istat
working
papers

N. 4
2012

La funzione su web per l'individuazione del codice ATECO sulla base di una descrizione sintetica e monitoraggio delle performance

Angelina Ferrillo, Stefania Macchia, Loredana Mazza,
Alberto Valery e Paola Vicari

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 4
2012

La funzione su web per l'individuazione del codice ATECO sulla base di una descrizione sintetica e monitoraggio delle performance

Angelina Ferrillo, Stefania Macchia, Loredana Mazza,
Alberto Valery e Paola Vicari

Istat

Istat Working Papers

La funzione su web per l'individuazione del codice ATECO
sulla base di una descrizione sintetica e monitoraggio delle performance

N. 4/2012

ISBN 88-458-1708-3

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 – Roma

La funzione su web per l'individuazione del codice ATECO sulla base di una descrizione sintetica e monitoraggio delle performance[1]

Angelina Ferrillo, Stefania Macchia, Loredana Mazza, Alberto Valery e Paola Vicari

Sommario

L'utilizzo di sistemi di codifica automatica per attribuire codici secondo classificazioni ufficiali a risposte testuali fornite nei questionari di indagine è ampliamente diffuso in Istituto. Questo lavoro descrive una particolare applicazione del sistema di codifica messo a punto per la classificazione delle attività economiche e finora utilizzato per codificare i testi rilevati nelle indagini: l'algoritmo per il matching testuale e la base informativa implementata per l'ATECO sono stati adattati e arricchiti di un'apposita interfaccia per l'ambiente web, in modo da fornire una funzione che consenta agli utenti web di individuare il codice ATECO corrispondente all'attività da loro espletata e fornita con una descrizione a testo libero. Viene inoltre presentato il sistema di monitoraggio della qualità messo a punto per analizzare i risultati prodotti da tale funzione ed per aggiornare costantemente la base informativa. Si prospetta infine come l'utilizzo di tale sistema, essendo generalizzato, possa essere esteso a diverse classificazioni e costituire uno strumento standard di interrogazione delle classificazioni.

Parole chiave: codifica automatica, matching testuale, ATECO.

Abstract

The use of automatic coding systems to assign classifications codes to textual responses given in survey questionnaires is widely adopted in Istat. This work describes a particular use of the coding application developed for the Economic Activity classification to code textual responses: the generalised software for textual matching and the informative base already implemented for ATECO have been adapted and enriched with an interface to be used on the web to provide a function which allow users to identify the ATECO code corresponding to their activity they describe with free text. The paper presents also the quality monitoring system designed to analyse the results of this web function and to constantly update the informative base. Finally, it is outlined how this system, being generalised, could be used for other classifications so as to constitute a standard tool to navigate in the classifications data bases.

Keywords: automatic coding, textual matching, Economic Activity classification.

[1] Il lavoro è frutto dell'attività congiunta degli autori. In ogni caso, ai soli fini dell'attribuzione, i capitoli 2 e 5 e paragrafi 4.1 e 4.3 sono da attribuirsi ad Angelina Ferrillo paragrafi 1.1, 1.3, 3.1 e 6 a Stefania Macchia, il paragrafo 3.3 a Loredana Mazza, il paragrafo 3.2 Alberto Valery e il paragrafo 1.2 a Paola Vicari.

Indice

1. L'applicazione ACTR su Web

1.1 Il sistema di codifica ACTR e il suo utilizzo in Istat

ACTR (Automatic Coding by Text Recognition) è un sistema che consente automaticamente l'attribuzione di codici, secondo classificazioni predefinite, ai dati rilevati tramite quesiti a testo libero. Progettato e commercializzato da Statistics Canada, è ampliamente utilizzato non soltanto in Istat, che se ne avvale per diverse classificazioni in numerose indagini, ma anche in diversi Istituti Nazionali di Statistica.

E' un sistema generalizzato (indipendente dalla lingua e dalla classificazione di riferimento), quindi sono a carico dell'utilizzatore la costruzione della base informativa per ciascuna classificazione e l'adattamento alla lingua.

L'individuazione dei codici da associare alle descrizioni avviene tramite un processo in *batch* che realizza il *matching* tra testi da codificare e quelli della base informativa di riferimento avvalendosi di una metodologia che rientra tra i cosiddetti *weighting algorithms;* questi algoritmi, in sintesi, individuano *match* esatti o parziali sulla base di funzioni di similarità tra i testi, dove alle parole è attribuito un peso, empirico o probabilistico, proporzionale al loro grado di informatività.

L'attività di codifica è preceduta da una fase di standardizzazione dei testi, chiamata *parsing*, per la quale ACTR fornisce 14 differenti funzioni quali, ad esempio: la mappatura dei caratteri, la cancellazione delle parole inutili, la definizione di sinonimi, la rimozione di suffissi/prefissi, ecc.. In sintesi, il *parsing* ha l'obiettivo di rimuovere le differenze grammaticali e/o sintattiche in modo da rendere uguali due descrizioni con lo stesso contenuto semantico.

Il testo sottoposto al *parsing* viene quindi confrontato con i testi della base informativa, che hanno a loro volta subito lo stesso trattamento. Se da questo confronto emerge un abbinamento esatto (*direct match*), viene assegnato un unico codice, altrimenti il sistema utilizza un algoritmo per individuare il *match* 'più simile'. A seguito di una misura della similarità tra i testi messi a confronto e al confronto di tale misura con appositi parametri soglia, definiti dall'utente, ACTR produce i seguenti possibili risultati:

- *match* unico, se viene assegnato un singolo codice al testo da codificare;
- *match* multipli, se viene individuata una serie di possibili codici corrispondenti al testo da codificare;
- *match* fallito, se non è possibile alcun *match*.

In Istat ACTR è stato ed è tuttora utilizzato per codificare i dati di numerose indagini nelle quali vengono rilevate, con quesiti a testo libero, variabili da ricondurre alle seguenti classificazioni ufficiali:

- Professione;
- Attività economica;
- Comune/Provincia;
- Stato estero/Cittadinanza;
- Titolo di studio;
- Cause di morte.

Relativamente all'attività economica (ATECO 2007), è stata implementata nel corso degli anni una base informativa molto ricca, costituita da un dizionario (*reference file*) di oltre 33.000 descrizioni associate ai codici della classificazione (1.893 voci) e dai file di *parsing* per un totale di circa 15.680 sinonimi.

Sono state codificate tramite ACTR le risposte testuali fornite nelle seguenti indagini:

- Censimento intermedio dell'industria;
- Indagine di qualità del censimento della popolazione 1991;
- Prima indagine pilota sulle Forze di Lavoro;
- Prima indagine pilota sul Censimento della Popolazione 2001;
- Seconda indagine pilota sul Censimento della Popolazione 2001;
- Censimento della popolazione (modello CP29);
- Censimento dell'Industria 2001;
- Indagini Eusilc (anni 2008, 2009 e 2010).

Gli indicatori abitualmente utilizzati per misurare la qualità dei risultati della codifica automatica sono due:

- tasso di codifica → percentuale dei codici assegnati automaticamente sul totale dei testi sottoposti a codifica;
- tasso di precisione o accuratezza → percentuale dei codici corretti assegnati automaticamente sul totale dei testi codificati..

Come può vedersi dalla tabella 1, i tassi di codifica ottenuti variano da un minimo del 43,5% ad un massimo dell'80%, a seconda che si trattasse di dati rilevati su famiglie/individui (per le quali si rileva una particolare difficoltà nel rispondere al quesito) o sulle imprese.

Tabella 1 - Performance della codifica automatica delle attività economiche sulle indagini Istat

	L'applicazione di codifica automatica delle attività economiche		
	Testi da codificare	Tasso di codifica	Accuratezza
Censimento Intermedio dell'Industria	1.793	58,8	91,0
Indagine Qualità Censimento Popolazione 1991	6.288	54,5	85,0
I Indagine Pilota Forze di Lavoro	-	43,5	85,0
I Indagine Pilota Censimento Popolazione 2001	-	51,2	93,7
II Indagine Pilota Censimento Popolazione 2001	-	51,9	90,0
Censimento Popolazione 2001 (Convivenze)	-	53,6	92,31
Censimento dell'Industria 2001	1.130.693	80,7	-
Eusilc 2008 (Ateco Attuale)*	1.068	48%	
Eusilc 2008 (Ateco Precedente)*	990	56%	
EuSilc 2009 (Ateco Attuale)*	18.236	72%	
EuSilc 2009 (Ateco Precedente)*	14.034	73%	
EuSilc 2010 (Ateco Attuale)*	17.037	72%	
EuSilc 2010 (Ateco Precedente)*	12.867	73%	

(*) Nell'indagine Eusilc vengono rilevate sia l'Ateco attualmente espletata dal rispondente che quella relativa alla eventuale precedente attività.

1.2 Perché mettere l'applicazione di codifica a disposizione degli utenti Web

Il progetto di mettere a disposizione di tutti gli utenti lo strumento di codifica automatica dell'ATECO è stato concepito nel 2004 dopo aver aggiornato il sistema ACTR con la classificazione ATECO 2002. Lo strumento ormai era stato utilizzato nella sua forma *batch* in molte indagini e ne era stata apprezzata l'utilità sia per la classificazione delle attività economiche sia per altre classificazioni. Già con la nuova versione del 2002 si era visto che la pagina del sito Istat dedicata all'ATECO era molto consultata e quindi sembrava opportuno mettere a disposizione degli utenti quanti più strumenti possibile. Soprattutto, dopo ogni aggiornamento della classificazione, anche i semplici cittadini hanno bisogno di trovare una risposta ai loro quesiti.

Per la prima volta, con la nuova classificazione del 2007, l'ATECO è unica e utilizzata da tutti gli Enti interessati (Agenzia delle Entrate, Camere di Commercio, Inail e Inps). In precedenza l'Agenzia delle Entrate pubblicava su Gazzetta Ufficiale l'ATECOFIN che differiva dall'ATECO per alcuni maggiori dettagli individuati dalle lettere dell'alfabeto al posto della quinta cifra numerica; le Camere di Commercio applicavano l'ATECORI nella quale erano presenti diversi ulteriori dettagli, rispetto all'ATECO, individuati da seste cifre numeriche. In quest'ultimo caso anche l'interpretazione della classificazione non era esattamente la stessa; ad esempio le Camere di Commercio classificavano le pizzerie a taglio nel Manifatturiero mentre l'ISTAT le classificava nella Ristorazione. Uno dei compiti del sotto Comitato ATECO[2] era definire l'interpretazione comune della classificazione che, in alcuni punti, poteva risultare ambigua o non chiara, anche per problemi di traduzione dall'originale inglese.

[2] Il sotto Comitato è stato costituito all'interno del Comitato ATECO con il mandato di discutere la classificazione e di definire la versione italiana della classificazione. Affinché le riunioni fossero proficue, il sotto Comitato era costituito da un gruppo ristretto di esperti di ISTAT, Unioncamere, Agenzia delle Entrate (Studi di Settore) e Inps più esperti di settori particolari convocati di volta in volta.

I primi dati statistici con la nuova classificazione sono stati diffusi nel 2009 ma, per gli adempimenti fiscali, essa era già in vigore dal 1 gennaio 2008. Il sito dell'Agenzia delle Entrate era collegato direttamente alla pagina dell'ISTAT per quanto riguardava l'ATECO 2007; diventava pertanto prioritario per l'ISTAT fornire più strumenti possibili per la consultazione dell'ATECO. La pagina relativa alla classificazione ATECO risultava essere quella più consultata dopo quella relativa ai prezzi.

I primi adempimenti fiscali dell'anno cadevano a fine febbraio; in quel periodo la casella di posta elettronica dedicata ha ricevuto molti quesiti. ACTR su Web è stato realizzato anche pensando all'alleggerimento del lavoro quotidiano per gli esperti codificatori dell'ISTAT; infatti, dopo la messa in linea dell'applicazione, gli interrogativi inviati all'apposita casella di posta elettronica ATECO 2007@ISTAT.it sono notevolmente diminuiti.

L'utilità di una tale funzione sul sito WEB istituzionale dell'Istat,[3] inoltre, può andare oltre la specifica esigenza espressa per l'ATECO. In merito alle rilevazioni sulle imprese, per esempio, nei questionari di alcune indagini è riportata per ciascuna azienda l'attività economica risultante dagli archivi delle imprese e si richiede agli intervistati di aggiornarla se non corretta, oppure se variata. In presenza di tale funzione, nei questionari stessi si può riportare il riferimento al sito Web cui accedere per descrivere l'attività economica espletata e verificarne il codice corrispondente.

L'altro obiettivo fondamentale di questa applicazione, per l'ATECO ed in prospettiva per le altre classificazioni, è che i dizionari di queste possono essere aggiornati sulla base di un ritorno controllato delle richieste più significative effettuate dagli utenti. Questa attività richiede una gestione non eccessivamente onerosa ma costante di ciò che perviene dall'esterno e rappresenta forse l'implicazione più interessante offerta dal progetto, in quanto, una volta realizzata, consente di tener conto di tutto un complesso di informazioni aggiuntive sulla cui base operare aggiornamenti ad integrazione della classificazione standard delle attività economiche, necessità che si presenta con una cadenza interna più frequente delle versioni ufficiali delle classificazioni stesse.

1.3 L'applicazione di codifica sul Web

Il migrare in ambiente Web l'applicazione di codifica utilizzata in *batch* per codificare i dati rilevati nelle indagini (chiameremo quest'ultima ACTR_indagini) ha reso necessari alcuni interventi sia di tipo tecnico che derivanti dall'analisi delle esigenze dei potenziali utenti che accedono al sito.

Infatti mentre la finalità principale di un'applicazione *batch* è di massimizzare i codici univoci assegnati automaticamente, un utente che consulta il sito Web per individuare il codice ATECO corrispondente all'attività economica da lui espletata può trarre vantaggio dal poter analizzare diverse descrizioni di attività (corrispondenti a diversi codici) affini a quella espletata e di selezionare quindi, tra queste, quella a lui più attinente.

Inoltre, mentre nel codificare i dati di un'indagine, può essere di utilità anche individuare un codice non al massimo dettaglio, qualora la descrizione fornita sia generica (ad esempio perché i dati vengono pubblicati ad un certo livello di dettaglio, oppure perché ci si può avvalere di ulteriori informazioni disponibili sul questionario di rilevazione per completare un codice generico), un utente che consulta il sito Web per individuare il codice ATECO corrispondente all'attività economica da lui espletata ha bisogno necessariamente del codice al massimo livello di dettaglio.

Per queste due motivazioni sono state introdotte le seguenti varianti all'applicazione su Web che influiscono sui risultati, differenziandola dall'applicazione *batch*:

- sono stati modificati i parametri soglia utilizzati da ACTR per misurare la similarità tra i testi ed individuare il codice, in modo da aumentare le possibilità che, laddove la descrizione fornita non realizzi un *direct match*, il sistema proponga un ventaglio di codici possibili piuttosto che uno solo corrispondente alla descrizione più simile a quella fornita;
- è stato elevato a 7 il set di codici con le corrispondenti descrizioni proposte dal sistema (il parametro utilizzato nel *batch* è usualmente 5);

[3] http://www3.istat.it/strumenti/definizioni/ateco/atecoactr.php

- non è stato messo in linea il dizionario completo utilizzato da ACTR per l'applicazione *batch*, ma un estratto di quest'ultimo contente esclusivamente le descrizioni corrispondenti ai codici a cinque digit;
- questo dizionario è stato collegato ad una tabella, contente le descrizioni corrispondenti ai codici a sei digit, tale che l'utente, una volta individuato il codice a cinque digit pertinente alla propria attività, possa visualizzare anche i codici a sei digit (con le relative descrizioni) corrispondenti a quest'ultimo.

Il fatto di inibire l'attribuzione di un codice che non sia al massimo dettaglio è stato inoltre gestito non soltanto fornendo all'utente indicazioni specifiche su come descrivere la propria attività economica (ad esempio non fornire descrizioni generiche, non utilizzare abbreviazioni che potrebbero risultare ambigue per il sistema, eccetera), ma anche tramite una messaggistica di errore che, in caso di mancata attribuzione del codice, rimandasse a suggerimenti su come esplicitare meglio la propria attività.

Dal punto di vista tecnico è stato necessario gestire aspetti non previsti dall'applicazione *batch* che gira in modalità *stand alone*, quali la gestione di più accessi contemporanei e il non poter utilizzare direttamente l'interfaccia grafica di ACTR.

Nella pratica, l'applicazione Web è costituita da pagine HTML dinamiche scritte in PHP e consta essenzialmente di due parti.

La prima parte comprende le pagine Web che si trovano sul server che ospita il sito dell'Istituto; si tratta delle pagine in cui vengono acquisite le *query* dell'utente, la pagina di pubblicazione dei risultati ottenuti con il sistema ACTR che fornisce un elenco di descrizioni delle possibili attività tra le quali l'utente deve poi scegliere da quest'elenco la descrizione che ritiene più vicina alla propria, la pagina di pubblicazione dei risultati finali che fornisce il codice a 6 cifre e il relativo titolo ufficiale della classificazione, nonché la pagina statica contenente i suggerimenti per un corretto utilizzo dello strumento.

La seconda parte dell'applicazione comprende i programmi in PHP che si trovano su un altro server, quello su cui gira il software ACTR. Questi programmi vengono chiamati in HTTP dalle pagine del sito e si preoccupano di costruire gli input per il software ACTR, eseguire il software, recuperare gli output, formattarli e farli pubblicare nella pagina dei risultati. Si noti che il software ACTR è tale che non può accettare più richieste in parallelo, per cui è stato necessario serializzare le richieste utente provenienti dalla pagina Web principale.

Inoltre, è stata introdotta un'ulteriore funzione, ossia la memorizzazione delle *query* effettuate dagli utenti per consentire di valutare la qualità dei risultati e per aggiornare sistematicamente la base informativa utilizzata dal sistema.

Tutte le *query* digitate dagli utenti vengono scritte in un file memorizzato nel server windows; con periodicità settimanale, in modalità automatica, il file viene inviato via e-mail ad una persona designata (esperto della classificazione) che si occupa di una prima analisi di questo file.

Per automatizzare questo processo è stato creato un demone in VBscript (ossia un processo sempre attivo) che con una fissata periodicità:

- verifica l'esistenza del file da inviare;
- crea e invia la e-mail con il file allegato;
- rinomina il file inviato aggiungendo nel nome l'informazione della data di invio.

Questi file così rinominati vengono storicizzati per un periodo di tempo opportuno e poi cancellati in modo automatico da un altro demone.

Come di seguito descritto, sono proprio questi i dataset dai quali si estrae periodicamente un campione ragionato per un'analisi della qualità del processo di codifica ed i conseguenti interventi di manutenzione della base informativa, sia in termini di correzione di eventuali errori che di arricchimento del dizionario.

2. Le finalità del monitoraggio della qualità dell'applicazione

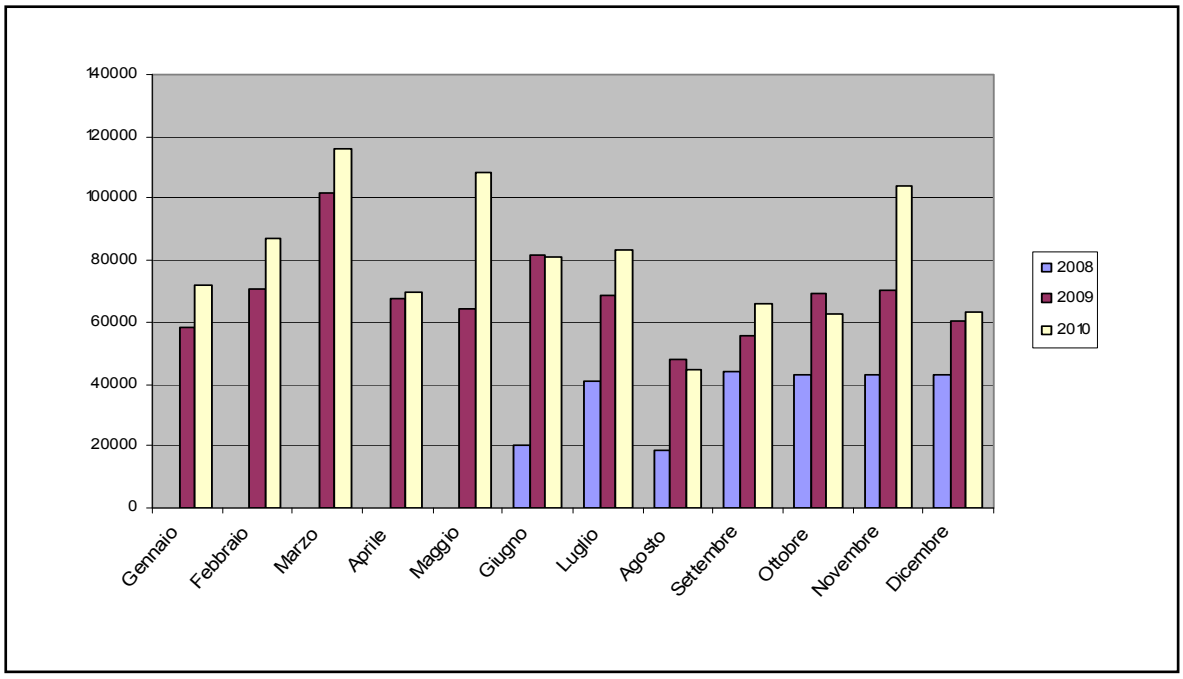
Come accennato nel paragrafo 1, già prima di mettere in produzione la funzione di codifica dell'ATECO su Web se ne prevedeva un'utenza ampia, sia dal punto di vista quantitativo che come tipologia di soggetti interessati.

In effetti, già dalle prime settimane è stata registrata una media di circa 10.000 *query* a settimana che, con il tempo, si è attestata attorno alle 15.000.

Tabella 2 - Numero Query mensili per gli anni 2008-2010

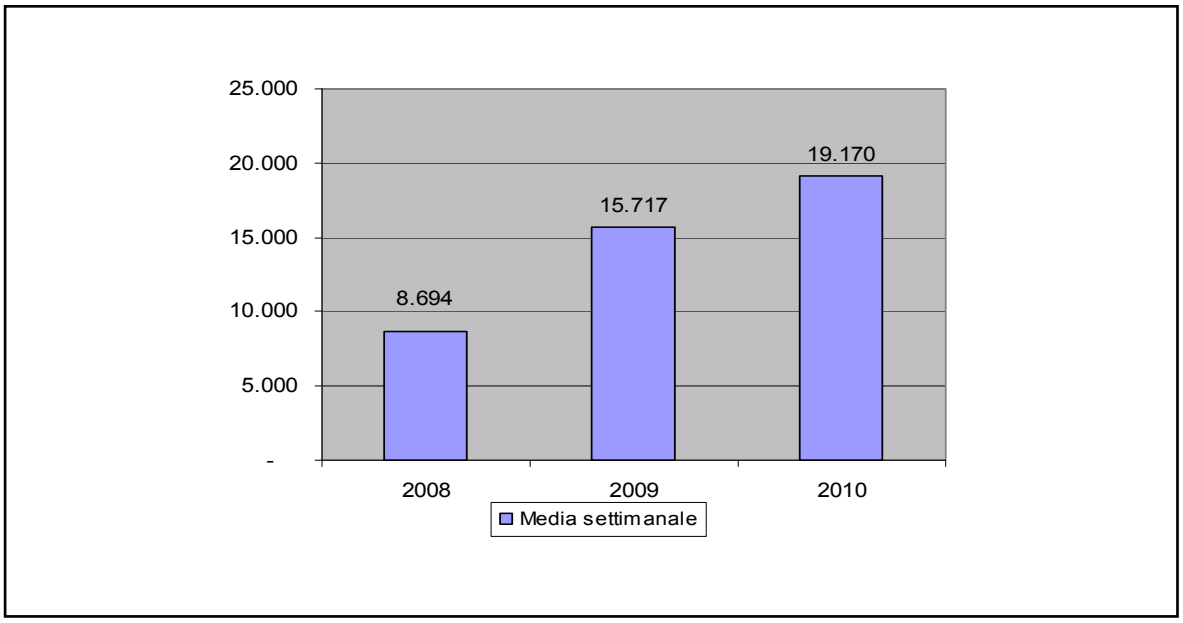
	Anni		
	2008	2009	2010
Gennaio		58.204	71.965
Febbraio		70.802	87.033
Marzo		101.990	116.091
Aprile		67.431	69.857
Maggio		64.514	108.193
Giugno	20.020	81.862	81.087
Luglio	40.618	68.858	83.454
Agosto	18.426	47.719	44.799
Settembre	43.950	55.581	66.162
Ottobre	42.927	69.323	62.838
Novembre	42.919	70.254	104.057
Dicembre	43.269	60.733	62.972
Totale	**252.129**	**817.271**	**958.508**
Media settimanale	**8.694**	**15.717**	**19.170**

Figura 1 - Numero query (collegamenti al sito Web Istat) per gli anni 2008-2010 per mese



Come può vedersi, l'utilizzo medio della funzione mostra una crescita costante di anno in anno.

Figura 2 - Media settimanale delle query per gli anni 2008-2010



Subito, quindi, si è ritenuto opportuno procedere al monitoraggio della qualità della funzione di codifica, per garantire da un lato in termini di correttezza dei codici assegnati e dall'altro per fornire elementi che consentano di aggiornare la base informativa utilizzata dall'applicazione così da sanare eventuali errori ed allinearla al modo di esprimersi degli utenti.

Gli indicatori abitualmente utilizzati per misurare la qualità dei risultati della codifica automatica, così come già riportato nel paragrafo 1.1, sono due:

- tasso di codifica → percentuale dei codici assegnati automaticamente sul totale dei testi sottoposti a codifica;
- tasso di precisione o accuratezza → percentuale dei codici corretti assegnati automaticamente sul totale dei testi codificati.

Nel caso in esame, però, il monitoraggio della qualità ha seguito un'ottica un po' diversa, dal momento che, come già detto, la finalità principale non è quella di massimizzare i codici univoci assegnati automaticamente, così come nelle applicazioni *batch*, ma di far sì che l'utente che consulta il sito Web per individuare il codice ATECO corrispondente all'attività economica da lui espletata possa analizzare diverse descrizioni di attività (corrispondenti a diversi codici) affini alla sua e di selezionare tra queste quella a lui più attinente.

Da un lato, quindi, si è tenuto sotto controllo che il tasso di codifica (individuazione univoca del codice) si mantenga a livelli coerenti con le applicazioni *batch*, dall'altro si è lavorato per tentare di escludere la possibilità che il sistema non proponga alcuna soluzione ad una descrizione fornita dall'utente qualora questa contenga un'informazione significativa al fine dell'attribuzione del codice.

A tal fine, utilizzando come input le *query* fornite dagli utenti, è stata messa a punto una procedura che, con una cadenza prefissata, espleta due funzioni:

- rielabora in *batch* il processo di codifica per verificare il tasso di codifica e per stimare l'accuratezza dei record codificati, quantificando il numero di *match diretti*
- analizza i casi non codificati univocamente per individuare eventuali descrizioni fornite dall'utente che riportino un significato utile per l'assegnazione del codice e possano quindi essere utilizzate per arricchire la base informativa.

3. La procedura di monitoraggio

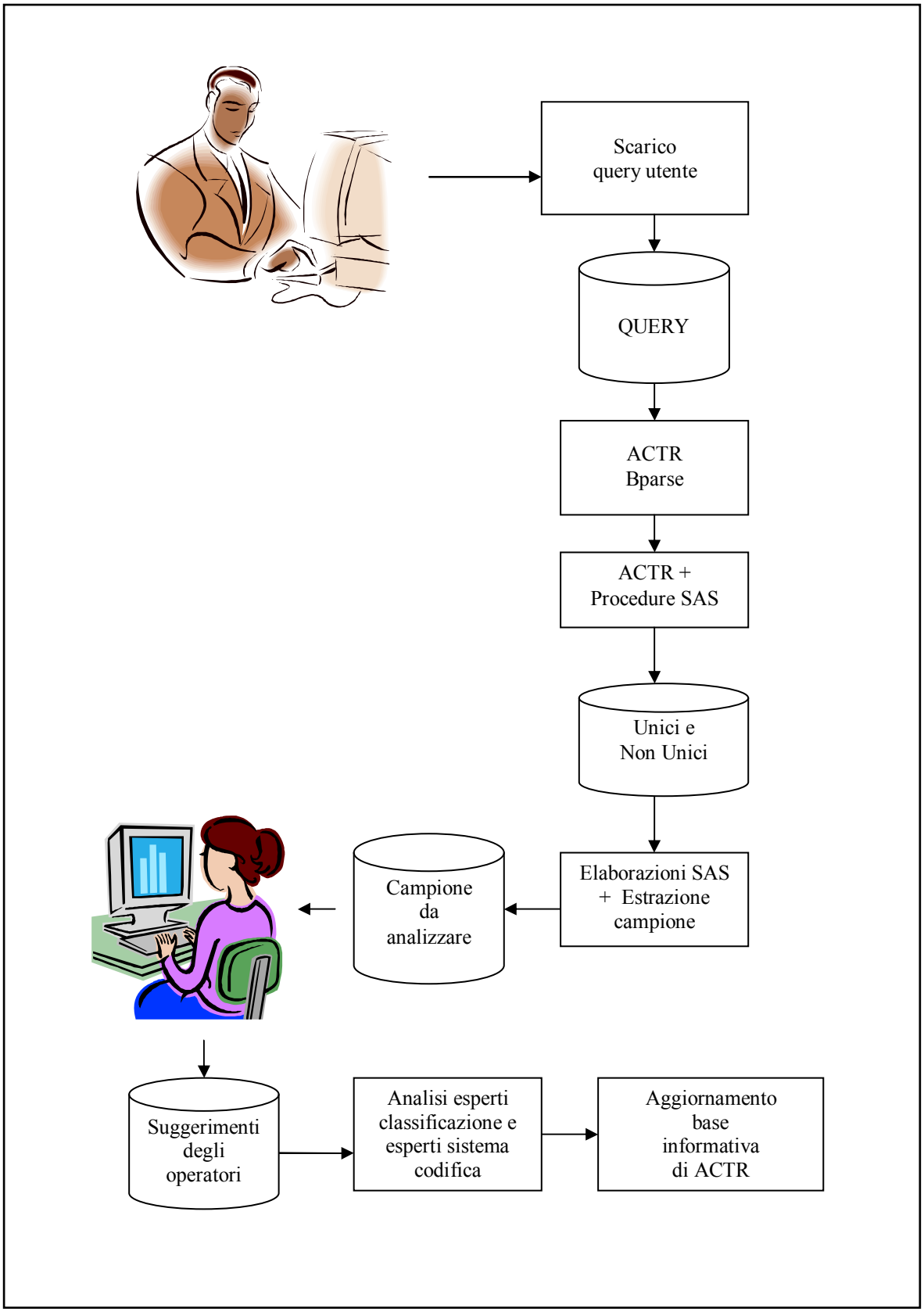
3.1. Gli aspetti tecnico/metodologici

La procedura di monitoraggio prevede una serie di passaggi e coinvolge diversi attori, come mostrato nella figura 3.

Si possono riassumere gli *step* come segue:

1. le *query* degli utenti vengono memorizzate e, con cadenza settimanale, inviate automaticamente ad un esperto della classificazione;
2. l'insieme di *query* collezionate nell'arco di uno o, al massimo, due mesi viene sottoposto ad un *parsing* semplificato, che ha il fine di rimuovere gli elementi che possono rendere diverse due descrizioni di fatto uguali (punteggiatura, articoli, preposizioni, genere e numero); l'output di detta procedura viene elaborato con un programma SAS creato ad hoc che consente di ripulire il file dai record vuoti, da quelli che presentano solo codici numerici, dai record il cui contenuto informativo risulta minore di tre lettere, ecc. Viene quindi prodotto un file che contiene sia le descrizioni originarie che quelle processate a seguito del *parsing* semplificato;
3. viene calcolata la frequenza di descrizioni uguali che vengono aggregate in un unico record (nel record, che aggrega *n* descrizioni iniziali, ricondotte ad un'unica descrizione a seguito del *parsing* semplificato e delle elaborazioni SAS, è riportata la prima delle descrizioni originarie che, a seguito del *parsing*, è diventata uguale agli altri *n* testi);
4. il file così elaborato viene sottoposto ad ACTR in *batch*, ottenendo così i file: Unici, Multipli, Possibili e Falliti;
5. i file di output vengono elaborati con programmi SAS creati ad hoc, ottenendo un unico dataset; poi, vengono eliminate le descrizioni già controllate nelle precedenti occasioni di monitoraggio, accoppiando questo file con quello dei testi già analizzati fino a quel momento;
6. dal file così ottenuto viene estratto un campione da sottoporre a verifica, la cui dimensione è definita in funzione del carico di lavoro attribuibile agli operatori da dedicare all'analisi. L'estrazione tiene conto della classe di frequenza delle descrizioni, favorendo le classi più ampie; in particolare non sono mai stati estratti testi con classi di frequenza inferiori a 9. Finora l'estrazione non ha incluso gli Unici, perché si è voluto privilegiare l'analisi dei non Unici in modo da favorire interventi di aggiornamento che massimizzano il tasso di codifica;
7. il campione così ottenuto viene inviato ad alcuni operatori appositamente formati (cfr, par. 3.3), perché questi diano indicazioni sull'eventuale possibilità di associare un codice ATECO a queste descrizioni;
8. il frutto dell'analisi di questi operatori viene restituito agli esperti della classificazione e agli esperti del sistema di codifica, che validano il lavoro degli operatori, apportano le opportune modifiche e lo utilizzano per l'aggiornamento della base informativa;
9. le osservazioni di questi esperti vengono comunque restituite agli operatori che ne tengono conto per le analisi da effettuare nel periodo successivo.

Figura 3 - la procedura di monitoraggio



3.2. Le attività degli operatori

Relativamente alle attività degli operatori, ad essi viene fornito un file così costituito:

- identificativo del record originario;
- frequenza della descrizione;
- descrizione;
- flag da apporre (C = codificabile, I = non codificabile);
- codice proposto;
- testo proposto;
- note.

Gli operatori devono analizzare la descrizione fornita. Viene data loro informazione sulla frequenza della descrizione stessa perché possano tenere conto di quanto un eventuale intervento sulla base informativa da loro proposto potrebbe impattare sull'applicazione in produzione (è presumibile che la frequenza di una descrizione registrata in un certo periodo non sia casuale, ma configuri una tendenza).

Per effettuare l'analisi dei testi gli operatori possono procedere:

- utilizzando ACTR stesso, nella versione online; a tal fine è stata messa a loro disposizione un'applicazione ACTR in cui i valori dei parametri soglia sono stati abbattuti, impostandoli in modo tale che il sistema proponga in visualizzazione tutti i testi del dizionario che abbiano almeno una parola in comune col testo da codificare. In tal modo gli operatori sono facilitati nel verificare a quali codici possano corrispondere i diversi elementi del testo digitato, nonché verificare quali trasformazioni subisca il testo a seguito del *parsing*;
- consultando il manuale ufficiale della classificazione per chiarimenti di tipo definitorio e contenutistico;
- consultandosi l'un l'altro per confrontare le interpretazioni di ciascuno.

La prima discriminazione che si richiede loro a seguito di questa analisi è di dichiarare, tramite il flag, se la descrizione sia o meno codificabile.

Qualora lo sia, l'operatore deve inserire nel campo 'testo proposto' la descrizione che suggerisce di aggiungere nel dizionario e/o nel campo 'note' l'eventuale sinonimo con cui arricchire le regole di *parsing*.

Infatti, non è detto che sia opportuno inserire nella base informativa una descrizione così come fornita dall'utente finale; si deve infatti tenere in considerazione che la descrizione eventualmente inserita è funzionale al fatto di riuscire a codificare non soltanto il testo in esame, ma testi simili, caratterizzati dallo stesso significato, che possono essere espressi anche in forma leggermente diversa. Potrebbe quindi essere opportuno rendere il testo più 'pulito' dal punto di vista grammaticale o sintattico, oppure, qualora si verifichi che la non individuazione di un codice univoco sia dovuta alla mancanza di un sinonimo, limitarsi a proporne l'inserimento.

Inoltre, anche nel caso in cui l'operatore dichiari che il testo non sia univocamente codificabile, ossia esprima un concetto generico associabile a più di un codice, potrebbe comunque proporre delle integrazioni della base informativa. In tal caso, dovrebbe verificare, utilizzando ACTR, se l'applicazione già produca un Multiplo e, in particolare, contempli tutti i codici a cui il concetto generico può far riferimento; qualora non fosse così, l'operatore dovrebbe proporre una o più descrizioni (associate ad altrettanti codici) e/o uno o più sinonimi.

Infine, il campo 'note' può essere utilizzato dall'operatore per mettere in luce eventuali errori del sistema di codifica, in termini, per esempio di sinonimi non corretti o non generalizzabili rispetto a tutti i codici.

3.3. La formazione degli operatori

Per le attività sopra descritte sono state individuate persone abitualmente adibite ad effettuare controlli di qualità sulla registrazione (in forza presso l'unità operativa "Controlli di qualità della registrazione e della codifica dei dati e test dei questionari"), che si sono dimostrate interessate ad ampliare le proprie conoscenze per realizzare un controllo di qualità che richiedeva competenze su una particolare tematica, quale l'ATECO.

A tale scopo è stato predisposto un corso interno di formazione su *"L'analisi di qualità nell'attribuzione del codice ATECO 2007 tramite un software di codifica automatica"*. Il corso, della durata di tre giorni (18-20 novembre 2009), è stato tenuto a Roma presso la sede centrale di via Balbo. La formazione è stata effettuata in due momenti con il duplice scopo di:

- esplicitare i criteri guida della codifica dell'attività economica secondo la nuova Classificazione ATECO 2007;
- illustrare le peculiarità del software di codifica automatica ACTR in maniera tale da formare il partecipante a saper trattare le descrizioni delle attività economiche fornite dagli utenti.

I docenti che si sono alternati nella formazione sono stati quindi sia gli esperti della Classificazione ATECO 2007 che quelli del software di codifica automatica ACTR.

Come già detto, il corso si è articolato in due momenti di formazione. Nel primo, curato dagli esperti della **Classificazione ATECO 2007**, ci si è preoccupati di definire i **principi della Classificazione, la sua struttura e i contenuti.**

In dettaglio è stata data, quindi, la definizione di *"attività economica"* intesa come '*la combinazione di differenti risorse, quali materie prime, prodotti intermedi, attrezzature, lavoro, tecnologie, che da luogo alla produzione di specifici beni o servizi destinati a terzi*'. Le unità impiegate nello stesso tipo di attività economica sono classificate con lo stesso codice ATECO.

E' stato quindi trattato il tema della 'prevalenza' in caso di più di una attività; infatti, poiché una unità può svolgere una o più attività, descritte in una o più categorie della classificazione, è necessario individuare l'attività prevalente che normalmente viene stabilita in base al contributo maggiore sul valore aggiunto totale (VA). In assenza di informazioni sulla composizione in termini di VA delle diverse attività svolte da una unità, si sono utilizzate le seguenti regole empiriche di prevalenza (il segno '>' sta ad indicare la prevalenza):

- attività manifatturiere (produzione)>commercio;
- attività manifatturiere (produzione)>installazione;
- attività manifatturiere (produzione)>riparazione;
- attività manifatturiere (produzione)>manutenzione;
- installazione>riparazione;
- riparazione>manutenzione;
- commercio all'ingrosso>commercio al dettaglio;
- albergo>ristorante;
- ristorante>bar;
- albergo>terme.

E' stata inoltre descritta la struttura della Classificazione ATECO 2007, che è articolata in sei livelli, comprendenti, rispettivamente, le voci identificate da un codice:

1. alfabetico di una lettera (**Sezioni)**;
2. numerico a due cifre (**Divisioni**);
3. numerico a tre cifre (**Gruppi**);
4. numerico a quattro cifre (**Classi**);
5. numerico a cinque cifre (**Categorie**);
6. numerico a sei cifre (**Sottocategorie**).

La struttura di classificazione è ad "albero" e parte dal livello 1, il più aggregato, distinto in 21 sezioni fino a giungere al livello massimo di dettaglio, punto 6, comprendente 1.224 sottocategorie.

La classificazione è standardizzata a livello europeo fino alla quarta cifra, mentre le categorie e le sottocategorie (rispettivamente livello 5 e 6) possono differire tra i singoli Paesi per meglio cogliere le specificità nazionali.

I docenti hanno illustrato in dettaglio le 88 Divisioni dell'ATECO 2007, mettendo in risalto per ognuna:

- le varie attività produttive che ricadono nella relative Sezioni;
- le attività che ne sono escluse;
- le eventuali eccezioni presenti all'interno delle Divisioni.

Come strumenti di consultazione dell'ATECO 2007 sono stati forniti ai partecipanti al corso sia il volume della Classificazione ufficiale delle attività economiche 2007, che l'applicazione di codifica automatica tramite ACTR. Per la consultazione del volume ci si è soffermati soprattutto, nella fase di formazione, sulla ricerca per parola chiave all'interno dell'elenco alfabetico delle voci comprese nelle sottocategorie delle attività economiche, mentre la formazione sull'utilizzo della procedura di codifica interattiva con ACTR è stato oggetto della seconda parte del corso.

Nella seconda parte del corso, prima di entrare in merito all'uso del **modulo di codifica con ACTR**, ci si è soffermati sui vantaggi che comporta il processo di automatizzazione della codifica rispetto alla codifica manuale. Tramite esso infatti:

- si velocizza l'attività;
- si riducono i costi di formazione del personale;
- si evitano interpretazioni soggettive dei testi;
- si possono includere nella base informativa alcuni criteri classificatori;
- si ottengono standard qualitativi di codifica più elevati.

Si è passati quindi a descrivere il software utilizzato per mettere a punto il contesto di codifica automatica, che, come già detto nel capitolo 1, è ACTR. Esso si ispira alla metodologia sviluppata presso US Census Bureau ed utilizza gli algoritmi d'abbinamento di dati testuali messi a punto da ricercatori di Statistics Canada.

Il contesto di codifica della variabile ATECO implementato è costituito da:

- un dizionario informatizzato;
- una serie di file ausiliari (*parsing)* utilizzati per rimuovere stringhe ininfluenti, definire sinonimi a livello di stringa, rimuovere spazi multipli, rimuovere parole ininfluenti come articoli, preposizioni…, definire sinonimi a livello di parola o di coppie di parole, rimuovere suffissi/prefissi;
- alcuni parametri 'soglia' che il software utilizza per definire il livello di 'similitudine' tra il testo da codificare e quello contenuto nel dizionario.

E' stato descritto il processo per la costruzione dell'ambiente di codifica, il cui punto di partenza è il manuale ufficiale della classificazione (in questo caso logicamente *Classificazione delle attività economiche ATECO 2007*). Il manuale è stato rielaborato per renderlo leggibile dal software ed arricchito con risposte empiriche fornite dagli intervistati nel corso di precedenti rilevazioni. Sono stati predisposti i file di *parsing* per adattare l'applicazione alla lingua e al contesto applicativo. E' stato deciso quali funzioni usare ed in quale ordine sottometterle nella *parsing strategy*; inoltre è stata adottata la scelta di utilizzare i file di *parsing* per inserire regole classificatorie, intese come prevalenze, ovvero criteri generali validi in tutto il contesto di codifica come ad esempio: produzione>commercio; coltivazione>commercio; commercio>manutenzione; commercio>riparazione (cfr. regole empiriche di prevalenza sopra esplicitate).

Si è provveduto ad esporre come si realizza la fase di codifica vera e propria: la risposta testuale viene confrontata con il database (dizionario informatizzato standardizzato) alla ricerca di un abbinamento esatto (*direct match*), ma se il tentativo fallisce viene ricercato un abbinamento parziale (*indirect match*). In questo caso il software individua tramite una misura di similarità tra testi di tipo empirico il/i codice/i del dizionario con descrizione più simile alla risposta fornita dal rispondente. La misura di similarità (S) è normalizzata ed assume valori compresi nell'intervallo [0,10] dove, logicamente, se S=0 abbiamo un abbinamento nullo, se S=10 abbiamo un abbinamento esatto. La regione di accettazione per la misura di similarità è costruita utilizzando tre parametri soglia: Smin, Smax; e ΔS (minima distanza richiesta tra testo a punteggio massimo e il successivo). La strategia di codifica si estrinseca nella definizione dei suddetti parametri; si parlerà allora di 'strategia standard' (parametri 8, 6, 0.2) utilizzata nella codifica *batch* e 'strategia *deep*' (parametri 10, 0, 0) per la navigazione interattiva nell'ambiente di codifica.

A fine corso, per consentire agli operatori di effettuare con successo tutte le attività descritte nel paragrafo precedente, si è provveduto a fornire agli stessi, oltre al manuale della classificazione ufficiale ATECO 2007, il 'progetto' di codifica automatica interattivo, installandolo sui PC di ciascuno; tale progetto utilizza come database il dizionario informatizzato completo, che prevede cioè

anche codici non al massimo dettaglio (codici previsti< 5 digit), con la strategia *deep,* in maniera tale da fornire agli operatori una casistica di codifica più completa.

Contemporaneamente alla formazione teorica e tecnica sono state inoltre previste, all'interno delle giornate di corso, delle esercitazioni in aula con il duplice scopo di far familiarizzare i partecipanti con la classificazione e contemporaneamente di valutare l'apprendimento degli stessi. Durante le esercitazioni in aula si è provveduto a dividere i partecipanti in gruppi fornendo loro degli esempi di descrizioni digitate dagli utenti Web, nello stesso formato excel così come descritto nel paragrafo precedente, in modo che simulassero il lavoro che avrebbero successivamente dovuto espletare. A seguito di queste esercitazioni, si è discusso in aula sui risultati della eventuale codifica delle singole descrizioni fornite e sui vari dettagli tecnici proposti dagli operatori o che sarebbe stato opportuno proporre.

Considerata ancora la difficoltà della codifica della variabile ATECO, si è pensato di verificare le conoscenze acquisite nella prima tranche del corso tramite dei debriefing tecnici, a distanza di qualche settimana dal corso. A tale scopo, al termine del corso, è stata assegnata a gruppi costituiti da due partecipanti ciascuno una lista di descrizioni digitate dagli utenti Web non codificate con ACTR. Gli operatori dovevano, coerentemente con lo schema e le informazioni fornite durante il corso:

- verificare se i testi fossero effettivamente privi di contenuto informativo per assegnare un codice e quindi dovesse essere loro attribuito un flag come incodificabili (flag I);
- proporre l'inserimento di una empirica nel dizionario informatizzato con il relativo codice, qualora ritenessero il testo codificabile (flag C);
- segnalare eventuali sinonimi da prevedere nei file di *parsing*;
- segnalare eventuali errori ed incompatibilità di assegnazione dei codici da parte del sistema.

Il lavoro prodotto dagli operatori su queste descrizioni è stato successivamente analizzato dagli esperti della classificazione e del sistema di codifica che, nel corso di un apposito de briefing, hanno illustrato e discusso gli errori più frequenti e fornito ulteriori approfondimenti.

4 Il monitoraggio

4.1 Cicli di analisi effettuate (query a settimana e campioni estratti fino a novembre 2010)

Il file che accumula le query rilevate dal 9-6-2008 al 22-11-2010 è costituito da 1.931.678 record. Nel 2010, infatti, la media dei collegamenti settimanali al sito Web dell'Istat per l'individuazione del codice ATECO, è stata di 19.170, che corrispondono ad una media giornaliera di 2.718 query.

Per ogni ciclo di monitoraggio, come descritto nel capitolo 3, l'insieme di query rilevate fino a quel momento viene sottoposto ad un processo di *parsing* semplificato che consente l'eliminazione degli articoli, della punteggiatura e poco altro, al fine di individuare le descrizioni tra loro differenti. Questa procedura restituisce come output un file con la descrizione originale così come digitata dagli utenti Web, la descrizione normalizzata e il numero progressivo record.

Detto output viene elaborato con una procedura SAS creata ad hoc che consente di ripulire ulteriormente il file dai record vuoti, dai record che presentano solo codici numerici e da quelli il cui contenuto informativo risulta minore di tre lettere (tra cui vari segni tipo *, :, ", eccetera); vengono quindi eliminate le descrizioni che, a seguito di queste elaborazioni, sono uguali e a ciascuna di esse viene associata la frequenza con la quale è stata rilevata; si produce così un output con numero progressivo, testo originale digitato dagli utenti Web (corrispondente alla prima descrizione tra quelle aggregate in quanto uguali a seguito delle elaborazioni descritte) e frequenza del testo. Per fornire un ordine di grandezza, si riporta che il file così prodotto, corrispondente all'insieme del 1.931.678 query originarie, è costituito da 368.222 record.

Detto file viene quindi sottoposto a codifica automatica con l'applicazione *batch* speculare a quella Web e gli output di ACTR vengono elaborati con programmi SAS con lo scopo di avere un unico dataset con tutte le informazioni necessarie per l'estrazione del campione da analizzare ai fini del monitoraggio.

Poiché, come già detto, questo file si ottiene in maniera incrementale, aggiungendo di volta in volta le query rilevate nelle settimana, al fine di evitare che le stesse descrizioni vengano controllate più volte, con conseguente perdita di tempo da parte dei codificatori e perdita di interesse per il lavoro da fare, vengono eliminate le descrizioni già controllate così da ottimizzare la fase di analisi e controllo.

In pratica, per eliminare tali descrizioni, il file comprendente i testi rilevati dal rilascio dell'applicazione al tempo *t* e quello dei testi già analizzati nei precedenti cicli di monitoraggio vengono accoppiati per testo (testi normalizzati a seguito del *parsing*) e tolti i record già analizzati.

Dal file così ottenuto, viene estratto il campione da sottoporre a controllo per ciascun ciclo di monitoraggio.

Nella tabella 3 si riportano i campioni estratti nei cinque cicli di monitoraggio ed i criteri di estrazione adottati.

Tabella 3 - Cicli di estrazione e campioni estratti

DATA	N° descrizioni originali	Descrizioni differenti tra loro	Campione estratto (rk)	Criteri estrazione
09/11/2009	964.959	210.307	389	falliti, multipli, possibili con frequenza testo>39
14/12/2009	1.052.568	223.406	727	falliti e multipli con frequenza testo>15; possibili con frequenza testo>80
25/01/2010	1.147.533	239.476	701	Falliti con frequenza testo >9
17/05/2010	1.485.821	294.757	733	Falliti con frequenza testo >8
22/11/2010	1.931.678	368.222	701	Falliti con frequenza testo >8

Come si può notare, i campioni hanno sempre avuto una dimensione di circa 700 descrizioni, a meno di quello utilizzato nel primo ciclo per il quale, per far prendere confidenza a tutti i partecipanti al corso e farli lavorare singolarmente, è stato estratto un campione più piccolo, contenente Multipli, Possibili e Falliti con frequenza superiore a 39.

Per il secondo ciclo, invece, sono stati selezionati quattro codificatori che hanno lavorato in gruppi di due e l'estrazione è stata effettuata dai Falliti e Multipli con frequenza >15 e Possibili con frequenza >80.

Per il terzo, quarto e quinto ciclo, viste le difficoltà dei codificatori, soprattutto nell'individuare le segnalazioni sia di empiriche sia di trasformazione al *parsing* per le frasi che provenivano dai Multipli e Possibili, sono stati estratti solo i record dai Falliti con frequenza superiore a 8 o 9. La frequenza è stata decisa durante l'elaborazione dei file di output di ACTR in modo da estrarre sempre un campione di circa 700 record.

4.2 Il punto di vista degli operatori[4]

Come già descritto, per effettuare l'analisi del campione di query inviate per ciascun ciclo di monitoraggio è stato utilizzato il sistema di codifica automatica ACTR; si è infatti provveduto ad installare sulle postazioni degli operatori l'applicazione di codifica ATECO completa, che utilizza, cioè il dizionario comprendente le descrizioni associate ai codici a tutti i livelli di dettaglio. Si è provveduto inoltre ad impostare le soglie utilizzate per il *matching* in modo tale che quando l'utente inserisce una descrizione, ACTR estrae dal dizionario tutti i testi con almeno una parola in comune con quella digitata (*strategia deep*). Questa scelta è finalizzata a consentire all'operatore la navigazione nel dizionario in modo più agile ed esteso possibile.

Si riporta di seguito la descrizione fornita dagli operatori sulle modalità con le quali hanno effettuato l'attività di analisi, nonché su alcune difficoltà incontrate.

[4] Questo paragrafo raccoglie le riflessioni degli operatori addetti al monitoraggio, che hanno provveduto alla redazione dello stesso (S. Alunni, C. Dominici, M. R. Federico, M. M. Iacomino).

Laddove il sistema non consentiva l'associazione univoca di un codice a ciascuna attività, si è provveduto a consultare le note esplicative contenute all'interno del manuale ATECO 2007, al fine di individuare la soluzione più idonea tra quelle proposte.

Il citato manuale è stato altresì adoperato per evidenziare e codificare le attività economiche non contemplate all'interno del dizionario informatizzato, ovvero segnalare eventuali errori di ACTR nell'assegnazione dei codici, consentendo di ottimizzare l'implementazione del sistema di codifica stesso.

Si è dimostrato di notevole utilità anche il continuo e sistematico accesso al Web, orientato principalmente alla ricerca di acronimi, testi in inglese (p.e. RENDERING o SYSTEM INTEGRATOR) e/o espressioni inusuali (tipo PROTOTIPAZIONE).

La codifica ha presentato alcune difficoltà iniziali dovute alla scarsa esperienza posseduta dagli operatori nello specifico settore, inoltre l'inevitabile interruzione di cicli di lavoro, che non sono stati continuativi ma si sono alternati con una cadenza mensile o bimestrale, non ha agevolato l'acquisizione di sicurezza e padronanza dell'attività svolta.

E' risultato inoltre difficoltoso acquisire dimestichezza con la logica di trasformazione del *parsing*, oltreché attribuire un codice univoco a ciascuna attività analizzata.

E' stato particolarmente costruttivo ed utile il costante confronto e la collaborazione tra i colleghi impegnati nella fase di codifica, che hanno permesso di chiarire i dubbi (p.e. proporre un'empirica o suggerire un sinonimo al testo digitato dall'utente) e risolvere i problemi relativi alle situazioni più controverse.

4.3 Analisi del lavoro dei codificatori da parte degli esperti della classificazione e del software ACTR

Le proposte fornite dagli operatori addetti al controllo di qualità vengono quindi analizzate, supervisionate e validate dagli esperti della classificazione, in forza presso la Direzione dei Dati Amministrativi e dei registri Statistici. L'analisi consiste nella rielaborazione dei testi eventualmente proposti dagli operatori addetti al monitoraggio, nella validazione del codice ATECO, nonché nel riepilogo degli interventi inerenti:

- empiriche nuove da inserire;
- empiriche già presenti nel dizionario da correggere o eliminare;
- inserimenti o correzioni nei file di *parsing*.

Questo materiale viene quindi inviato agli esperti dell'ambiente di codifica del settore Metodi, strumenti e supporto metodologico che lo analizzano per verificare la congruenza con le regole di codifica già implementate nel sistema, ne apportano le eventuali modifiche e provvedono all'aggiornamento dell'ambiente di codifica.

Dopo ogni aggiornamento dell'ambiente di codifica, per verificare eventuali errori o incongruenze, gli esperti della classificazione provvedono ai seguenti controlli:

- accoppiamento tra il dizionario utilizzato dal sistema di codifica automatica e la tabella ATECO 5 digit;
- accoppiamento degli output della codifica prima e dopo l'aggiornamento dell'ambiente di codifica per l'estrazione delle descrizioni che prima risultavano uniche e dopo non più uniche, al fine di controllare (le frequenze più alte) se la codifica nuova risulta esatta.

Per il lavoro di analisi e supervisione da parte degli esperti è necessaria una conoscenza approfondita della classificazione e del software ACTR, poiché la bontà della codifica automatica è attribuibile "esclusivamente" alla costruzione di un buon dizionario e alla corretta standardizzazione del testo nei file di *parsing*.

La logica usata durante la supervisione è stata quella di lavorare il più possibile sulle trasformazioni da inserire nei file di *parsing* evitando, quando possibile, di inserire nuove empiriche.

Come già descritto, il lavoro effettuato dagli operatori sui campioni estratti per l'analisi di qualità viene supervisionato dagli esperti della classificazione e da quelli del sistema di codifica, prima di effettuare gli aggiornamenti dell'ambiente di codifica, in modo da vagliare l'appropriatezza delle osservazioni dei codificatori ed individuare l'intervento più opportuno.

I suggerimenti forniti, in termini di empiriche da aggiungere o modifiche dei file di *parsing*, sono diventati più pertinenti man mano che gli operatori hanno preso dimestichezza sia con la classificazione, sia con la logica alla base del sistema di codifica; tuttavia la rielaborazione da parte degli esperti del lavoro dei codificatori è essenziale per delineare il tipo di aggiornamento da effettuare.

Quella che segue è una descrizione, sicuramente non esaustiva, di problematiche che sono emerse dalla supervisione del lavoro svolto dai codificatori, da parte degli esperti:

- Una prima problematica che ci si è trovati ad affrontare, infatti, consiste nel fatto che non sempre, anche qualora i suggerimenti forniti dagli operatori fossero contenutisticamente validi, è stato possibile recepirli così come formulati; spesso un'analisi del loro impatto sull'ambiente di codifica ha, infatti, reso necessario un intervento diverso;
- Ad esempio:
  - la segnalazione da apportare ai file di *parsing* "unione comuni=ente locale" non è stata recepita in quanto si è ritenuto più corretto inserire l'empirica "84.11.1 unione di comuni" in modo che l'utente che digita la frase si possa riconoscere nell'attività esplicata;
  - la segnalazione dell'inserimento dell'empirica "85.51.0 - pilates" non può essere recepita così come segnalata, in quanto con il codice proposto si classifica "istruttore di pilates", ma "pilates" può significare anche "gestione centro per pilates" con codice ATECO 93.13.0 o anche "produzione commercio e noleggio macchine per pilates" con codici ATECO diversi tra loro; risulta quindi ovvio che la singola parola "pilates" non può essere inserita come empirica. Infatti, in questo caso si è deciso di intervenire sul *parsing*, segnalando la trasformazione della parola "pilates" in "fitness". In questo modo gli utenti che digitano la singola parola "pilates" vedranno a video una serie di codici tra cui scegliere la propria attività;
  - La segnalazione dell'inserimento dell'empirica "45.20.2 levabolli" è formalmente corretta, in quanto dalla consultazione di wikipedia risulta che il "levabolli" (detto anche tirabolli) è un particolare artigiano che ripara la carrozzeria delle automobili. Tuttavia, potendo identificare il "levabolli" come carrozziere, già presente nel dizionario, si è deciso di segnalare la trasformazione "levabolli=carrozziere" nei file di *parsing* piuttosto che inserire una nuova empirica;
  - La segnalazione di inserimento dell'empirica "*90.03.0* mosaicista" non si può prendere in considerazione in quanto il termine mosaicista indica una professione e il professionista è specializzato sia nella realizzazione di mosaici sia nel restauro degli stessi, quindi **dovendo inserire nel dizionario informatizzato solo attività economiche univoche e non ambigue** e visto che la sola parola "mosaicista" si può classificare sia nel codice segnalato dai codificatori sia nel codice 23.70.2, è stato deciso di segnalare due empiriche: "*23.70.*2 realizzazione di mosaici"; "*90.03.*0 restauro di mosaici";
- Correlato a quanto detto nei punti precedenti, è emerso che non sempre i meccanismi del *parsing* sono risultati chiari agli operatori che, non essendo esperti dell'ambiente di codifica, hanno trovato a volte difficoltà nell'individuare l'effetto a catena che un intervento sul *parsing* può comportare su tutta l'applicazione di codifica;
  - Ad esempio, la proposta "educatore cinofilo = addestratore cinofilo" non è stata recepita in questi termini in quanto "educatore" si trasforma in "educazione" e, visto che in base alla strategia messa a punto la trasformazione delle parole singole viene eseguita prima della trasformazione delle due parole, se si fosse inserita la proposta dei codificatori non si sarebbe ottenuto alcun effetto; quindi per poter trasformare "educatore cinofilo" in "addestratore cinofilo" è stato proposto di introdurre la trasformazione "educazione cinofilo = adde-

stratore cinofilo";

```
Parsing Results for Context "db_ateco":
Word Breaking ..... "EDUCATORE CINOFILO"
   Replacement Words ..... "EDUCAZIONE CINOFILO"
   Double Words ..... "ADDESTRATORE CINOFILO"
```

– Altro esempio riguarda la proposta "istruttore cinofile = addestratore cinofilo"; in questo caso per far trasformare anche "istruttore cinofile" bisogna segnalare la trasformazione di "cinofile=cinofilo" e "istruttore cinofilo=addestratore cinofilo"

```
Parsing Results for Context "db_ateco":

Word Breaking ..... "ISTRUTTORE CINOFILE"
   Replacement Words ..... "ISTRUTTORE CINOFILO"
   Double Words ..... "ADDESTRATORE CINOFILO"
```

In merito alla trasformazione di "*istruttore cinofilo*" in "*addestratore cinofilo*", bisogna precisare che la scelta è stata fatta per evitare di inserire altre empiriche con la parola "*istruttore*", in quanto nel dizionario informatizzato la stessa è associata a parole come tennis, sub, nuoto, ceramica, scuola guida, ecc, attività che si codificano con codice 85.51.0, 85.52.0, 85.53.0, ossia a livello di gruppo di attività economica (tre digit) nell' 85.5, mentre istruttore cinofilo si codifica con codice 96.09.0, che è completamente differente da quelli sopramenzionati; quindi per avere la certezza di una corretta codifica quando la parola "*istruttore*" è associata a "*cinofilo*", si è deciso di segnalare le due trasformazioni già citate;

– la trasformazione "*tacchificio* = *produzione tacchi*", così come segnalata dagli operatori non avrebbe prodotto nessun effetto in quanto la singola parola "*tacchi*" viene trasformata nel file `Replacement Words` in "*tacco*" quindi per avere l'effetto desiderato bisogna introdurre "*tacchificio = produzione tacco*";

- Un altro aspetto non di così facile comprensione per chi affronta per le prime volte una classificazione complessa come l'ATECO, è la gestione delle 'prevalenze' nel caso di più attività citate nella descrizione da codificare. Questo tema è stato affrontato nel corso di formazione, tuttavia, sebbene siano state individuate alcune regole generali, nella realtà sono possibili numerosissime eccezioni che soltanto con un'esperienza di lungo periodo possono essere ricondotte ad un quadro di insieme.
  Ad esempio:
  – descrizioni "`ristorante albergo centro benessere`" o "`albergo con annesso centro benessere`". Nella prima troviamo tre attività economiche con codici ATECO diversi:
    `56.10.1 ristorante`;
    `55.10.0 albergo`;
    `56.10.2 centro benessere`.
    Poiché la descrizione "`albergo ristorante`" si classifica con il codice "55.10.0 `alberghi`", come da specifiche della nota esplicativa del volume ATECO 2007 (inclusi quelli con attività mista di fornitura di alloggio e somministrazione di pasti e bevande), per consentire la codifica automatica era stata prevista la prevalenza dell'attività di albergo su ristorante (tramite i file di *parsing*) e, per far sì che la frase "`albergo centro benessere`" fosse codificata con il codice dell'albergo, si è deciso di segnalare la prevalenza di `albergo>centro benessere`.
    La seconda descrizione, in cui sono citate attività, rientra nella casistica della prima. Non sempre i suggerimenti degli operatori sono stati in linea con questi criteri di prevalenza.
  – Altre descrizioni che sono risultate problematiche sono le seguenti:
    `Commercio e assistenza personal computer`;

Commercio al dettaglio e assistenza computer;
Commercio all'ingrosso e assistenza hardware e software;
In questo caso è stata segnalata la prevalenza dell'attività di commercio, dettaglio e ingrosso su assistenza;
commercio>assistenza;
dettaglio>assistenza;
ingrosso>assistenza.

- Infine, in alcuni casi sono state proposte empiriche non univoche, ossia alle quali, a seconda dei contesti non deducibili dalla sola descrizione fornita dall'utente, sarebbe possibile associare più di un codice. In questi casi, quindi, il lavoro dell'esperto si è concretizzato:
  1. nel prevedere più di un'empirica, ciascuna corredata degli elementi di dettaglio che consentano l'individuazione di un singolo codice. In tal modo, nel caso di una descrizione fornita dall'utente che non contempli tali elementi, il sistema genera un set di multipli tra i quali l'utente stesso si potrà riconoscere;
  2. nel decidere di non considerare la segnalazione da parte dei codificatori. Ad esempio:
     - generica è l'empirica segnalata "90.02.0 impresario"; in questo caso, infatti, non era opportuno inserire in un dizionario un'empirica che avrebbe ricondotto ad una codifica univoca una dizione generica come "impresario", quindi sono state proposte alcune empiriche mancanti, quali "96.03.0 impresario di pompe funebri" e "90.02.0 impresario teatrale";
     - la segnalazione "10.84.0 produzione di aromi" non avrebbe ricondotto ad un codice univoco, quindi sono state segnalate anche le empiriche sotto-riportate:
       20.14.0 produzione di aromi sintetici;
       20.53.0 produzione di essenze e aromi naturali;
       10.84.0 produzione di aromi e spezie.
     - la segnalazione "25.62.0 taglio al plasma" non è un'empirica con codice univoco, quindi sono state segnalate anche le empiriche sotto-riportate:
       28.41.0 produzione di macchine utensili per il taglio al plasma;
       25.62.0 lavorazione metalli con taglio al plasma.
     - l'inserimento dell'empirica "94.99.1 organizzazione non governativa (ong)" non è stato preso in considerazione in quanto la descrizione indica una forma giuridica che può essere operativa in varie attività economiche.

5. I risultati del monitoraggio

5.1 Effetti del monitoraggio/aggiornamento dell'applicazione di codifica in termini di tassi di codifica ottenuta

Come descritto nel paragrafo 4.1 i codificatori dovevano indicare con un flag se la descrizione era codificabile o incodificabile. Dall'analisi di questo flag è emerso che delle 2.550 descrizioni controllate e supervisionate il 23,5%, con le opportune correzioni al sistema di codifica, risultava essere codificabile, mentre il 76,5% non codificabile.

Infatti, dall'analisi effettuata, sono emerse diverse tipologie di casi non codificabili, tra le quali si citano di seguito le più frequenti:

- esplicitazione di una singola parola, priva del contenuto informativo per l'individuazione di un codice (ad esempio escavatore, smerigliatrice, eccetera senza specificare l'attività svolta esempio produzione, vendita, noleggio, commercio);
- fornitura della descrizione della propria professione che, nella maggior parte dei casi, non è riconducibile ad un codice univoco di attività economica (ad esempio tecnico can-

`tiere, tecnico della prevenzione, addetto alla sicurezza operatore personal computer`, eccetera);

- descrizioni assolutamente carenti in termini di informatività (ad esempio `altre, attività di gestione, porta, agente,affitto, ateco, attività di gestione, attività di libero professionista`, eccetera).

Nella tabella 4 sono riportate, a titolo esemplificativo, alcune descrizioni non codificabili, con la relativa frequenza con cui sono state digitate nell'applicazione Web.

Tabella 4 - esempio di query frequenti e non codificabili

Frequenza	Testo digitato dagli utenti Web
3130	Software
3672	Altri servizi
4281	Associazione
4565	Appalti di costruzioni
4669	Attivia di servizi alle imprese
4757	agricoltura
9952	Attivita consulenza

Nonostante il numero elevato di testi incodificabili, l'analisi di quelli codificabili ha portato ad un arricchimento della base informativa di:

- 457 righe di trasformazioni nei vari file di *parsing*;
- 408 empiriche nuove;
- circa 27 empiriche da correggere o eliminare.

L'impatto di questa attività è stato quello di comportare un certo innalzamento del tasso di codifica, come si evince dalla tabella 5, in cui viene riportato un riepilogo con le percentuali di unici e di falliti ottenute prima e dopo gli aggiornamenti dell'ambiente di codifica effettuati a seguito delle analisi dei campioni estratti.

In particolare, questi tassi sono stati ottenuti con due applicazioni di codifica, che chiamiamo:

- ACTR Web (applicazione disponibile sul sito Web dell'Istat);
- ACTR indagini (applicazione utilizzata in *batch* per codificare l'attività economica nelle indagini statistiche).

Come descritto nel paragrafo 1.3, infatti, l'applicazione ACTR utilizzata sul Web e quella in *batch* funzionale alla codifica delle indagini sono state diversificate, in quanto diverse sono le finalità di ciascuna di esse. Per l'applicazione sul web la finalità è quella di proporre un ventaglio di descrizioni di attività economiche (corrispondenti a diversi codici, tutti al massimo dettaglio) simili all'attività digitata dall'utente che consulta il sito; per l'applicazione in batch è quella di massimizzare i codici univoci (anche non al massimo dettaglio) assegnati automaticamente.

Tali varianti, quindi, comportano che il passaggio di codifica di uno stesso file effettuato con le due applicazioni possa produrre tassi lievemente diversi.

Dalla tabella 5 si osserva infatti che:

- con ACTR Web (contesto con soglie ridotte e dizionario con codici a cinque digit, privi delle descrizioni corrispondenti a codici fittizi previsti per l'applicazione ACTR indagini per evitare falsi *match*, 'n.c.'), dopo l'aggiornamento dell'ambiente di codifica, si rileva un incremento degli unici in media dello 0,4%, mentre per i falliti la percentuale si riduce in media dello 0,5% e i codici 'n.c.' risultano in media del 5,5%;
- con ACTR indagini la percentuale degli unici risulta incrementata in media dello 0,7%, mentre la percentuale dei falliti si riduce in media dell'1,3%.

Tabella 5 - Percentuale di unici (tasso di codifica) e di falliti prima e dopo l'aggiornamento dell'ambiente di codifica per cicli di analisi effettuate

DATA	N° record digitato dagli utenti Web	N° rk standardizzato	Tasso di codifica prima delle correzioni del reference e *parsing*						Tasso di codifica dopo le correzioni del reference e *parsing*					
			ACTR Web			ACTR indagini			ACTR Web			ACTR indagini		
			Unici	di cui: n.c.	Falliti	Unici	di cui: unici<5 digit e n.c.	Falliti	Unici	di cui: n.c.	Falliti	Unici	di cui: unici<5 digit e n.c.	Falliti
			%											
09/11/2009	964.959	210.307	45,9	5,5	12,3	54,1	12,0	11,6	46,5	5,7	11,8	55,4	12,7	11,1
14/12/2009	1.052.568	223.406	46,0	5,5	12,3	54,1	12,0	11,6	46,6	5,5	12,0	55,3	12,4	11,1
25/01/2010	1.147.533	239.476	46,3	5,5	12,2	54,8	12,3	11,3	46,6	5,5	12,1	55,3	12,4	11,2
17/05/2010	1.485.821	294.757	46,3	5,5	12,2	54,8	12,1	11,4	46,5	5,5	11,9	55,1	12,1	11,1
22/11/2010	1.931.678	368.222	46,3	5,5	13,1	54,8	12,1	12,3	46,7	5,6	11,9	55,0	12,1	11,1

Per avere un quadro completo dell'impatto delle modifiche apportate, è inoltre stato sottoposto ad ACTR l'insieme delle query digitate dagli utenti Web (fino al 22-11-2010), utilizzando le due applicazioni relative a giugno 2009 (riferita a prima dell'inizio del corso di formazione) e a dicembre 2010 (ultima applicazione disponibile, dopo tutti gli aggiornamenti provenienti dal lavoro dei codificatori e degli esperti).

Come si può osservare dalla tabella 6, l'arricchimento della base informativa effettuato a seguito del monitoraggio ha consentito di incrementare ulteriormente l'efficacia dell'applicazione di codifica e, con essa, la soddisfazione degli utilizzatori esterni.

In termini numerici l'impatto degli interventi effettuati sulla base informativa si può quantificare come percentuali dei casi di attribuzione o non attribuzione del codice ottenibili con l'applicazione antecedente gli aggiornamenti (2009), rispetto a quelli ottenibili con l'ultima applicazione (che include gli aggiornanti del 2010).

Dalla citata tabella emerge che, relativamente all'applicazione ACTR Web:

- nel 2010 i casi in cui l'applicazione riesce ad individuare un codice corrispondente alla descrizione digitata oppure produce un ventaglio di possibilità (unici + multipli + possibili) sono pari all'88,1% delle descrizioni da codificare, a fronte dell'86,9% registrato nel 2009; di questi la percentuale di unici al 2010 è del 46,7%, con un incremento rispetto al 2009 dell'1% pari in termini assoluti a 18.955 descrizioni codificate con codice univoco
- la percentuale di falliti nel 2010 è pari all'11,9% a fronte del 13,1% registrato nel 2009, corrispondente ad un decremento dell'1,2% (pari a 22.742 descrizioni di attività economiche).

Anche relativamente all'applicazione ACTR indagini il tasso di codifica degli unici risulta incrementato dal 2009 al 2010 dell'1,5 % (pari a 28.597 descrizioni di attività economiche).

Tabella 6 - Tasso di codifica sull'insieme delle query con applicazione Web e applicazione indagini prima e dopo aggiornamenti del dizionario e *parsing*

	ACTR Web				*ACTR indagini*			
	Applicazione giugno 2009		Applicazione dicembre 2010		Applicazione giugno 2009		Applicazione dicembre 2010	
	Numero di descrizioni originali	%	Numero di descrizioni originali	%	Numero di descrizioni originali	%	Numero di descrizioni originali	%
Unici	882.878	45,7	901.833	46,7	1.032.689	53,5	1.061.286	55,0
Di cui Unici con codice ateco n.c. e/o <5 digit	*106.140*	*5,5*	*107.264*	*5,6*	*222.903*	*11,5*	*233.229*	*12,1*
Multipli	112.804	5,8	118.791	6,1	43.144	2,2	49.894	2,6
Multipli con 5 codici ateco	2.286	0,1	2.163	0,1	1.355	0,1	1.599	0,1
Possibili	667.287	34,5	665.896	34,5	603.408	31,2	591.740	30,6
Possibili con 5 codici ateco	13.363	0,7	12.677	0,7	12.845	0,7	12.412	0,6
Falliti	253.060	13,1	230.318	11,9	238.237	12,3	214.747	11,1
Totale	**1.931.678**	**100,0**	**1.931.678**	**100,0**	**1.931.678**	**100,0**	**1.931.678**	**100,0**

5.2 I test sui dati censuari

Al fine di misurare le performance dell'applicazione su dati rilevati in indagini statistiche, si è pensato di utilizzare un file di descrizioni alternativo rispetto ai testi digitati su Web. A tal fine sono stati utilizzati i testi del Censimento dell'industria e dei servizi 2001.

L'universo era costituito da 1.130.570 descrizioni che, a seguito del *parsing* ridotto, corrispondevano a 228.738 testi tra di loro diversi. Il file è stato sottoposto a codifica con le due applicazioni (ACTR Web e ACTR_indagini) prima e dopo l'aggiornamento dell'ambiente di codifica ottenendo i risultati riportati nella tabella 7

Tabella 7 - Risultati codifica con applicazione Web e applicazione indagini prima e dopo aggiornamenti del dizionario e *parsing*

	Risultati codifica con applicazione Web				Risultati codifica con applicazione indagini			
	Applicazione giugno 2009		Applicazione dicembre 2010		Applicazione giugno 2009		Applicazione dicembre 2010	
	N° rk	%	N° rk	%	N° rk	%	N° rk	%
Unici	808.766	71,5	808.633	71,5	913.307	80,8	917.857	81,2
Unici 5 digit	*781.498*	*69,1*	*781.488*	*69,1*	*814.795*	*72,1*	*817.242*	*72,3*
Di cui Unici con codice ateco n.c. e/o <5 digit	*27.268*	*2,4*	*27.145*	*2,4*	*98.512*	*8,7*	*100.615*	*8,9*
Multipli	92.945	8,2	92.838	8,2	25.757	2,3	24.871	2,2
Multipli con 5 codici ateco uguali	3.682	0,3	3.587	0,3	535	0,0	507	0,0
Possibili	198.654	17,6	200.586	17,7	166.557	14,7	164.572	14,6
Possibili con 5 codici ateco uguali	8.535	0,8	8.257	0,7	8.342	0,7	8.033	0,7
Falliti	17.988	1,6	16.669	1,5	16.072	1,4	14.730	1,3
Totale	**1.130.570**	**100,0**	**1.130.570**	**100,0**	**1.130.570**	**100,0**	**1.130.570**	**100,0**

Come può vedersi dalla tabella, l'efficacia ottenuta su questo file è indubbiamente più che soddisfacente, tanto più se si analizzano i dati per classe di frequenza. Emerge infatti che mentre gli unici si distribuiscono su tutte le classi, comprese quelle corrispondenti a valori più elevati, non si rilevano falliti corrispondenti alle classi superiori a 301 occorrenze.

Inoltre analizzando gli Unici per punteggio (tabella 8), l'84,1% con ACTR Web e l'80,5% con ACTR_indagini (corrispondenti rispettivamente a 680.145 e 738.690 match) hanno punteggio = 10, ossia sono frutto di un abbinamento diretto con testi presenti nel dizionario; ciò sta a significare che il dizionario è altamente rappresentativo del modo di esprimersi dei rispondenti.

Tabella 8 - Unici per punteggio con applicazione Web e applicazione indagini prima e dopo aggiornamenti del dizionario e *parsing*

UNICI PER PUNTEGGIO	Unici per punteggio ACTR Web				Unici per punteggio ACTR indagini			
	Unici prima dell'aggiornamento dei file di *parsing*		Unici dopo aggiornamento dei file di *parsing*		Unici prima dell'aggiornamento dei file di *parsing*		Unici dopo aggiornamento dei file di *parsing*	
	n° rk	%	n° rk	%	n° rk	%	n° rk	%
Punteggio=10	677.295	83,7	680.145	84,1	733.915	80,4	738.690	80,5
Punteggio=8-9	131.471	16,3	128.488	15,9	179.392	19,6	179.167	19,5
Totale	**808.766**	**100**	**808.633**	**100**	**913.307**	**100**	**917.857**	**100**

Tabella 9 - Risultati della codifica automatica per classi di frequenza e punteggio

CLASSE DI FREQUENZA TESTO	Codifica con applicazione giugno 2009						Codifica con applicazione dicembre 2010					
	Unici				Falliti		Unici				Falliti	
	N. rk con testo diverso tra loro	% Sul totale degli Unici	*Match diretti* (Punteggio = 10)		N. rk con testo diverso tra loro	% Sul totale dei Falliti	N. rk con testo diverso tra loro	% Sul totale degli Unici	*Match diretti* (Punteggio = 10)		N. rk con testo diverso tra loro	% Sul totale dei Falliti
			N. rk con testo diverso tra loro.	% sul totale dei Match Diretti					N. rk con testo diverso tra loro.	% sul totale dei Match Diretti		
1	80.695	67,8	23.498	53,2	9.274	85,9	81.432	67,8	24.058	53,4	8.915	86,5
2-8	30.078	25,3	14.500	32,8	1.388	12,9	30.287	25,2	14.753	32,8	1.291	12,5
9-30	5.307	4,5	3.712	8,4	101	0,9	5.346	4,5	3.761	8,3	79	0,8
31-50	1.025	0,9	796	1,8	22	0,2	1.042	0,9	813	1,8	15	0,1
51-100	858	0,7	699	1,6	5	0,0	864	0,7	706	1,6	7	0,1
101-300	642	0,5	567	1,3	4	0,0	647	0,5	571	1,3	2	0,0
301-500	147	0,1	132	0,3	.		148	0,1	134	0,3	.	
501-1000	147	0,1	137	0,3	.		147	0,1	137	0,3	.	
1001-2000	70	0,1	68	0,2	.		70	0,1	68	0,2	.	
≥ 2001	46	0,0	45	0,1	.		46	0,0	45	0,1	.	
Totale	**119.015**	**100,0**	**44.154**	**100,0**	**10.794**	**100,0**	**120.029**	**100,0**	**45.046**	**100,0**	**10.309**	**100,0**

6. Conclusioni

In sintesi, la funzione di individuazione del codice ATECO sulla base di una descrizione sintetica messa a disposizione degli utenti Web ha registrato un grande successo sia in termini di numero di accessi che di rispondenza alle esigenze degli utenti, che sono riusciti nella stragrande maggioranza dei casi a ritrovare il codice corrispondente alla loro attività.

Quest'ultimo aspetto ci porta a concludere che la base informativa, grazie anche al monitoraggio e al conseguente aggiornamento della stessa, è in linea con il modo di esprimersi degli utenti.

Sulla base di queste considerazioni si sottolinea come questa funzione possa in prospettiva costituire uno standard che l'Istituto potrebbe mettere a disposizioni anche per altre classificazioni. Infatti, l'applicazione si avvale del sistema generalizzato ACTR, già utilizzato in Istat per la codifica di risposte testuali fornite in molteplici indagini e afferenti a diverse classificazioni; già esistono quindi le basi informative per numerose classificazioni che dovrebbero soltanto essere adattate alle esigenze degli utenti Web o, per alcune classificazioni, quali la Professione, aggiornate in funzione dell'ultimo release della classificazione stessa. Anche l'interfaccia Web è da considerarsi ormai un'architettura consolidata e necessiterebbe soltanto di qualche piccola modifica.

Riferimenti bibliografici

C. Colasanti, S. Macchia, P.Vicari. 2009. *The automatic coding of Economic Activities descriptions for Web users*, NTTS 2009 New techniques and technologies for statistics, Bruxelles

Eurostat. 2007. *NACE Rev. 2. Introductory Guidelines*, Statistical governance, quality and evaluation division.

S. Macchia, e al. 2007. *Metodi e software per la codifica automatica e assistita dei dati,* Istat: Tecniche e strumenti n. 4. Roma

P. Vicari e al. 2009. *L'ambiente di codifica automatica dell'Ateco 2007*. Istat: Metodi e Norme n. 41. Roma

M.J. Wenzowski (1988), *ACTR – A Generalised Automated Coding System*. Survey Methodology, vol. 14: 299-308. Statistics Canada

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.