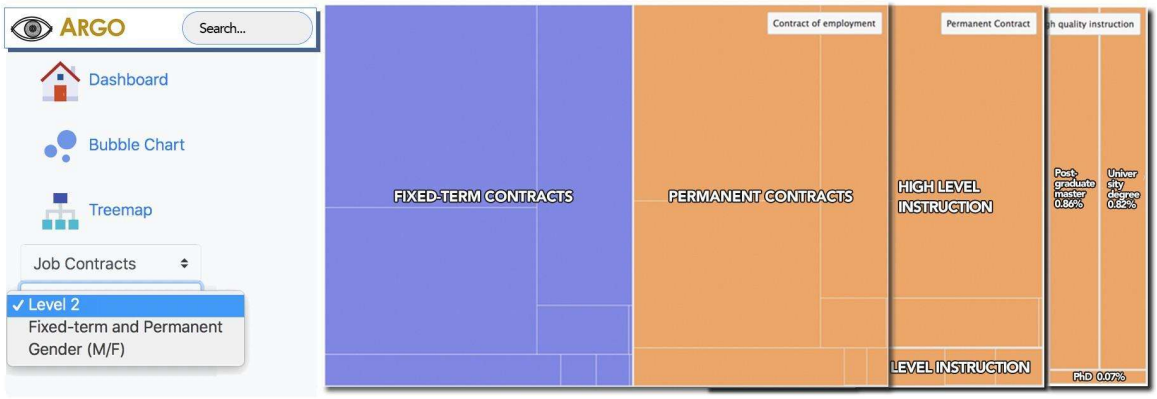

Figura 12: Treemap generata in *real time* in base alle scelte dell'utente

4.2 Rappresentazione dinamica di analisi multidimensionali longitudinali

Obiettivo

Consentire un'esplorazione intuitiva dell'evoluzione di fenomeni multidimensionali in modo da rendere evidenti le variazioni, nel tempo, di variabili tra di loro in qualche modo correlate.

Nel caso di specie la nostra attenzione ci è concentrata sull'obiettivo di rappresentare le variazioni nell'andamento dei contratti di lavoro a tempo determinato (TD) ed indeterminato (TI) nel periodo 2008-2015 offrendo contestualmente informazioni utili all'interpretazione del fenomeno (tanto in una prospettiva economica quanto in quella si una scienza giuridica empirica) e relative a[26]:

- andamento dei contratti TD/TI nelle singole regioni italiane
- andamento dei contratti TD/TI nelle aree Nord/Centro/Sud
- introduzione delle principali riforme del mercato del lavoro
- data di inizio e fine delle legislature

Soluzione

La scelta è caduta sulla *motion bubble chart,* grafico a bolle dinamico che consente una visualizzazione chiara ed efficiente di dati multivariati longitudinali. Integrata da due *timeline* usate per associare l'andamento dei dati sui contratti all'entrata in vigore delle diverse normative. La *motion chart* messa a punto a partire dalla struttura XML del dataset COB sembra essere in grado di fornire una visione dinamica del mercato del lavoro in Italia e delle sue evoluzioni nel corso del tempo. Sull'asse x abbiamo i contratti a tempo determinato e sull'asse y i contratti a tempo indeterminato.

Nell'esperimento fatto (figura 12), le *bubble* rappresentano le regioni italiane, le loro

[26] La selezione delle informazioni da associare a specifici aspetti della visualizzazione (colore e dimensione delle bubble, contenuto delle timeline visualizzate sotto la *motion chart*) è parametrica. Ferme restando le limitazioni dovute alla tipologia/semantica dei dati (non si può usare un dato non temporale per generare una *timeline*), l'utente può scegliere qualsiasi informazione presente nel dataset

dimensioni sono proporzionali gli altri tipi di contratti conclusi nella stessa regione, mentre i colori si riferiscono alla posizione geografica o alla regione (Nord, Centro, Sud). L'utente può evidenziare una (o più) regioni di interesse e seguire la sua tendenza nel corso degli anni attraverso un grafico a linee animato. Grazie all'animazione, si possono osservare le evoluzioni del mercato del lavoro messe in relazione con il succedersi delle diverse legislature e con l'intervento delle principali leggi in materia di lavoro.

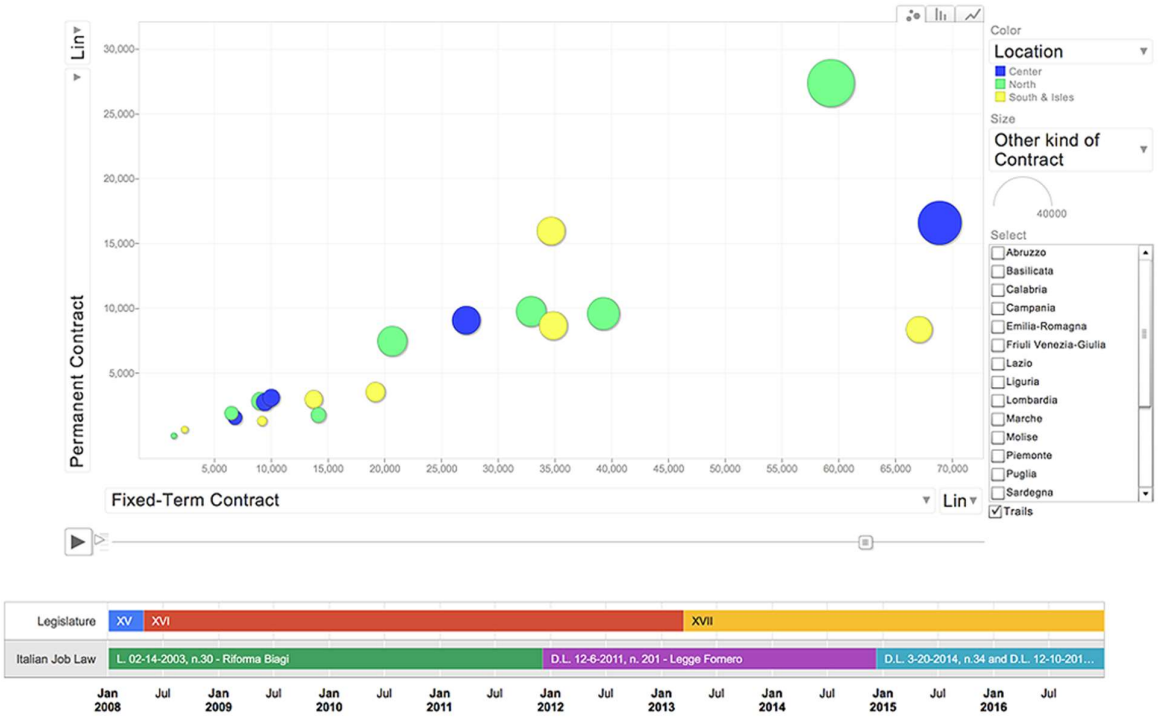

Figura 13. *Motion chart*: andamento dei contratti di lavoro TI/TD nelle regioni italiane dal 2008 al 2016.

4.3 Visualizzazione dinamica e georeferenziata dei dati

Obiettivo

Offrire una visualizzazione interattiva e animata dei dati relativi dell'andamento del mercato del lavoro in modo da permettere all'utente di mettere in relazione la dimensione spaziale e la dimensione temporale dei mutamenti in atto.

Soluzione

La scelta è caduta sulla associazione di mappe georeferenziate e *timeline* (implementate utilizzando la libreria D3.js) suscettibili di essere animate (con lo spostamento dell'indice della timeline e il mutamento del colore delle aree della mappa) così da rappresentare:

- le variazioni dei valori presi in esame in specifici contesti territoriali (con la geomap)
- la possibile correlazione degli stessi con interventi normativi indicati (con la timeline)

Un primo esperimento è consistito nella implementazione di una mappa che rappresenta, regione per regione, l'evoluzione del rapporto contratti TD/contratti TI. Le possibilità sono in ogni caso innumerevoli. Conoscendo i dati relativi alla provincia di provenienza del lavoratore e al luogo di conclusione dei contratti, si potrebbe generare, solo per fare un esempio, una mappa relativa alla dinamica, alla direzione ed agli effetti dei flussi di spostamento dei lavoratori. Nell'esperimento fatto, le mappe sono state generate con una risoluzione spaziale che si ferma al livello delle regioni. Con dati a disposizione, potrebbero essere generate mappe con risoluzione di livello provinciale. Il modulo è ancora in fase di sviluppo.

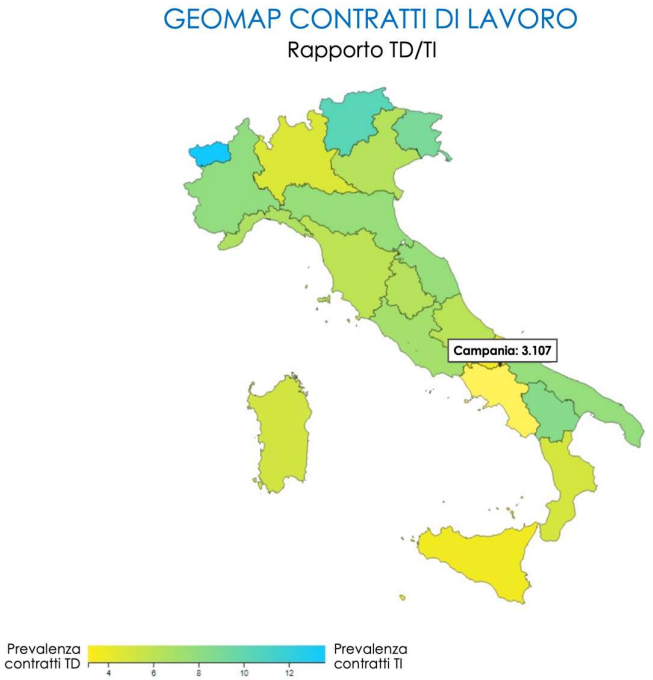

Figura 14. Geomappa rapporto contratti TD/contratti TI

4.5. Primi risultati e prospettive di sviluppo

Nonostante alcune difficoltà iniziali nell'analisi della struttura dati e alcuni rallentamenti sul piano dell'implementazione (dovuti essenzialmente ad alcune complessità riscontrate sul piano della progettazione e scrittura del codice), il lavoro ha prodotto risultati interessanti che suggeriscono di procedere nello sviluppo della piattaforma.

La possibilità di esplorare visivamente e in modo interattivo grandi quantità di dati amministrativi mettendo in relazione informazioni diverse (es. scelte di politica legislativa e potenziali effetti delle stesse sulle dinamiche del mercato del lavoro), appare promettente con riguardo tanto alle finalità di valorizzazione e condivisione del patrimonio informativo

dell'istituto, quanto all'evoluzione della valutazione dell'impatto delle politiche normative verso euristiche *data-driven*. In questa prospettiva, le piattaforme capaci di integrare grandi dataset e analisi altamente personalizzate sembrano destinate a svolgere un ruolo centrale. Gli sviluppi immaginabili (e in parte sono già stati considerati) sono molteplici ma due sembrano particolarmente interessanti e vengono riportati di seguito

4.5.1 Machine learning per feature ranking e pattern recognition

Gli ultimi anni sono stati segnati da un'attenzione sempre più accentuata - non priva, in alcuni casi, di spunti critici (Abadie & Kasy, 2017) -all'utilizzo di tecniche di *machine learning* (e, più in generale, di intelligenza artificiale) nell'analisi dei dati economici (Chernozhukov et al. 2016). Il *machine learning,* d'altra parte, rappresenta una soluzione ottimale quando si ha a che fare con enormi quantità di dati e ci permetterà di costruire un modello a partire dalla rappresentazione di questi ultimi. A differenza delle tecniche statistiche, che formalizzano delle relazioni tra variabili nei dati sotto forma di equazioni matematiche, gli algoritmi di *machine learning* "imparano" dai dati a disposizione.

Una prospettiva interessante, da questo punto di vista, sarebbe quella di dedicare un lavoro specifico (un primo esperimento è già in corso) all'implementazione di funzionalità incentrate sull'apprendimento automatico - con algoritmi di apprendimento supervisionato (Witten et al. 2016) - e/o su tecniche note in letteratura come il *feature ranking* e la f*eature selection*[27] (Novakovic, 2016). La cosa potrebbe condurre a risultati interessanti in termini di estrazione di caratteristiche (es. quali sono, tra le tante a disposizione, le caratteristiche che definiscono i lavoratori di "maggior successo"?) e il riconoscimento di *pattern* (es. identificare schemi comportamentali precisi nelle decisioni dei lavoratori o delle aziende)

4.5.2 Inferenze *network-based* per lo studio del mercato del lavoro

I dati COB contengono una quantità notevole di informazioni di tipo associativo/relazionale (dati relativi a rapporti tra individui, datori di lavoro, titoli di studio, regioni etc. arricchiti, tra l'altro, da metadati temporali) suscettibili di studi basati sulle tecniche di analisi di rete. La letteratura sedimentatasi nel campo della sociologia economica, rivela un interesse crescente verso l'impiego della network analysis (Scott, 2000; Wasserman, Faust, 1994) per lo studio del mercato del lavoro e delle sue dinamiche.

Se, storicamente, molti studi si sono concentrati sul ruolo delle reti relazionali informali (amicizia e parentela) nella dinamica delle posizioni lavorative (Granovetter, 1973; Corcoran et al. 1980; Topa, 2001; Mouw, 2003), altri lavori, anche molto recenti (Giannelle, 2014), mostrano

[27] La *feature selection/ranking* (estrazione di caratteristiche) è una particolare forma di riduzione della dimensionalità utilizzata nelle applicazioni di intelligenza artificiale quali il riconoscimento di pattern o l'elaborazione delle immagini quando i dati in ingresso sono troppi per l'esecuzione di un algoritmo e c'è il sospetto di ridondanza. Nella feature selection, i dati vengono convertiti in una rappresentazione ridotta di un insieme di caratteristiche (il vettore delle caratteristiche o *feature vector*) che permette di semplificare il costo delle risorse richieste per descrivere un grande insieme di dati accuratamente. Quando si eseguono analisi di dati complessi, uno dei più grandi problemi sta nel ridurre il numero di variabili coinvolto dato che l'analisi di un gran numero di variabili generalmente richiede molta memoria e capacità di calcolo nonchè algoritmi di classificazione che hanno bisogno di un'alta soglia di adattamento con i campioni di prova e generalizzano in modo povero nuovi campioni.

come l'analisi di dati molto simili a quelli sulle COB possa essere utilizzata per meglio comprendere i flussi di assunzione relative a categorie di lavoratori dotate di specifici titoli di studio.

Un possibile sviluppo del *tool* potrebbe consistere nella realizzazione di un modulo per l'applicazione di tecniche di Network Analysis ai dataset COB. Un primo *step* potrebbe essere l'utilizzo di dati INAPP per replicare studi di economia del lavoro interessanti già presenti in letteratura (Cappellari, 2016)

4.6 Applicazione ad altre tipologie di dati e documenti

Alla luce di quanto emerso nei paragrafi precedenti è possibile formulare delle considerazioni che prescindono dall'esperimento fatto con i microdati delle COB e guardano, in senso molto più ampio e generale, a tutto l'insieme di dati, informazioni e documenti generati, acquisiti e utilizzati dall'istituto nello svolgimento delle proprie attività.

L'esperienza sperimentale e di ricerca portata avanti anche per lo sviluppo del prototipo, ha messo in evidenza come le soluzioni di *visual analytics* dischiudano opportunità inedite e molto interessanti per l'accesso, il trattamento e l'analisi potenzialmente di tutte le categorie di dati e documenti presenti in istituto.La letteratura in materia mostra applicazioni[28] interessanti di tool di visualizzazione a tipologie di dati e documenti estremamente eterogenei incluse categorie documentali (oltre ai dati statistici, documenti di carattere giuridico, fonti bibliografiche) che si sovrappongono in larga parte con tipologie di fonti documentali e di dati già presenti in istituto.

Da questo punto di vista, sembra del tutto ragionevole iniziare a riflettere operativamente, anche in considerazione degli orientamenti assunti negli ultimi PTA, a piattaforme più complesse sul piano non solo delle funzionalità, ma anche su quello della varietà dei dati utilizzati e connessi. La messa a punto di soluzioni integrate e consentirebbe di mettere in relazione tra di loro e rendere accessibili non solo all'interno dell'istituto ma anche all'esterno, documenti e conoscenze diversi, con interessanti ricadute sul piano della valorizzazione dei dati e della comprensione dei processi del mondo reale.

Un'ultima osservazione, che estende in qualche modo i punti precedenti, riguarda il ruolo potenziale delle piattaforme analitiche nel promuovere l'adozione di approcci più interdisciplinari per la ricerca, tema che ha assunto una rilevanza crescente nelle scienze economiche e sociali. Negli ultimi anni l'interdisciplinarietà è stata oggetto di una crescente attenzione che ha indotto a vederla non solo come un'opzione scientifica, ma anche come un passaggio obbligato per gestire questioni complesse e urgenti del mondo reale che non possono essere adeguatamente affrontate da una sola disciplina (Ledford, 2015).

La capacità di integrare in modi nuovi saperi e discipline diventa in questa prospettiva cruciale e le piattaforme analitiche sembrano poter svolgere un ruolo rilevante a tal fine. Le piattaforme analitiche *online* non sono certamente l'unico mezzo per innescare questa transizione, ma possono senz'altro contribuire in maniera significativa alla creazione di nuovi modi per generare conoscenze nell'era dei Big Data.

---

[28] Molto interessanti, in questa prospettiva, i *proceedings* pubblicati ogni anno da conferenze internazionali di settore quali *IV - International Conference Information Visualisation*. Per un'analisi introduttiva allo sviluppo di piattaforme analitiche destinate in maniera specifica al trattamento della documentazione giuridica v. Lettieri, N. (2019). Knowledge machineries. introducing the instrument-enabled future of legal research and practice. In Knowledge of the Law in the Big Data Age (pp. 10-23). IOS Press.

5. *Literature review*

A completamento delle riflessioni svolte sinora, si riporta di seguito una primissima selezione di riferimenti bibliografici utili all'approfondimento dei temi trattati. La bibliografia è stata strutturata per aree tematiche al fine di agevolare la consultazione.

Tra le voci elencate non compaiono ancora quelle relative ad aspetti non ancora divenuti oggetto di approfondimento o sperimentazione (es. l'implementazione di algoritmi di *machine learning*). La bibliografia non ha alcuna pretesa di esaustività o rappresentatività rispetto agli ambiti disciplinari e tematici citati.

5.1 *Open data,* digitalizzazione PA, valorizzazione patrimonio informativo pubblico

- Agnoloni, T. (2011). Linked Open Data nel dominio giuridico. *Informatica e diritto*, (1-2), 411-430.
- Agnoloni, T., Francesconi, E., Sagri, M. T., & Tiscornia, D. (2011). Linked data in the legal domain. *Proceedings of ITAIS*.
- Agnoloni, T. , Sagri, M. & Tiscornia, D. (2009). Open data: nuova frontiera della libertà informatica?. *Informatica e diritto*, (2), 7-19.
- Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government information quarterly*, *32*(4), 399-418.
- Carloni, E. (2019). Algoritmi su carta. Politiche di digitalizzazione e trasformazione digitale delle amministrazioni. *Diritto Pubblico*, (2), 363-391.
- Carloni, E. (2014). L'amministrazione aperta: regole strumenti limiti dell'open government. Maggioli.
- Carotti, B. (2018). Il correttivo al Codice dell'amministrazione digitale: una meta-riforma. Giornale di diritto amministrativo, (2), 131-141.
- Carpentieri, R. (2013). L'Agenda digitale italiana. Giornale di diritto amministrativo, (3), 225-233.
- Cavallo Perin, R. (2020). Ragionando come se la digitalizzazione fosse data. *Diritto amministrativo*, (2), 305-328.
- Clarizia, P. (2020). La digitalizzazione della pubblica amministrazione. *Giornale di diritto amministrativo*, (6), 768-781.
- Costantino, F. (2019). Gli "open data" come strumento di legittimazione delle istituzioni pubbliche?. *Diritto e società*, (3), 443-475.
- Costantino, F. (2017). Lampi. Nuove frontiere delle decisioni amministrative tra "open" e "big data". *Diritto amministrativo*, (4), 799-836.
- De Simone, C. (2021). Dal riuso delle fonti pubbliche alla "European strategy of data". *ambientediritto.it*, (1), *ambientediritto.it*, 2021, 1, pp. 621-646. Disponibile a: https://www.ambientediritto.it/dottrina/dal-riuso-delle-fonti-pubbliche-alla-european-strategy-of-data/
- De Vivo, M. C., Polzonetti, A. & Tapanelli, P. (2011). Open Data, Business Intelligence e Governance nella Pubblica Amministrazione. *Informatica e diritto*, (1-2), 239-262.
- Falcone, M. (2017). "Big data" e pubbliche amministrazioni: nuove prospettive per la funzione conoscitiva pubblica. Rivista trimestrale di diritto pubblico, (3), 601-639.
- Giorio, D. (2021). Trasparenza e diritto d'accesso nei servizi demografici: gli "open

data". *Lo Stato Civile Italiano*, (2), 72-75.
- Gobbato, S. (2020). Verso l'attuazione della direttiva (UE) 2019/1024 sul riutilizzo degli open data della PA: nuove opportunità per le imprese. *Rivista di diritto dei media*, (2), Retrieved from: http://www.medialaws.eu
- Harrison, T. M., Guerrero, S., Burke, G. B., Cook, M., Cresswell, A., Helbig, N., ... & Pardo, T. (2012). Open government and e-government: Democratic challenges from a public value perspective. *Information Polity*, *17*(2), 83-97.
- Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European journal of ePractice*, *12*(1), 4-16.
- Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.
- Lathrop, D., & Ruma, L. (2010). *Open government: Collaboration, transparency, and participation in practice*. " O'Reilly Media, Inc.".
- Marzano, F. & Buongiovanni, A.M. (2008). Storia dell'open source nella pubblica amministrazione italiana. Informatica e diritto, (1-2), 389-406.
- Marzano, F. (2011). La trasparenza nella Pubblica Amministrazione passa dall'Open Data o l'Open Data passa dalla trasparenza?. *Informatica e diritto*, (1-2), 287-303.
- McDermott, P. (2010). Building open government. *Government Information Quarterly*, *27*(4), 401-413.
- Meijer, A. J., Curtin, D., & Hillebrandt, M. (2012). Open government: connecting vision and voice. *International review of administrative sciences*, *78*(1), 10-29.
- Obama, B. (2009). Transparency and open government. *Memorandum for the heads of executive departments and agencies*.
- Orsoni, G. & D'Orlando, E. (2019). Nuove prospettive dell'amministrazione digitale: "Open Data" e algoritmi. *Istituzioni del Federalismo,* 3, pp. 593-617. Disponibile a: http://www.regione.emilia-romagna.it/affari_ist/rivista_3_2019/Indice.html
- Pagnanelli, V. (2016). Accesso, accessibilità, "Open Data". Il modello italiano di "Open Data" pubblico nel contesto europeo. *Giornale di storia costituzionale*, (31), 205-215.
- Peruzzi, M. (2020). Il dialogo sociale europeo di fronte alle sfide della digitalizzazione. *Diritto delle relazioni industriali*, (4), 1213-1219.
- Ponti, B. (2007). Il patrimonio informativo pubblico come risorsa. I limiti del regime italiano di riutilizzo dei dati delle pubbliche amministrazioni. *Diritto pubblico*, (3), 991-1013.
- Prato F. (2021), Il fenomeno della digitalizzazione nei rapporti con la Pubblica Amministrazione: il caso italiano tra criticità e nuovi punti di partenza. *GiustAmm.it,* fasc. 3, pp. 20
- Tiscornia, D. Open Data and Re-use of Public Sector Information. *Informatica e diritto*, 1
- Tresca M.(2021), Big data, open data e algoritmi: i dati al servizio della pubblica amministrazione**.** *Rivista trimestrale di diritto pubblico,* fasc. 2, pp. 545-557
- Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives
- Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, *31*(2), 278-290.

- Votta, F. (2019). "Distributed Ledger Technology" e "Blockchain": considerazioni sulla possibile evoluzione della digitalizzazione delle amministrazioni. *GiustAmm.it*, (11), Retrieved from: http://www.giustamm.it
- Zeno-Zencovich, V. & Giannone Codiglione, G. (2016). Ten legal perspectives on the "big data revolution". *Concorrenza e mercato*, 29-57.
- Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information polity*, *19*(1, 2), 17-33.

5.2 Data science, e-science, computational science

– Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., ... & Walport, M. (2012). Science as an open enterprise. The Royal Society.
– Dhar, V. (2013). Data science and prediction. Communications of ACM, 56(12), 64-73.
– Evans, J., & Rzhetsky, A. (2010). Machine science. Science, 329(5990), 399-400.
– Hey, T., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.
– Hey, T., Tansley, S., & Tolle, K. M. (2009). Jim Gray on eScience: a transformed scientific method.
– Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science*, *308*(5723), 817-821.
– Hine, C. (Ed.). (2006). New infrastructures for knowledge production: Understanding e-science. IGI Global.
– Humphreys, P. (2004). Extending ourselves: Computational science, empiricism, and scientific method. Oxford University Press.
– Jankowski, N. W. (2007). Exploring e-science: an introduction. *Journal of Computer-Mediated Communication*, *12*(2), 549-562
– Ledford, H. (2015). How to solve the world's biggest problems. Nature News, 525(7569), 308.
– Newman, H. B., Ellisman, M. H., & Orcutt, J. A. (2003). Data-intensive e-science frontier research. *Communications of the ACM*, *46*(11), 68-77.
– Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1).
– Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), 51-59.
– Reed, D. A., Bajcsy, R., Fernandez, M. A., Griffiths, J. M., Mott, R. D., Dongarra, J., ... & Ponick, T. L. (2005). Computational science: Ensuring America's competitiveness. President's Information Technology Advisory Committee Arlington VA.
– Van Der Aalst, W. (2016). Data science in action. In *Process mining* (pp. 3-23). Springer, Berlin, Heidelberg.
– Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management.
– Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

5.3 Computational social science: temi e applicazioni

- Alvarez, R. M. (Ed.). (2016). *Computational social science*. Cambridge University Press.
- Cioffi-Revilla C. (2014), Introduction to computational social science, New York: Springer.
- Cioffi-Revilla, Claudio. (2010). Computational Social Science. Wiley Interdisciplinary Reviews (WIREs): Computational Statistics, 2(3), 259–271.
- Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., ... & Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, *214*(1), 325-346.
- Epstein, Joshua. (2006). Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton, NJ: Princeton University Press.
- Faro, S., & Lettieri, N. (2013). Law and computational social science. ESI
- Gilbert, G. N. (Ed.). (2010). *Computational social science* (Vol. 21). Sage
- Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. Annual Review of Sociology, 43, 19-39.
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915), 721.
- Lettieri, N. (2020). *Law In The Turing's Cathedral. Notes on the Algorithmic Future of the Legal Research,* In W. BARFIELD (ed.) *Cambridge Handbook on The Law Of Algorithms*, Cambridge University Press pp. 32-95
- Lettieri, N. (2017).*Scienze sociali computazionali e policy innovation*. *Nuove frontiere dell'elaborazione delle politiche pubbliche*. *Sinappsi*,1, 2017, pp. 97-119.
- Lettieri, N. (2016). Computational social science, the evolution of policy design and rule making in smart societies. *Future internet*, *8*(2), 19.
- Lettieri, N., & Parisi, D. (2013). Neminem laedere. An evolutionary agent-based model of the interplay between punishment and damaging behaviours. *Artificial intelligence and law*, *21*(4), 425-453.
- Lettieri, N., & Faro, S. (2012). Computational social science and its potential impact upon law. European Journal of Law and Technology, 3(3).
- Sagarra, O., Gutiérrez-Roig, M., Bonhoure, I., & Perelló, J. (2016). Citizen science practices for computational social science research: The conceptualization of pop-up experiments. *Frontiers in physics*, *3*, 93.
- Squazzoni, Flaminio. (2008). A (computational) social science perspective on societal transitions. Computational and Mathematical Organization Theory, 14(4), 266–282.
- Trobia, Alberto. (2001). La Sociologia Come Scienza Rigorosa: Modelli simulativi, intelligenza collettiva, forme del mutamento [Sociology as a Rigorous Science: Simulation Models, Collective Inteligence, and Patterns of Change]. Milan, Italy: Franco Angeli.

5.4 Information retrieval, interactive data exploration, visualizzazione, visual analytics,

- Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., ... & Tominski, C. (2010). Space, time and visual analytics. International journal of geographical information science, 24(10), 1577-1600.

- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.
- Chowdhury, G. G. (2010). Introduction to modern information retrieval. Facet publishing.
- Dill, J., Earnshaw, R., Kasik, D., Vince, J., & Wong, P. C. (Eds.). (2012). Expanding the frontiers of visual analytics and visualization. Springer Science & Business Media.
- Dimitriadou, K., Papaemmanouil, O., & Diao, Y. (2014). Explore-by-example: An automatic query steering framework for interactive data exploration. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 517-528). ACM.
- Dix, A. (2009). Human-computer interaction. In Encyclopedia of database systems (pp. 1327-1331). Springer, Boston, MA.
- Gupta, A., & Jain, R. (1997). Visual information retrieval. Communications of the ACM, 40(5), 70-79.
- Hao, M. C., Dayal, U., Keim, D. A., Morent, D., & Schneidewind, J. (2007, October). Intelligent visual analytics queries. In Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on (pp. 91-98). IEEE.
- Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. ACM Computing Surveys (CSUR), 32(2), 144-173.
- Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015, May). Overview of data exploration techniques. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 277-281). ACM.
- Murray, D.G. (2013) Tableau your data!: fast and easy visual analysis with Tableau software.
- Mazza, R. (2009). Introduction to information visualization. Springer Science & Business Media.
- McCandless, D. (2010). The beauty of data visualization. Talk at TED Global Oxford (http://goo. gl/7MzQ). Based on work by Tor Nørretranders.
- Fekete, J. D., Van Wijk, J. J., Stasko, J. T., & North, C. (2008). The value of information visualization. In Information visualization (pp. 1-18). Springer, Berlin, Heidelberg.
- Keim, D. A. (2002). Information visualization and visual data mining. IEEE Transactions on Visualization & Computer Graphics, (1), 1-8.
- Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H (2008) Visual analytics: Scope and challenges, Springer.
- Keim ED, Kohlhammer J, Ellis G (2010) Mastering the information age: Solving problems with visual analytics eurographics association
- Nazemi, K.: Adaptive Semantics Visualization. Studies in Computational Intelligence, p. 422. Springer International Publishing (2016). ISBN: 978-3-319-30815-9. https://doi.org/10.1007/978-3-319-30816-6
- Keim, D. A., Mansmann, F., & Thomas, J. (2010). Visual analytics: how much visualization and how much analytics?. ACM SIGKDD Explorations Newsletter, 11(2), 5-8.
- Kohlhammer J, Keim D, Pohl M, Santucci G, Andrienko G (2011) Solving Problems with Visual Analytics. Procedia Comput Sci 7:117–120. Proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (FET 11)
- Nazemi, K. (2018, January). Intelligent Visual Analytics–a Human-Adaptive Approach for

Complex and Analytical Tasks. InInternational Conference on Intelligent Human Systems Integration (pp. 180-190). Springer, Cham.
– Rosling, H., & Zhang, Z. (2011). Health advocacy with Gapminder animated statistics. Journal of epidemiology and global health, 1(1), 11-14.
– Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., & Keim, D. A. (2014). Knowledge generation model for visual analytics. IEEE transactions on visualization and computer graphics, 20(12), 1604-1613.
– Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on graphics (TOG), 11(1), 92-99.
– Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In The craft of information visualization (pp. 364-371). Morgan Kaufmann.
– Spence, R. (2001). Information visualization (Vol. 1). New York: Addison-Wesley.
– Summa, B., Gyulassy, A., Bremer, P. T., & Pascucci, V. (2016). Interactive Data Exploration. In Data-Intensive Science (pp. 370-401). Chapman and Hall/CRC.
– Tufte, E. R., & Graves-Morris, P. R. (1983). The visual display of quantitative information (Vol. 2, No. 9). Cheshire, CT: Graphics press.
– Tufte, E. R., Goeler, N. H., & Benson, R. (1990). Envisioning information (Vol. 2). Cheshire, CT: Graphics press.
– Zambrano, R, Engelhardt, Y (2008) "Diagrams for the masses: Raising public awareness–from neurath to gapminder and google earth," in Int. Conference on Theory and Application of Diagrams
– Zhang, Q. (Ed.). (2010). Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications: Data, Text and Web Mining Applications. IGI Global.

5.5 Data science, VA, IA: proiezioni negli studi sociali, giuridici ed economici

- Abadie, A., & Kasy, M. (2017). The risk of machine learning. arXiv preprint arXiv:1703.10935.
- Ashley, K. D. (2017). Artificial intelligence and legal analytics: new tools for law practice in the digital age. Cambridge University Press.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. K. (2016). Double machine learning for treatment and causal parameters (No. CWP49/16). cemmap working paper, Centre for Microdata Methods and Practice.
- Kleven, H. J. (2018) Languge trends in Public Economics, https://www.henrikkleven.com/papers.html
- Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1-24.
- Mullainathan, S. and Spiess, J (2017) Machine Learning: An Applied Econometric Approach, Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017-Pages 87–106
- Guarino, A., Lettieri, N., Malandrino, D., Russo, P., & Zaccagnino, R. (2019, July). Visual analytics to make sense of large-scale administrative and normative data. In 2019 23rd International Conference Information Visualisation (IV) (pp. 133-138). IEEE.
- Lettieri, N., Guarino, A., Malandrino, D., & Zaccagnino, R. (2021, July). The sight of Justice.

Visual knowledge mining, legal data and computational crime analysis. In *2021 25th International Conference Information Visualisation (IV)* (pp. 267-272). IEEE.
- Lettieri, N., Guarino, A., Malandrino, D., & Zaccagnino, R. (2020, September). The Affordance of Law. Sliding Treemaps browsing Hierarchically Structured Data on Touch Devices. In 2020 24th International Conference Information Visualisation (IV) (pp. 16-21). IEEE.
- Lettieri, N. (2019). Knowledge machineries. introducing the instrument-enabled future of legal research and practice. In Knowledge of the Law in the Big Data Age (pp. 10-23). IOS Press.
- Lettieri, N., & Malandrino, D. (2018, July). Cartographies of the legal world. rise and challenges of visual legal analytics. In 2018 22nd International Conference Information Visualisation (IV) (pp. 241-246). IEEE.
- Lettieri, N., Guarino, A., & Malandrino, D. (2018). E-science and the law. three experimental platforms for legal analytics. In Legal Knowledge and Information Systems (pp. 71-80). IOS Press.
- Lettieri, N., Altamura, A., & Malandrino, D. (2017). The legal macroscope: Experimenting with visual legal analytics. Information Visualization, 16(4), 332-345.
- A. J. Simoes and C. Hidalgo, "The economic complexity observatory: An analytical tool for understanding the dynamics of economic development," in Scalable Integration of Analytics and Visualization, 2011.
- Simoes, A.J.G., & Hidalgo, C. A. (2011, August). The Economic Complexity Observatory: An Analytical Tool for Understanding the Dynamics of Economic Development. In Scalable Integration of Analytics and Visualization.

5.6 Analisi di microdati amministrativi

- Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. *American economic association, ten years and beyond: Economists answer NSF's call for long-term research agendas*.
- Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social science research*, *59*, 1-12.
- Dejuan-Bitria, D., & Mora-Sanguinetti, J. S. (2021). Which legal procedure affects business investment most, and which companies are most sensitive? Evidence from microdata. *Economic Modelling*, *94*, 201-220.
- Dolls, M., Fuest, C., Neumann, D., & Peichl, A. (2018). An unemployment insurance scheme for the euro area? A comparison of different alternatives using microdata. *International Tax and Public Finance*, *25*(1), 273-309.
- Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1-24.
- Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.
- Korbmacher, J. M., & Schroeder, M. (2013). Consent when linking survey data with administrative records: The role of the interviewer. In *Survey Research Methods* (Vol. 7, No. 2, pp. 115-131). Southampton: European Survey Research Association.

- Künn, S. (2015), The challenges of linking survey and administrative data. IZA World of Labor 2015: 214 doi:10.15185/izawol.214
- Sestito, P., & Viviano, E. (2016). Hiring incentives and/or firing cost reduction? Evaluating the impact of the 2015 policies on the Italian labour market. *Evaluating the Impact of the 2015 Policies on the Italian Labour Market (March 18, 2016). Bank of Italy Occasional Paper*, (325).
- Trivellato, U. (2012). Verso politiche basate sull'evidenza: il ruolo di base di microdati. *SSI&*, 34.

5.7 Social Network Analysis: applicazioni allo studio del mercato del lavoro

– Scott, J. (2000). Social Network Analysis: A Handbook 2nd Ed. Newberry Park, CA: Sage
– Wasserman S., Faust K. (1994). Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge.
– Faust K. (2005). Using Correspondence Analysis for Joint Displays of Affiliation Networks. In: Models and Methods in Social Network Analysis, Carrington P., Scott J. and Wasserman S. (Eds.), Cambridge University Press, Cambridge, 117-147.
– Capellari, S., Chiesa, L., De Stefano, D., & Puggioni, A. (2016). L'analisi di rete per capire il mercato del lavoro. I flussi di assunzione di laureati e dottori di ricerca nel Friuli Venezia Giulia nel periodo 2005-2014.
– Mouw T. (2003). Social Capital and Finding a Job: Do Contacts Matter? American Sociological Review, 68, 868-898.
– Granovetter M.S. (1973). The Strength of Weak Ties, American Journal of Sociology, 78, 1360-1380.
– Topa G. (2001). Social Interactions, Local Spillovers and Unemployment. Review of Economic Studies, 68, 261-295.
– Corcoran M., Datcher L., Duncan G. (1980). Information and Influence Networks in Labor Markets. In: Five Thousand American Families: Patterns of Economic Progress, Duncan G. and Morgan J. (Eds), Vol.7, Ann Arbor, MI: Institute for Social Research, 1-37.
– Gianelle G. (2014) Discovering the Regional Small World of Labour Mobility. Evidence from Linked Employer–Employee Data, Regional Studies, 48(7): 1263-1278.
– Stoloff, J. A., Glanville, J. L., & Bienenstock, E. J. (1999). Women's participation in the labor force: the role of social networks. Social networks, 21(1), 91-108.

6. Normativa e atti di indirizzo in tema di valorizzazione patrimonio informativo PA

6.1 Normativa e atti di indirizzo di livello europeo

- Direttiva 2003/98/CE relativa al riutilizzo dell'informazione del settore pubblico
- Comunicazione delle Commissione UE "*Dati aperti. Un motore per l'innovazione, la crescita e una governance trasparente*" COM(2011)882
- Direttiva 2013/37/UE del Parlamento europeo e del Consiglio, del 26 giugno 2013, che modifica la direttiva 2003/98/CE relativa al riutilizzo dell'informazione del settore pubblico

- Direttiva 2019/1024 del Parlamento europeo e del Consiglio del 20 giugno 2019 relativa all'apertura dei dati e al riutilizzo dell'informazione del settore pubblico
- Comunicazione delle Commissione UE "*Orientamenti sulle licenze standard raccomandate, i dataset e la tariffazione del riutilizzo dei documenti* 2014/C 240/01
- Comunicazione delle Commissione UE "*Verso una florida economia basata sui dati*" COM(2014)442
- Comunicazione della Commissione UE "*Digital Single Market Strategy*" COM(2015)192
- Regolamento (Ue) 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016 "*Protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati*" (GDPR)
- Comunicazione delle Commissione UE "*Building a European data economy*" COM (2017)09
- Documento di lavoro dello Staff della Commissione UE "*Free flow of data and emerging issues of the European data economy*" - Allegato alla Comunicazione "Building a European data economy SWD(2017)02
- Comunicazione delle Commissione UE "*Verso uno spazio comune europeo dei dati*" COM/2018/232 final

6.2 Normativa nazionale

- D.lgs. 7 marzo 2005, n. 82 s.m.i recante Codice dell'amministrazione digitale – CAD.
- Legge, 7 agosto 2012 n.134 - Conversione in legge, con modificazioni, del decreto-legge 22 giugno 2012, n. 83, recante "*Misure urgenti per la crescita del Paese*" (si v., in particolare, le disposizioni relative all'Agenzia per l'Italia digitale).
- LEGGE 17 dicembre 2012, n. 221 - Conversione in legge, con modificazioni, del decreto-legge 18 ottobre 2012, n. 179, recante "*Ulteriori misure urgenti per la crescita del Paese*".
- Decreto Legislativo 18 maggio 2015, n. 102 - Attuazione della direttiva 2013/37/UE che modifica la direttiva 2003/98/CE, relativa al riutilizzo dell'informazione del settore pubblico
- Legge, 11 settembre 2020, n. 120 - Conversione in legge, con modificazioni, del decreto-legge 16 luglio 2020, n. 76, recante «*Misure urgenti per la semplificazione e l'innovazione digitali*» *(Decreto Semplificazioni)*
- Legge 11 agosto 2014, n. 114 (legge di conversione del decreto-legge 24 giugno 2014, n. 90 «Misure urgenti per la semplificazione e la trasparenza amministrativa e per l'efficienza degli uffici giudiziari. (si v. in particolare il Capo III recante "*Strategia di gestione del patrimonio informativo pubblico per fini istituzionali*".
- Decreto Legislativo 24 gennaio 2006, n. 36 - Attuazione della direttiva 2003/98/CE relativa al *Riutilizzo di documenti nel settore pubblico*.
- Agenzia per l'Italia digitale - Linee guida nazionali per la valorizzazione del patrimonio informativo pubblico 2017/2019, https://docs.italia.it/italia/daf/lg-patrimonio-pubblico/it/stabile/index.html

7. Riferimenti web

Observatory on Economic Complexity
https://atlas.media.mit.edu/en/

GapMinder
https://www.gapminder.org/tools/

United States Census Bureau
https://www.census.gov/library/visualizations.html

European Data Portal
https://www.europeandataportal.eu/

Agenzia Italiana per il Digitale - AGID
https://www.agid.gov.it/

Portale Dati.gov
https://www.dati.gov.it/

LexMex
http://www.lexmex.fr/

TED Talk su data visualization
Una selezione di TED Talk dedicati ai metodi ed ai contesti di utilizzo delle tecniche di visualizzazione:

Jer Thorp (2011), *Make Data More Human*
https://www.ted.com/talks/jer_thorp_make_data_more_human

D. McCandles (2010), *The beauty of data visualization*.
https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization

Chris Jordan (2008), *Turning Powerful Stats into Art*
https://www.ted.com/talks/chris_jordan_pictures_some_shocking_stats

Hans Rosling (2006), *The Best Stats You've Ever Seen*
https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Google Charts Tools
https://developers.google.com/chart/

D3.js - Data driven visualization library

https://d3js.org/

Tableau+ Visualizazion software
https://www.tableau.com/

WEKA - Data Mining with Open Source Machine Learning Software
https://www.cs.waikato.ac.nz/ml/weka/