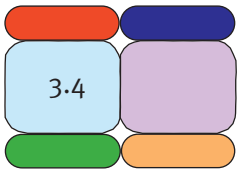
# UN'INTRODUZIONE RAGIONATA AL MONDO DEI WEB SERVICE

**Barbara Pernici**
**Pierluigi Plebani**

Il paradigma del Service Oriented Computing è visto come una rivoluzione nella comunità informatica e i Web Service una sua realizzazione. La possibilità di vedere il Web come un grande sistema informativo in cui sono forniti innumerevoli servizi offre agli utenti finali un potentissimo strumento che va al di là del mero scambio di informazioni che al momento rappresenta il Web. Si propone, qui, una panoramica sui Web Service, le attuali tecnologie e i problemi aperti su cui la ricerca si sta muovendo.




## 1. INTRODUZIONE

L'esplosione di Internet, avvenuta negli anni novanta, è andata di pari passo con il forte interesse che le aziende hanno mostrato verso questo nuovo strumento. Il potenziale pubblico che può attraverso Internet relazionarsi con le aziende è vastissimo e, di conseguenza, anche la possibilità di concludere trattative commerciali aumenta in modo considerevole. Si è assistito, quindi, a una integrazione di sistemi informativi sia orizzontalmente sia verticalmente che ha portato alla proliferazione di termini quali *e-commerce*, *e-business*, *e-government*, *e-procurement*. In realtà, già negli anni ottanta, dopo aver visto una forte diffusione dei sistemi informativi all'interno delle aziende, si è sentita la necessità di poter integrare piattaforme differenti. Questo tipo di esigenza sorge non solo tra aziende ma addirittura all'interno della medesima azienda in cui, per motivi storici e organizzativi, diverse divisioni hanno operato scelte tecnologicamente differenti nella realizzazione del proprio sistema informativo.

Ora, complice Internet, le aziende perseguono il medesimo obiettivo in modo da poter automatizzare le filiere produttive o creare dei *marketplace* virtuali. Tale integrazione ha richiesto, e tuttora richiede, notevoli sforzi sia di tipo organizzativo che di tipo tecnologico in grado di mediare tra la necessità delle aziende in questione di scambiarsi informazioni e la necessità opposta di mantenere una certa autonomia.
Dal punto di vista organizzativo è plausibile immaginare che ogni azienda possa fornire servizi e al contempo utilizzarne. In particolare, i servizi forniti, che rappresentano le funzionalità che l'azienda intende esportare verso l'esterno, delineano il confine tra ciò che del proprio sistema informativo aziendale è pubblico e ciò che al contrario rimane privato. D'altro canto, dal punto di vista tecnologico affinché questa interoperabilità sia fattibile, le aziende devono accordarsi su un linguaggio comune di descrizione dei servizi in modo tale da riconoscere cosa un sistema mette a disposizione. Ciò deve essere accompagnato anche da un meccanismo di ricerca dei servi-

zi esistenti e dalla possibilità di utilizzare il Web come canale di trasmissione.
Sebbene gli aspetti organizzativi siano molto importanti per capire le politiche che sottendono alla decisione su quando una funzionalità possa essere esportata e come questo impatta sul *core-business* aziendale, nel presente articolo ci si concentrerà unicamente sugli aspetti tecnologici soffermandoci, quindi, sui Web Service quale soluzione tecnologica adatta all'interoperabilità dei sistemi. Va altresì sottolineato che il ruolo dei Web Service non si limita solo a un discorso di interoperabilità, bensì permette di descrivere nuovi servizi realizzati *ad hoc*, sempre però con l'intento di fornire una soluzione *platform-independent*. Il primo passo è, quindi, quello di separare nettamente, come del resto avviene anche in modelli di programmazione basati su componenti, la logica di presentazione da quella che è la logica applicativa. I Web Service, infatti, si occupano solamente della logica applicativa, e sarà, quindi, il fruitore del servizio a presentare i dati ottenuti utilizzando il servizio con stili e grafica propri.
Si immagini, ad esempio, che l'azienda ACME voglia, all'interno della propria pagina Web, inserire una piccola casella in cui mostrare il valore della quotazione in tempo reale delle proprie azioni. L'ACME per avere questo tipo di informazione dovrà necessariamente richiedere i dati periodicamente al sistema informativo della Borsa utilizzando un servizio appositamente fornito da quest'ultima. Tale servizio richiederà, quindi, in *input* il simbolo dell'azione e restituirà, in *output*, la quotazione attuale. Il Web master di ACME non dovrà fare altro che invocare il servizio e far sì che una porzione della pagina Web dell'azienda sia alimentata dai dati ottenuti in risposta.
Prima però di discutere dei dettagli prettamente tecnici, e per inquadrare meglio lo spazio d'azione, si vuole dare una definizione di più ampio respiro di Web Service per non correre il rischio, come spesso avviene, di limitare la descrizione di questa tecnologia alla descrizione delle tre tecnologie di base quali *Web Services Description Language* (WSDL) [1], *Simple Object Access Protocol* (SOAP) o [7] e *Universal Description, Discovery and Integration* (UDDI) o [11, 12] che comunque verranno affrontate in seguito. A tal proposito, si partirà da una architettura di riferimento, la *Service Oriented Architecture* (SOA) [3, 4], grazie alla quale è possibile identificare gli scopi, gli utilizzi e gli sviluppi futuri che caratterizzano i Web Service e si potrà notare come l'architettura si inserisce perfettamente all'interno delle considerazioni fatte in precedenza, in particolare di quelle relative al mondo *business-to-business* (B2B).
Focalizzando l'attenzione sul concetto di servizio è ovvio immaginare, anche alla luce di quanto detto finora, come gli attori in causa siano necessariamente il fornitore e il richiedente. Questo tipo di paradigma è il medesimo che si riscontra nella tipica interazione di tipo *client-server*. Attraverso la SOA questa interazione viene arricchita con un ulteriore attore detto *Service Directory o Service Broker* che, come mostrato in figura 1, si inserisce all'interno della comunicazione tra fornitore e fruitore del servizio. Di seguito, viene riportata una descrizione approfondita del ruolo di ognuno dei tre attori coinvolti nella SOA:

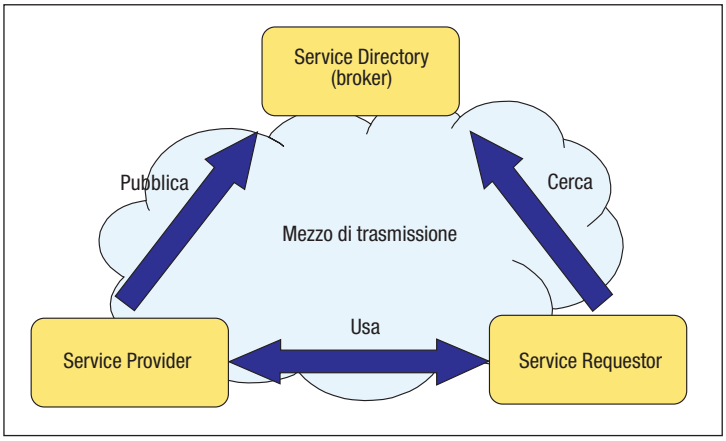



**FIGURA 1**
*Service Oriented Architecture*

### *Service Provider*

Chi realizza e mette a disposizione un servizio. Tramite l'operazione di *publish* il servizio viene "pubblicizzato", in quanto le caratteristiche del servizio realizzato vengono memorizzate all'interno di un *registry* accessibile pubblicamente. Il Service Provider rimane, quindi, in attesa che un utente richieda tale servizio.

### *Service Directory o Service Broker*

Questo componente si occupa della gestione del registry, permettendo, a chi ha necessità, di ricercare un servizio sulla base delle caratteristiche con le quali è stato definito e memorizzato. Naturalmente, il Service Directory può seguire politiche di controllo degli accessi sulle interrogazioni in modo da limitare la visibilità sui servizi inseriti. Nel presente lavoro il registry, salvo diversa indicazione, viene considerato totalmente accessibile.

### *Service Requestor*

Rappresenta un potenziale utente che richiede un servizio. A tale scopo, tramite la primitiva di *find* l'utente interagisce con il Service Directory per ottenere il servizio più adatto ai propri obiettivi. Una volta individuato si collega al Service Provider corrispondente (*bind*) e inizia a fruire del particolare servizio (*use*).

La SOA definisce, quindi, "*chi fa che cosa*" all'interno di una serie di interazioni in cui il servizio ricopre il ruolo principale. Va notato che i tre attori interessati possono essere distribuiti sul territorio e possono utilizzare piattaforme tecnologiche differenti, con l'unico vincolo però di dover utilizzare tutti e tre un canale trasmissivo comune. Rimanendo sul mezzo trasmissivo, questo risulta essere un parametro dell'architettura, quindi l'approccio adottato dalle SOA ha il vantaggio di potersi integrare con diversi ambienti quali la telefonia mobile, il Web, o paradossalmente anche la posta ordinaria, permettendo in tal modo di realizzare applicazioni multi-canale, fruibili cioè attraverso diversi dispositivi.

Partendo da questa considerazione si può dire che una architettura per *e-Service* è un'istanza di una SOA dove il mezzo di comunicazione è di tipo elettronico, mentre una architettura per Web Service è un'istanza di una SOA dove il mezzo di comunicazione considerato è il Web. In figura 2 viene individuato il ruolo di SOAP, WSDL e UDDI, ovvero delle tre tecnologie fondamentali.

In particolare, SOAP è un protocollo, basato su XML (*eXtensible Markup Language*) e HTTP (*Hyper Text Transfer Protocol*), in grado di fare interagire componenti remoti attraverso il Web. Nonostante uno dei primi scopi di SOAP sia stato quello di supportare l'RPC (*Remote Procedure Call*) sul Web, questo protocollo è stato studiato anche per avere usi che possono comprendere una interazione di tipo asincrono e, quindi, basata su messaggi. In SOAP la specifica delle chiamate viene descritta in XML, mentre l'HTTP è il protocollo di trasporto su cui poggia. Questa ultima caratteristica pone SOAP in una posizione privilegiata rispetto ad altri meccanismi di invocazione presenti in standard di computazione distribuita quale COM+, Java RMI e CORBA, in quanto questi ultimi mostrano dei limiti nell'uso del Web durante la comunicazione visto che i loro messaggi vengono spesso bloccati dai *firewall*.

## 2. WSDL (WEB SERVICE DEFINITION LANGUAGE)

Si supponga che una banca voglia fornire un servizio di approvazione automatica prestiti di piccola entità che permette ai correntisti della banca di chiedere un prestito direttamente *on-line*. Dando per assunto che l'utente si sia precedentemente registrato, questi non dovrà far altro che indicare il proprio identificativo che corrisponderà al proprio numero di conto corrente e la somma di denaro richiesta.

Già da questa breve descrizione è possibile identificare come il servizio metta a disposizione dell'utente un'operazione (*approva*, per esempio) che richiederà dei dati in ingresso e in uscita. Attraverso WSDL è possibile formalizzare tutte queste caratteristi-

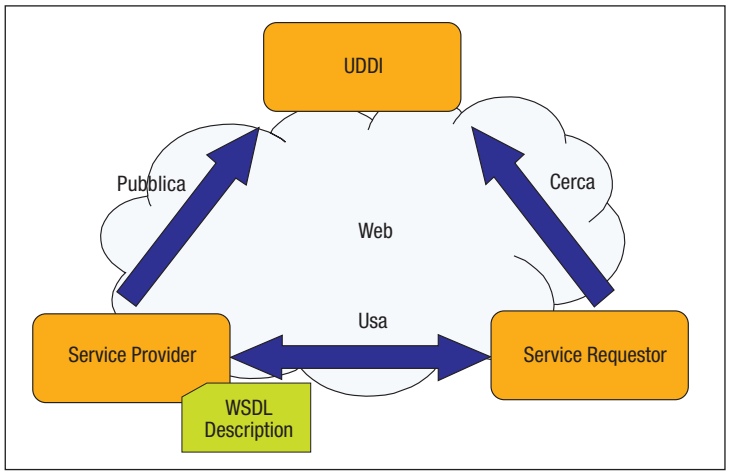



**FIGURA 2**
*Web Service nella SOA*

```
<definitions targetNameSpace="http://tempuri.org/services/approvaPrestiti"
             xmlns:tns="http://tempuri.org/services/approvaPrestiti"
             xmlns:xsd="http://www.w3.org/2001/XMLSchema"
             xmlns="http://schemas.xmlsoap.org/wsdl/">
  <types>
    ...
  </types>
  <message name="RichiedenteMsg">
    <part name="cc" type="xsd:integer" />
    <part name="ammontareRichiesto" type="xsd:integer" />
  </message>
  <message name="erroreMsg">
    <part name="errorCode" type="xs:integer" />
  </message>
  <message name="approvazioneMsg">
    <part name="risposta" type="xs:string" />
  </message>
  <portType name="approvaPrestitoPT">
    <operation name="approva">
      <input message="tns:richiedenteMsg" />
      <output message="tns:approvazioneMsg" />
      <fault name="errore" message="tns:erroreMsg" />
    </operation>
  </portType>
  <binding name="approvaPrestitoSOAPBinding" type="tns:approvaPrestitoPT">
    <soap:binding style="document"
                  transport="http://schemas.xmlsoap.org/soap/http"/>
    <operation name="approva">
      <soap:operation soapAction="http://tempuri.org/services/approva"/>
      <input>
        <soap:body use="literal"/>
      </input>
      <output>
        <soap:body use="literal"/>
      </output>
    </operation>
  </binding>
  <service name="approvaPrestitoService">
    <documentation>
        Servizio di approvazione prestiti automatico
    </documentation>
    <port name="approvaPrestito" binding="approvaPrestitoSOAPBinding">
      <soap:address location="http://tempuri.org/services/approva"/>
    </port>
  </service>
</definitions>
```

**TABELLA 1**
*Documento WSDL per il servizio di approvazione prestito*

che del servizio secondo uno schema che non differisce molto dalla specifica delle API (*Application Program Interface*) di un sistema. Come si può notare nella tabella 1, WSDL è un linguaggio basato su XML e permette di specificare quelle caratteristiche del servizio che prima sono state descritte solo a parole.

Si comincia con l'identificare le componenti fondamentali di un *file* WSDL partendo da *service* che identifica un insieme logico di servizi (un insieme di servizi e non solo uno!) ognuno dei quali è specificato da una *port*. È fondamentale ricordare che una port identifica l'indirizzo a cui il servizio risponde e il tipo di protocollo supportato dal servizio stesso e non le caratteristiche del servizio. Per questo motivo a ogni port è associata una *portType* il cui compito è quello di specificare la reale sintassi del servizio. La portType ha, quindi, il ruolo di dire cosa il servizio fa, mentre la port come il servizio possa essere acceduto. Questo tipo di specializzazione viene specificato attraverso il *binding* che data una portType descrive in che modo questa si inserisce all'interno di un particolare protocollo quale SOAP, HTTP o SMTP. Riferendosi, quindi, all'esempio, viene specificato un solo servizio (*ApprovaPrestito*) così come appare nel *tag port*. Questo servizio è accessibile all'indirizzo *http://tempuri.org/services/approva* e per invocarlo è necessario che il *client* inoltri le proprie richieste secondo il protocollo SOAP.
Nel dettaglio, una portType è composta da una serie di *operation* che rispecchiano le funzionalità fornite dal servizio e che interagiscono con l'utente e che possono basarsi su uno dei quattro pattern predefiniti:

- *One_way*: l'operazione è composta da un solo messaggio in ingresso al fornitore del servizio.
- *Request_response*: l'operazione prevede una risposta del fornitore del servizio successiva a un messaggio ricevuto dall'utente.
- *Solicit_Response*: l'operazione prevede l'attesa da parte del fornitore del servizio, di una risposta a fronte di una richiesta effettuata dal fornitore stesso.
- *Notification*: l'operazione è composta da un solo messaggio in uscita al fornitore del servizio.

Indipendentemente dal pattern, ogni interazione cliente-fornitore del servizio avviene attraverso uno scambio di messaggi opportunamente identificati in WSDL attraverso il *tag message*, grazie al quale è possibile specificare il formato del messaggio. In particolare, ogni messaggio è visto come un insieme

di una o più port ognuna delle quali aderente a un tipo di dato che può essere uno di quelli primitivi di XML (per esempio, *int* e *string*) oppure un tipo di dato definito in *types*; quest'ultimo non utilizzato all'interno dell'esempio visto che tutti i dati sono definiti secondo un tipo di base.
Al di là delle specifiche sintattiche di WSDL, soffermandosi ulteriormente sulla relazione che esiste tra port e portType è possibile dire che una port può essere vista come una specializzazione di una portType operata secondo un particolare protocollo di comunicazione. È quindi possibile che all'interno del medesimo file WSDL la stessa portType, quindi il servizio, possa essere reso accessibile, grazie ai *binding*, su diversi protocolli di trasporto. Al momento accanto alla specifica di WSDL sono stati specificati i binding su SOAP, HTTP e SMTP (*Simple Mail Transfer Protocol*) quindi è possibile solo specificare Web Service che comunicano attraverso questi protocolli. Sulla base di ciò un buon modo di operare prevede che una specifica WSDL di un servizio sia composta da almeno due file: il primo chiamato WSDL *Interface document* composto solo dalla specifica dei tipi, dei messaggi e delle portType, quindi per ogni tipo di binding verrà definito un documento apposito detto anche WSDL *implementation document*.
Sebbene risulti molto generico, WSDL è al tempo stesso molto esauriente e in grado di descrivere servizi di diverse tipologie. L'utilizzo dei quattro pattern di comunicazione permette, infatti, di specificare sia servizi asincroni che sincroni. Nel primo caso saranno utilizzati solo operazioni di tipo *one-way* e *notification*, mentre nel secondo caso di tipo *request-response* e *solicit-response*.
WSDL però propone solo una fotografia del servizio dandone quindi una visione statica. Si supponga, infatti, di aggiungere al WSDL l'operazione di *login*, come mostrato in tabella 2, in cui l'utente inserisce i propri *userID* e *password* che gli dà diritto di invocare l'operazione di richiesta prestito. In questo caso, solo una certa esperienza suggerisce che le operazioni previste devono essere invocate secondo un ordine preciso e a precise condizioni. L'utente, infatti, deve prima fare il login e solo dopo l'autenticazione richiedere un prestito. Ciò che manca, quindi, è la dinamica del servizio chiamato anche *comportamento*. Attraverso il comportamento è possibile sapere come un servizio funziona e quali sono le operazioni ammissibili in accordo con il suo stato interno. Per questo tipo di problema sono stati sviluppati diversi linguaggi. WSCL (*Web Service Conversation Language*) per esempio descrive, secondo una specifica XML, come un potenziale utente può conversare con il servizio descrivendolo come una macchina a stati finiti dove le operazioni rappresentano gli stati e, in aggiunta, vengono specificate le transizioni tra di essi. Un progetto più ambizioso, BPEL4WS (*Business Process Execution Language for Web Service*), definisce un linguaggio di composizione tra servizi che può anche essere utilizzato come specifica del comportamento di un singolo servizio.

```
...
<message name="loginReqMsg">
    <part name="userId" type="xsd:string" />
    <part name="password" type="xsd:string" />
</message>
<message name="loginResMsg">
    <part name="authRes" type="xsd:string" />
</message>
<portType name="approvaPrestitoPT">
    <operation name="login">
        <input message="tns:loginReqMsg" />
        <output message="tns:loginResMsg" />
        <fault name="errore" message="tns:erroreMsg" />
    </operation>
    <operation name="approva">
        ...
    </operation>
</portType>
<binding name="approvaPrestitoSOAPBinding"type="tns:approvaPrestitoPT">
    <soap:binding style="document"
                transport="http://schemas.xmlsoap.org/soap/http"/>
    <operation name="login">
        <soap:operation soapAction="http://tempuri.org/services/login"/>
        <input>
            <soap:body use="literal"/>
        </input>
        <output>
            <soap:body use="literal"/>
        </output>
    </operation>
    ...
</binding>
...
```

**TABELLA 2**
*Operazioni di autenticazione per il servizio di approvazione prestito*

## 3. BPEL4WS (BUSINESS PROCESS EXECUTION LANGUAGE FOR WEB SERVICES)

L'esigenza di coordinare in modo automatico attività svolte da attori diversi per il raggiungimento di un obiettivo comune quale, per esempio, l'approvazione di una richiesta, la concessione di un prestito e il rimborso di danni a un assicurato è caratteristica di processi con un elevato numero di istanze, regole di esecuzione precise e ripetibili. Le classiche applicazioni *software* per il supporto di flussi di lavoro con queste caratteristiche sono i sistemi di gestione di *workflow* (WfMS, *WorkFlow Management Systems* o), che consentono la gestione contemporanea di un numero, anche elevato, di istanze di processi seguendo schemi di processo predefiniti.

In questo ambito, il processo viene definito come una rete di attività e di relazioni esistenti tra di esse, criteri per indicare l'inizio e la fine di un processo, e informazioni riguardo alle singole attività, quali: i partecipanti, le applicazioni IT (*Information Technology*) associate, i dati, e così via. La rappresentazione di un processo in una forma che consente l'esecuzione automatica delle attività automatizzabili e la gestione automatica del passaggio tra un'attività e quelle che seguono nel flusso di lavoro è basata su alcuni costrutti fondamentali quali:

- *sequenza*: le attività vengono svolte l'una di seguito all'altra: una attività viene attivata solo al termine di un'attività precedente;
- *parallelo* (*AND split*): dopo la terminazione di una attività vengono attivate più attività in parallelo;
- *alternativa* (*OR split*): dopo la terminazione di una attività vengono attivate più attività in alternativa; vengono specificate le condizioni di attivazione delle attività oppure può essere effettuata una scelta non deterministica;
- *join* (*AND oppure OR*): per proseguire nel flusso di lavoro devono essere terminate tutte le attività precedenti (caso AND), o almeno una delle attività precedenti.

Sulla base di questi costrutti è possibile rappresentare reti di attività di tipo generale. Nei sistemi di gestione di workflow le attività vengono schedulate, attivate e gestite sulla base delle risorse disponibili e grazie a ciò è possibile ottenere l'automazione completa o parziale di un *business process* in cui documenti, informazioni o compiti sono passati da un partecipante a un altro per svolgere attività, secondo un insieme di regole procedurali. I prodotti sul mercato consentono, in genere, di coordinare processi all'interno di una sola organizzazione, seguendo una logica essenzialmente centralizzata, anche se le singole attività possono essere svolte dagli operatori su postazioni di lavoro diverse.

L'affermarsi dei Web Service rende naturale l'estensione dei concetti alla base dei sistemi di gestione di workflow anche per coordinare servizi in rete forniti da più organizzazioni. In questo ambito, l'obiettivo principale è quello di comporre più servizi forniti da fornitori diversi al fine di creare nuovi servizi a valore aggiunto. Tale composizione richiede però la definizione di standard per modellare le interazioni tra i servizi, e a tali standard stanno lavorando numerosi *vendor* e ricercatori.

Al momento, la letteratura propone due principali approcci al coordinamento dei servizi in rete, che vengono designati sotto le denominazioni di "orchestrazione" e "coreografia". Nell'orchestrazione, si fa riferimento a un processo di business che può interagire con altri servizi, interni o esterni. L'interazione avviene tramite messaggi scambiati tra i servizi. L'orchestrazione però presuppone il controllo sul processo da parte di una sola organizzazione e consente di gestire processi in esecuzione anche di lunga durata.

La coreografia indica, invece, un approccio più collaborativo, in cui ogni partecipante descrive il proprio ruolo nell'interazione e il compito del sistema di coreografia è principalmente quello di tenere traccia delle interazioni avvenute tra le parti.

BPEL4WS (o BPEL, in breve) è un linguaggio, nato grazie all'esperienza maturata sulla base di XLANG, una proposta di Microsoft, e WSFL (*Web Service Flow Language*), si colloca come la principale proposta in letteratura per l'orchestrazione di Web Service, che può essere vista come una na-

turale evoluzione dei modelli e sistemi per la gestione di workflow. La versione 1.1 di BPEL è stata progettata da Microsoft, IBM, Siebel Systems, BEA e SAP e rilasciata nel maggio 2003.
Essendo l'obiettivo di BPEL o quello di specificare un modello di comportamento dei servizi Web durante un processo di business, questo linguaggio si pone a un livello più alto rispetto ai linguaggi esaminati fin qui. Per questo BPEL presuppone che i servizi che dovrà orchestrare siano descritti secondo WSDL ed eventualmente memorizzati in un registry UDDI.
Dal punto di vista prettamente tecnico, BPEL esprime la logica di funzionamento del processo attraverso una grammatica basata su XML interpretabile da un *orchestration engine* opportunamente progettato per supportare questo linguaggio e il cui compito sarà quello di coordinare i vari servizi che compongono il processo.
Una delle particolarità di BPEL è la sua visione ricorsiva di composizione di Web Service. Infatti, Il processo definito come insieme di Web Service collaboranti, può essere esso stesso un Web Service. Questo nuovo servizio a valore aggiunto, così creato, sarà visibile all'esterno come un servizio semplice di cui è definita l'interfaccia WSDL. Ciò che avviene è, quindi, una divisione del processo secondo due punti di vista (Figura 3): quello dell'utente del processo, a sinistra, che può supporre di interagire con un singolo servizio, e quello dei servizi cooperanti all'interno del processo, a destra, che, descrivendo le proprie funzionalità attraverso WSDL, verranno invocati dall'*orchestration engine* nei tempi e nei modi definiti all'interno del processo.
Entrando più in dettaglio, il processo essenzialmente è composto da attività che possono gestire richieste da parte del client del processo (attività di *receive*) oppure inviare messaggi di risposta al client medesimo (attività di *reply*). Per poter soddisfare tale richiesta il processo si basa, come si è detto, su una serie di Web Service esterni chiamati *partner*, invocati attraverso l'attività di *invoke* che potrà corrispondere a un servizio sincrono o asincrono che dipende dalle caratteristiche del Web Service chiamato. Come è facile intuire, il paradigma sottostante basato sui messaggi ben si coniuga con i quattro modelli di interazione (*one-way, notification, request-response, solicit-response*) definiti in WSDL.

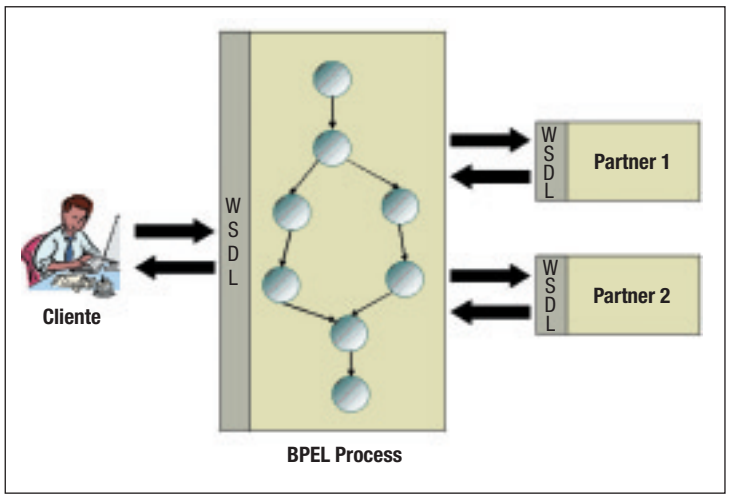



**FIGURA 3**
*Modalità di interazione con un processo BPEL*

I tre tipi attività descritti rappresentano le attività fondamentali che il processo deve svolgere per interagire con l'esterno. Tali attività dovranno essere svolte secondo uno schema ben preciso che rappresenta appunto il processo vero e proprio. A tal proposito, possono essere utilizzati i principali costrutti, precedentemente descritti, per i modelli di workflow e cui corrispondono i costrutti *sequence* per la sequenza, *flow* per l'esecuzione di attività in parallelo, *switch* e *pick* per le alternative e *while* come costrutto iterativo.
A supporto del mantenimento dello stato tra le diverse chiamate, il processo può, inoltre, contemplare la definizione di variabili che consentono di memorizzare i dati corrispondenti alle richieste ricevute o alle risposte di invocazioni di servizi. È anche possibile effettuare manipolazioni di tali variabili all'interno del processo, attraverso l'attività di *assign*, il che rende il linguaggio di descrizione del processo simile a un linguaggio di programmazione.
Partendo dall'esempio sul servizio di prestito illustrato per descrivere WSDL è possibile notare come questo può essere utilizzato all'interno di un processo BPEL. In prima battuta si consideri l'esempio di figura 4 dove si ipo-

tizza che una banca abbia definito un proprio processo per la gestione delle richieste di prestito [14].
Tale processo, definito in BPEL, delega l'approvazione del prestito a un servizio esterno che corrisponde all'*ApprovaPrestito* definito precedentemente in WSDL. In questo caso, il processo può essere visto come un semplice *mediator* che, ricevuta una richiesta la inoltra al servizio di competenza e, una volta ricevuta una risposta da quest'ultimo la inoltra al cliente. Il servizio visto in precedenza in poche parole potrebbe essere visto come un servizio che ora non è più invocabile direttamente ma che, supponendo sia fornito da un consulente finanziario, può essere utilizzato solo da istituti bancari.
Un esempio più complesso è mostrato in figura 5.
In questo caso, il processo risulta maggiormente ricco e si appoggia su due diversi servizi. Va notato che dal punto di vista del cliente, il processo precedente e quello attuale sono assolutamente identici. Ciò che cambia è come la banca gestisce le varie richieste di prestito. In questo caso infatti, sfruttando i costrutti tipici dei workflow, la banca è in grado di definire un processo BPEL in cui, dopo aver ricevuto la richiesta di prestito, seleziona uno dei due servizi di supporto sulla base dell'istanza di richiesta stessa. Nell'esempio, se l'ammontare del prestito richiesto è inferiore a 10.000 euro, la richiesta viene valutata da un "servizio di valutazione richiedenti", mentre il ricorso all'istituzione finanziaria avviene solo nei casi in cui il prestito sia di importo maggiore o il richiedente sia considerato a rischio. La risposta viene inviata, come nell'esempio precedente, al cliente, che in seguito potrà inviare la richiesta di attivazione del prestito. Quest'ultimo messaggio consente di esaminare una caratteristica di BPEL assente in generale nei sistemi di workflow: la seconda, *receive*, è anch'essa la ricezione di un messaggio, nel quale non viene effettuato un riferimento esplicito all'istanza attiva del processo. Il concetto di *correlationSet* di BPEL consente di definire una sorta di chiave per il processo, che permette di associare i messaggi alle istanze di processo attive. In questo esempio, si può fare l'ipotesi che venga gestita una richiesta di prestito alla volta da parte di uno stesso cliente e che, quindi, la chiave sia il nominativo del cliente. Pertanto, finché la richiesta di prestito precedente rimane attiva, ulteriori messaggi da parte dello stesso cliente verranno interpretati come continuazione dell'interazione nell'ambito del processo attivo. Per cui il cliente potrà inviare la richiesta di attivazione (o la rinuncia) senza avere la necessità di conoscere dettagli sulla modalità di gestione della sua pratica. Il meccanismo sopra descritto è particolarmente interessante poiché consente di gestire anche

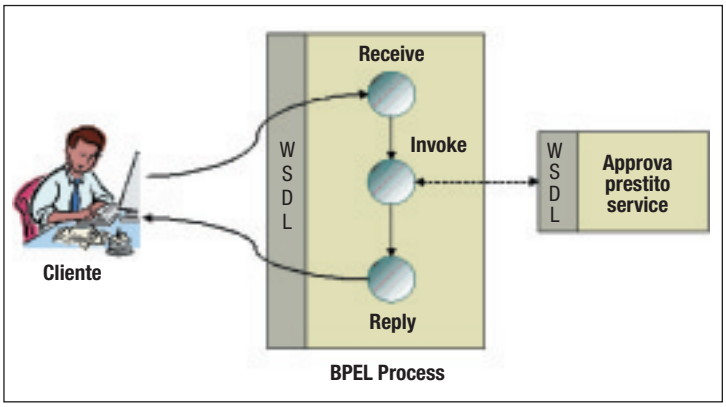



**FIGURA 4**
*Esempio di interazione semplice con un processo BPEL*

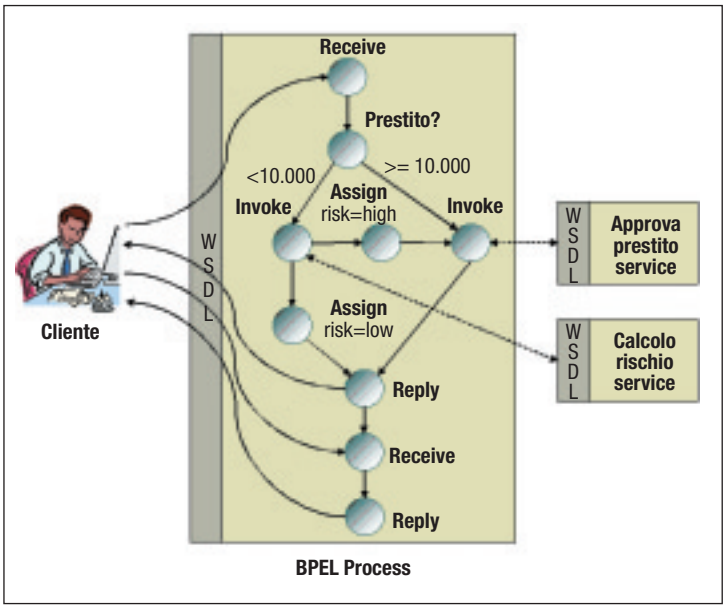



**FIGURA 5**
*Processo BPEL con costrutti decisionali*

processi e transazioni con lunga durata nel tempo.
Tra le funzionalità più avanzate, BPEL fornisce anche meccanismi per la gestione di transazioni ed eccezioni, basandosi sulle specifiche WS-coordination and WS-Transaction, sviluppate da IBM, Microsoft e BEA. Le eccezioni vengono gestite analogamente al linguaggio Java tramite costrutti di *throw* e *catch* identificando, quindi, anche degli *scope* all'interno dei quali la gestione delle eccezioni ha validità. Le eccezioni possono essere sollevate al verificarsi di particolari eventi, anche temporali. Poiché in un processo a lunga durata le eccezioni possono portare a interrompere transazioni in esecuzione, BPEL a supporto in particolare delle *long-transaction*, permette di specificare anche attività mirate alla compensazione e al *failure-recovery*.

## 4. UDDI (UNIVERSAL DESCRIPTION DISCOVERY AND INTEGRATION)

Dopo aver visto come un servizio può essere descritto secondo la visione chiamata statica e dinamica, va sottolineato come la probabilità di successo nella realizzazione di sistemi basati su Web Service dipende fortemente dalla facilità di reperimento dei servizi stessi. È di fondamentale importanza quindi, avere a disposizione una piattaforma in grado di reperire Web Service sulla base di diverse tipologie di ricerca.
UDDI è un progetto nato per questo scopo e iniziato da un gruppo di compagnie del mondo IT quali Microsoft, IBM e Ariba e al quale ora aderiscono circa 300 aziende.
Per meglio comprendere il ruolo di UDDI all'interno dell'ambiente tecnologico di interesse, occorre definire, in modo preciso, il dominio applicativo all'interno del quale si colloca. In prima analisi, questo scenario di riferimento può essere scomposto in tre fasi operative.
**1.** Un'azienda, o un gruppo di aziende, descrive le caratteristiche che una certa tipologia di servizio deve possedere. A questo livello ci si riferisce alla *tipologia di un servizio* come per esempio vendita, acquisto o noleggio e non a un servizio preciso quale può essere la vendita di software o l'acquisto di fiori.
**2.** Una qualunque azienda può, a questo punto, realizzare una delle tipologie di servizio definite, fornendo in tal modo un *servizio*.
**3.** L'insieme dei servizi, così come quello delle tipologie di servizio, deve essere consultabile. Questo permette, a chi vuole fruirne, di reperire il servizio desiderato, e agli sviluppatori di trovare le specifiche di una tipologia di servizio.
Partendo da questo scenario, UDDI affronta le problematiche di pubblicazione e reperimento di servizi, proponendo una architettura che permette di affrontare le tre problematiche riassunte nell'acronimo UDDI, e cioè:
- l'accesso alla descrizione di servizi, di tipologie di servizi e di fornitori di servizi secondo una struttura dati ben definita;
- l'astrazione dalla tecnologia utilizzata nella realizzazione del servizio. Questo al fine di permettere l'integrazione tra servizi realizzati in modo tecnologicamente differente;
- la ricerca di un servizio secondo differenti chiavi di ricerca.

UDDI, quindi, non è un contenitore di servizi o di descrizioni di servizi, bensì uno strumento che, utilizzando opportune strutture dati, tiene traccia della dislocazione dei servizi e delle loro descrizioni. Operando, inoltre, una classificazione sulle informazioni raccolte, UDDI permette l'esecuzione di ricerche efficaci ed efficienti.
Riguardo alla sua architettura UDDI prevede la creazione di un ambiente distribuito *peer-to-peer* in cui i vari *nodi*, che contengono una parte dei servizi disponibili, possano interoperare tra di loro allo scopo di soddisfare le richieste di pubblicazione e ricerca di servizi.
Per garantire l'interoperabilità, ogni nodo dell'infrastruttura, gestito da una figura che assume il ruolo di *operatore*, deve essere realizzato secondo le specifiche o rilasciate dal gruppo di lavoro UDDI. Tali specifiche definiscono sia la struttura delle informazioni che un registry UDDI deve memorizzare, sia il set minimo di API da implementare per l'accesso alle informazioni stesse.
Sulla base di queste specifiche ogni nodo può essere rappresentato secondo quanto schematizzato in figura 6, dove si possono

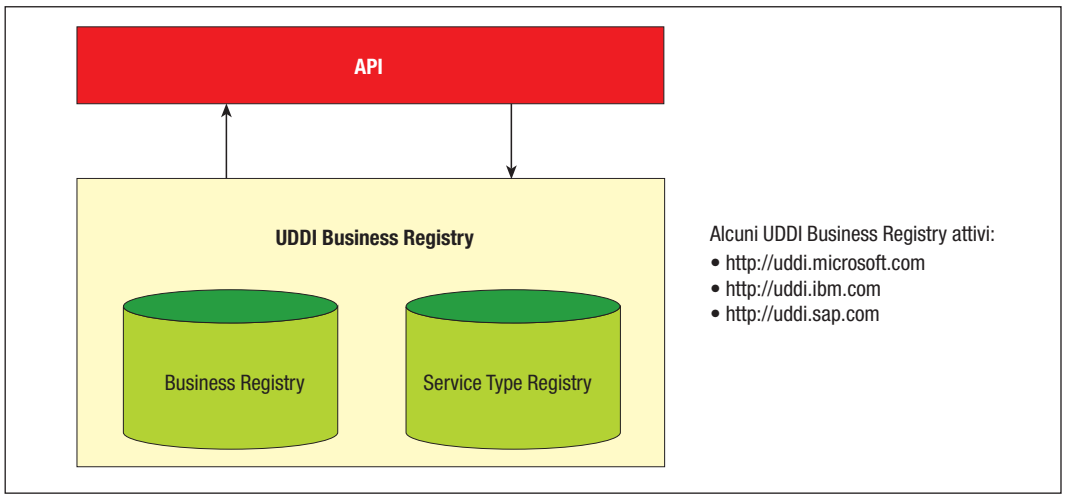



**FIGURA 6**
*Struttura di un UDDI Registry*

individuare le due componenti principali:

- *UDDI Business Registry* ulteriormente suddiviso in:
  - *Service Type Registry*: contenente un insieme di dati in grado di descrivere e localizzare le tipologie di servizio. Tali informazioni sono organizzate secondo una struttura detta *tModel*;
  - *Business Registry*: contenente informazioni sulle aziende che forniscono servizi e si sono registrate al nodo. Ogni entry di questo registry è detta *businessEntity*;
- libreria API per l'accesso, la ricerca e la manutenzione delle registrazioni.

Riprendendo, quindi, le tre fasi operative descritte nell'introduzione che caratterizzano il dominio applicativo di interesse, si può affermare che UDDI permette, grazie ai *tModel* memorizzati nel *Service Type Registry*, di ottenere le descrizioni di una tipologia di servizio, senza che esista alcun vincolo sul linguaggio di specifica utilizzato per redigere la descrizione stessa. Una volta realizzato un servizio, la sua descrizione può essere pubblicata su un nodo UDDI all'interno del *Business Registry*. Quindi UDDI, tramite il gruppo di API, mette a disposizione un meccanismo per la ricerca di servizi o tipologia di servizi secondo diverse chiavi di ricerca. In particolare, sono messe a disposizione tre tipologie di ricerca secondo una classificazione a *White page*, a *Yellow page* e a *Green page*. Nel primo caso, la metodologia di ricerca possiede una forte analogia con quella offerta dagli elenchi telefonici è possibile, quindi, conoscendo il nome del fornitore avere informazioni sui servizi forniti. Nel secondo caso, l'analogia si sposta verso le pagine gialle aziendali, la ricerca, quindi, avviene in prima istanza utilizzando una categorizzazione delle tipologie di servizio. Quello che invece è caratteristico di UDDI è il terzo tipo di classificazione che, basandosi sulle informazioni memorizzate nei tModel, permette di conoscere i fornitori di un servizio date le caratteristiche di alcune funzionalità. Inoltre, sfruttando le tassonomie e le classificazioni inserite nella *businessEntity*, è possibile organizzare l'elenco dei fornitori di servizio, in modo da permettere ricerche basate, per esempio, su l'ambito in cui si colloca il servizio, la localizzazione geografica dei fornitori oppure l'ambito in cui si colloca il fornitore del servizio.

In questa descrizione di UDDI è emerso come l'obiettivo dell'UDDI Registry sia duplice: rendere pubbliche le descrizioni sia di tipologie di servizio, sia di loro specializzazioni. Ciò si relaziona molto bene con la suddivisione della specifica WSDL esposta in precedenza in cui si suggeriva, per ogni servizio, di descriverlo secondo due documenti: il WSDL *Service Interface Document*, contenente solo i tag *type*, *message*, *portType* e *type* e il WSDL *Service Interface Document* che include il precedente e in più definisce i tag di binding e di service. Fatta questa suddivisione, alla luce delle caratteristiche di UDDI, è utile considerare il WSDL *Service Interface Document* come la descrizione della tipologia del servizio, mentre il WSDL *Service Implementation Document* come la descrizione del servizio vero e proprio.

Sempre in figura 6 sono elencati alcuni dei nodi UDDI al momento attivi. Data la natura ancora sperimentale, attualmente questi registri sono per la maggior parte popolati da servizi non funzionanti o non attivi. Questo sottolinea come i Web Service non siano ancora maturi per un utilizzo massiccio e gioca anche a sfavore di UDDI il fatto che da parte di alcuni venga considerato come un mezzo insufficiente e superato. Ciò è dovuto anche al fatto che purtroppo le funzionalità di UDDI, sebbene molto utili, sono considerate ancora troppo limitative. Un'effettiva ricerca di un servizio deve poter mettere a disposizione dell'utente sistemi di *querying* molto più efficaci del semplice *browsing* attraverso indici precostituiti. Ciò a cui si vuole arrivare è, quindi, una interrogazione di tipo *content-based* a cui la ricerca in questo momento sta lavorando.

## 5. DIREZIONI FUTURE

I linguaggi di specifica descritti nel presente lavoro sono ancora in una fase di continua evoluzione. Recente è la pubblicazione della proposta di standardizzazione del linguaggio WSDL nella sua versione 2.0, che prevede nuovi modelli di interazione con i Web Service, oltre a ridefinire alcuni costrutti precedentemente definiti. I motori di esecuzione di processi BPEL sono ancora allo stato di prototipi (per esempio, BPEL4J) anche se alcune proposte abbastanza solide, come Collaxa o, sono in commercio. In realtà, anche le specifiche di BPEL stesso sono in evoluzione, unitamente alle specifiche correlate.

Riguardo alla ricerca, particolare attenzione viene posta, attualmente, sui meccanismi di gestione dei servizi e sulle specifiche di caratteristiche dei servizi che consentano di rappresentare aspetti funzionali e non funzionali dei servizi stessi. In questo ambito, si possono prendere in esame alcune direzioni di sviluppo.

- Gestione dei servizi: è necessario ancora definire meccanismi di gestione e controllo per la pubblicazione e la validazione dei servizi che dovrebbero arricchire UDDI e renderlo uno strumento realmente efficace. Accanto a questo sono allo studio meccanismi per la definizione di politiche commerciali associate all'uso dei servizi in rete secondo strategie sviluppate appositamente per questo tipo di fornitura.
- Definizione delle caratteristiche riguardanti la qualità del servizio: tra le caratteristiche non funzionali dei servizi è particolarmente rilevante la definizione di livelli di qualità dei servizi e di meccanismi per il monitoraggio della qualità. Attualmente, sono state definite alcune proposte, quali WSLA o e WSOL [6, 10]. Per la definizione di caratteristiche di qualità dei servizi anche se in realtà sono necessarie ulteriori ricerche per la definizione di meccanismi di negoziazione e ottimizzazione dei parametri di qualità.
- Definizione di modalità di esecuzione dei servizi in ambienti distribuiti e mobili: sono allo studio, tra l'altro, possibili estensioni delle modalità computazionali utilizzate nel *grid computing* per utilizzare le risorse disponibili in rete per l'esecuzione di servizi distribuiti.
- Definizione di caratteristiche funzionali e non funzionali dei servizi per la ricerca in rete: le funzionalità di ricerca dei servizi disponibili in UDDI sono molto elementari. Proposte alternative sono state formulate nell'ambito del progetto Semantic Web del W3C (*World Wide Web Consortium*) per la caratterizzazione dei servizi basata su ontologie.

L'accesso a servizi in rete e la condivisione di risorse che è possibile effettuare tramite il *service oriented computing* è sicuramente un argomento di notevole interesse per lo sviluppo di nuove applicazioni. Tale interesse è stato dimostrato dalla partecipazione di tutti i principali produttori di software nella definizione di nuove specifiche e piattaforme di esecuzione. Di particolare attualità è l'utilizzo delle tecnologie sopra citate in realizzazioni nell'ambito di comunità chiuse, quali distretti virtuali o tra fornitori e clienti in una filiera produttiva. Le tecnologie basate su Web Service consentono un accesso più immediato alle informazioni e l'interazione automatizzata tra i sistemi informativi aziendali, superando le limitazioni dei sistemi proprietari. Pertanto, l'utilizzo delle tecnologie illustrate nel presente lavoro in ambito B2B è sicuramente di attualità e in forte crescita. Rimangono ancora

da esplorare le potenzialità dell'utilizzo di questi paradigmi nell'ambito di una interazione più globale.

## Bibliografia


[1] Christensen E., Curbera F., Meredith G., Weerawarana S.: *Web Services Description Language (WSDL)* 1.1. http://www.w3.org/TR/wsdl.

[2] Collaxa: http://www.collaxa.com

[3] Communications of the ACM: *Special issue on Service Oriented Architectures*. Ottobre 2003.

[4] IEEE Computer: *Special issue on Web Services*. Ottobre 2003.

[5] Leymann F., Roller D., Thatte S.: *Goals of the BPEL4WS Specification*. xml.coverpages.org/BPEL4WS-DesignGoals.pdf

[6] Ludwig H., Kelle A., Dan A., King R. P., Franck R.: *Web Service Level Agreement (WSLA) Specification*. 2003.

[7] Mitra N.: *SOAP Version* 1.2 Part 0: Primer. http://www.w3.org/TR/2003/REC-soap12-part0-20030624/

[8] Peltz C.: Orchestration and choreography. *IEEE Computer*, ottobre 2003, p. 46-52.

[9] Thatte S. (Editor): *Business Process Execution Language for Web Services*. Version 1.1, http://www-106.ibm.com/developerworks/webservices/library/ws-bpel/

[10] Tosic V., Patel K., Pagurek B.: WSOL - Web Service Offerings Language. In: Workshop "Web Services, e-Business, and the Semantic Web (WES)". In conjunction with CAISE '02. (2002).

[11] UDDI : http://www.uddi.org/

[12] UDDI Group, UDDI Version 2.0 XML Schema, http://www.uddi.org/schema/uddi_v2.xsd

[13] WfMC, Workflow Management Coalition, http://www.wfmc.org

[14] Weerawarana S., Curbera F.: *Business Processes: Understanding BPEL4WS*. http://www.ibm.com

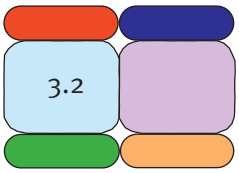


Barbara Pernici è professore ordinario di Sistemi per l'Elaborazione dell'Informazione al Politecnico di Milano. È laureata in Ingegneria Elettronica al Politecnico di Milano e ha un Master of Science in Computer Science della Stanford University. È attiva nella ricerca sulla progettazione di sistemi informativi, sui sistemi informativi mobili, sulla gestione di processi in rete e sulla qualità dei dati. Attualmente è responsabile scientifico del progetto FIRB MAIS (Multichannel Adaptive Information Systems).
barbara.pernici@polimi.it.

Pierluigi Plebani è laureato in Ingegneria Informatica presso il Politecnico di Milano dove attualmente sta seguendo il ciclo di studi per il dottorato in Ingegneria dell'Informazione.
Si occupa di tecnologie basate su Web Service per applicazioni inter-aziendali e applicazioni mobili, e di problematiche legate alla gestione delle politiche di sicurezza.
plebani@elet.polimi.it.

# ESTRAZIONE AUTOMATICA D'INFORMAZIONE DAI TESTI

**Sergio Bolasco**
**Bruno Bisceglia**
**Francesco Baiocchi**




L'analisi automatica dei testi si è sviluppata soprattutto in virtù della crescente disponibilità di tecnologie informatiche e linguistiche e consente di dare una rappresentazione dei testi estraendone alcune proprietà essenziali, capaci di descrivere e interpretare il loro contenuto. Nell'ambito di aziende e istituzioni, è diventato dunque prioritario far fronte alla massa di materiali testuali da gestire quotidianamente, estraendo solo l'informazione d'interesse.

## 1.INTRODUZIONE

Con il termine *Text Analysis* (TA) s'intende un'analisi del testo "mediata" dal computer, ossia basata non sulla lettura del testo, bensì su un'analisi automatica, utile soprattutto quando i testi sono di ampia dimensione[1]. In questi casi, infatti, ogni lettura diretta sarebbe limitata, lunga e difficoltosa, mentre un'analisi automatica è veloce e aperta a "infiniti" confronti, resi possibili dall'uso del computer. Questo approccio ha come obiettivo, fra gli altri, quello di fornire alcune rappresentazioni del contenuto della collezione di testi oggetto di studio (*corpus*) e di estrarre da questi una *informazione*, ossia alcune proprietà, attraverso misurazioni di tipo quantitativo. In una logica di tipo statistico, in questo ambito, si parla anche di "analisi dei dati testuali" (ADT), sottolineando la possibilità di ricavare informazioni strettamente qualitative, a partire da risultanze quantitative, quali sono quelle tipiche della statistica.

L'evoluzione storica degli studi quantitativi su dati espressi in linguaggio naturale, ha visto a partire dagli anni Settanta forti cambiamenti strettamente legati all'evoluzione dell'informatica e alla crescente disponibilità di risorse linguistiche [30] e più recentemente all'enorme dimensione dei testi da consultare *on-line*. Nel corso degli anni, l'interesse per gli studi quantitativi della lingua[2] si è spostato da una logica di tipo *lin-*

[1] È opportuno distinguere la Text Analysis dall'*analisi testuale*, poiché con questa espressione si indica generalmente quella ampia area di ricerca che ha le sue radici in analisi non automatiche, basate su una lettura a più riprese del testo tendente a categorizzare brani, a studiare l'accezione dei termini fino a sviluppare, in taluni casi, un'analisi semiotica d'interpretazione del testo.

[2] Contributi significativi si trovano in riviste quali, fra le altre, Cahiers de Lexicologie, Computers and Humanities, ACM Computing Surveys, Journal of Quantitative Linguistics, Linguisticae Investigationes, Literary and Linguistic Computing, Mots, TAL.

*guistico*[3] (sviluppata fino agli anni Sessanta) a una di tipo *lessicale*[4] (intorno agli anni Settanta), per approdare negli anni Ottanta e Novanta ad analisi di tipo *testuale*[5] o ancor meglio *lessico-testuale*[6].
Fin dall'inizio le applicazioni hanno interessato tutti i campi disciplinari: dai linguisti specialisti negli studi stilometrici o di autenticità dell'autore agli psicologi e antropologi interessati alle analisi di contenuto sia su testi che su materiali provenienti da indagini sul campo (interviste, storie di vita, *focus group*), dai sociologi che si occupano di discorso politico o di indagini qualitative agli specialisti di comunicazione orientati al marketing e al linguaggio veicolato sui principali tipi di media.
In questi ultimi anni, suscita molto interesse, nel filone statistico dell'analisi dei dati testuali, un ulteriore approccio noto con il termine di *Text Mining* (TM): esso è tipico di applicazioni indirizzate alle aziende e istituzioni, le quali, dovendo interagire con enormi masse di materiali testuali spesso disponibili in rete, hanno il problema di selezionare, all'interno di queste fonti smisurate, i dati di loro interesse, per estrarne *informazione* capace di produrre valore. Si tratta di soluzioni orientate al *Knowledge Management* (KM) e alla *Business Intelligence* (BI) che, nella gran parte dei casi, consistono nel ricavare da testi non strutturati quei dati essenziali utili ad alimentare i *database* operativi aziendali con informazioni strutturate, più facili da gestire nei processi decisionali a fini strategici.
Nel seguito, verranno ripresi punti di vista, aspetti teorici ed esempi di applicazione dei vari approcci qui accennati. In questa prospettiva, è opportuno richiamare preliminarmente alcuni concetti di base e relative definizioni dei principali oggetti, visti come utensili da tenere nella "cassetta degli attrezzi" per estrarre informazione dai testi.

## 2. CONCETTI E DEFINIZIONI

La prima lettura automatica dei testi oggetto di studio da parte del computer comporta la cosiddetta *numerizzazione* del corpus: operazione che, a ogni *forma* o **parola** diversa che appare nel testo, fa corrispondere un codice numerico e l'elenco delle collocazioni di tutte le sue *occorrenze* (*token*) nel corpus (Tabella 1), ossia delle loro posizioni lungo lo sviluppo del testo (*discorso*).
Il risultato di questa fase si traduce nella costruzione della lista (*indice*) di tutte le parole diverse che figurano nel testo, il cosiddetto *vocabolario* del corpus, espresso in *forme*

Il termine ***parola*** non trova una definizione soddisfacente: le parole sono gli oggetti linguistici che costituiscono il lessico e sono raccolti nel dizionario. La parola è un segno che ha un senso, è un segno che è simbolo di un concetto o almeno espressione di una conoscenza; si tratta di adottare delle convenzioni precise e il più possibile corrette dal punto di vista linguistico, ma comunque in parte arbitrarie. Una parola può denotare: un oggetto (sostantivo), un'azione o uno stato (verbo), una qualità (aggettivo, avverbio), una relazione (preposizione). Qui nel seguito, per semplicità, si indica con il termine *parola* l'unità di analisi del testo, qualunque essa sia. Va osservato che a seconda degli obiettivi dell'analisi questa unità lessicale può essere una forma grafica, un lemma, un poliforme o una "forma testuale", ossia un'unità di tipo misto in grado di catturare al meglio i contenuti presenti nel testo. In realtà, ormai si sceglie sempre come unità d'analisi del testo una mistura di tipi; quindi nell'articolo con *parola* s'intenda in generale una *forma testuale*.

[3] Per cogliere i rapporti fra lingua e sue concrete possibilità d'analisi (Guiraud, Herdan) [18, 19], si potrebbe seguire un'immagine di Tournier [29], sintetizzata in Bolasco [6]. La dimensione illimitata della lingua fa sì che non sia possibile, per definizione, associare alle parole una qualche "frequenza" in senso statistico-probabilistico. Quest'ultima è invece misurabile su una raccolta di testi, intesi come spezzoni di lessici, ovvero come "campioni" particolari della lingua. È così che ci si limita a considerare le occorrenze delle parole in un testo come un'approssimazione delle frequenze in un lessico, a patto che il corpus sia sufficientemente ampio (almeno 50.000 occorrenze).

[4] Cfr. per esempio Muller [24] e Brunet [8].

[5] In questo approccio l'attenzione sulla testualità del contenuto privilegia l'analisi statistica in forme grafiche (cfr. Lebart *et al.*) [21, 22].

[6] Recentemente si è visto che l'analisi dei dati testuali migliora di gran lunga con l'apporto di meta-informazioni di carattere linguistico (dizionari elettronici, lessici di frequenza, grammatiche locali) e con alcuni interventi sul testo (normalizzazione, lemmatizzazione e lessicalizzazione), cioè attraverso un'analisi statistico-linguistica integrata di tipo lessico-testuale.

| | |
|---|---|
| Occorrenze | Numero di volte in cui una parola appare nel corso del testo |
| Forma | Parola nella sua grafia originale nel testo (forma flessa assunta nel discorso dal corrispondente lemma): esempio *parlavo* |
| Lemma | Forma canonica corrispondente all'entrata del termine nel dizionario, che rappresenta tutte le flessioni con cui quell'unità lessicale può presentarsi nel discorso: esempio *parlare* |
| Tema | Famiglia lessicale di tutti i lemmi derivati da una medesima radice: esempio *parl>* (*parlare, parlato, parlatissimo, parlottante, parlucchiare, parlamentare, parlamento, parlamentarista,...*) |

**TABELLA 1**

*Definizione dei concetti basilari nell'analisi di un testo*

*grafiche* (*type*) con relative occorrenze, come illustrato in tabella 2.

Se si applicano al testo altre operazioni - quali la *normalizzazione*[7], il *tagging grammaticale*[8], la *lemmatizzazione*[9], e/o la categorizzazione *semantica*[10], si produce un corpus che si potrebbe definire "annotato", la cui numerizzazione dà luogo a un vocabolario diverso, per numero di voci (*entrate*) e per quantità di occorrenze a esse associate. Il vocabolario di un testo annotato ha *voci* meno ambigue delle forme grafiche originarie ed è più ricco d'informazioni sul testo[11]. Queste operazioni non sono tutte indispensabili e dipendono dal tipo di analisi di contenuto e dai suoi obiettivi. Si osserva, per esempio, che, analizzando corpus di grandi dimensioni per forme grafiche o per lemmi, i risultati sono sostanzialmente gli stessi. Al contrario, per testi di minori dimensioni o per analisi dei concetti, la riduzione delle parole alla radice comune (*tema*) o al *lemma* fa guadagnare in scoperta di significati e cattura di informazione.

Data la mole d'informazioni presenti nella collezione di testi considerati (*corpus*), non

| | | | |
|---|---|---|---|
| Antenne | 1055 | Contro | 379 |
| Antenna | 792 | Campi | 367 |
| Telefonia | 590 | Elettromagnetiche | 365 |
| Ripetitori | 508 | Telefonini | 361 |
| Cittadini | 499 | Radio | 360 |
| Tim | 498 | Inquinamento | 354 |
| Installazione | 471 | Ripetitore | 339 |
| Impianti | 458 | Cellulare | 332 |
| Salute | 453 | Omnitel | 315 |
| Onde | 442 | Metri | 306 |
| Cellulari | 414 | Legge | 287 |
| Elettrosmog | 407 | Elettromagnetico | 283 |
| Comune | 402 | Limiti | 274 |
| Sindaco | 388 | Wind | 251 |
| Mobile | 386 | | |

**TABELLA 2**

*Esempio di vocabolario: parole piene più frequenti in una rassegna stampa sull'elettrosmog (Fonte: dati Elettra2000)*

7 Per normalizzazione s'intende una serie di operazioni di *standardizzazione* del testo, effettuata sulle grafie attraverso il riconoscimento di nomi propri (persone, società, celebrità), toponimi, sigle, date, numeri (telefonici, prezzi, valute), percentuali, così come individuazione di locuzioni, di tipo avverbiale (in modo, per esempio), aggettivale (di massa, in via di sviluppo), o nominale (identificanti entità ricorrenti: per esempio, Capo dello Stato, Presidente del Consiglio, carta di credito).

8 Il *tagging* consiste nel marcare la *forma* con l'attribuzione della sua categoria grammaticale; per esempio: <parlavo> diventa *parlavo*_V.

9 Lemmatizzare significa trasformare la forma nel lemma corrispondente: per esempio <parlavo> diventa *parlare*_V.

10 L'attribuzione di una etichetta di tipo semantico permette di associare la forma ad altre appartenenti ad una stessa classe di equivalenza (per esempio, <cinema> categorizzato come spettacolo, sarà associabile a <circo>, <teatro> ecc.).

11 Per un riferimento sui corpus annotati si veda il sito: http://www.tei-c.org/.

Gli aspetti *grammaticali* sono risolti con i ***lemmatizzatori automatici***, strumenti per la lemmatizzazione del testo che raggiungono livelli di qualità superiori al 95% nella individuazione del giusto lemma. Questi tools sono basati su catene di Markov e/o sull'utilizzo di *grammatiche locali* che individuano nel testo strutture e regole sintattiche capaci di definire univocamente funzioni grammaticali diverse e quindi risalire correttamente al lemma di un termine (per l'italiano [17]). Questa funzione presuppone ovviamente la disponibilità di un *dizionario elettronico* (in grado di essere utilizzato dal computer) durante la lettura automatica del testo[1].
Gli aspetti *semantici* vengono risolti in parallelo con l'utilizzo di *basi di conoscenza*, dove sono inventariati via via per ogni vocabolo i diversi significati espressi nei dizionari (per esempio, il verbo <andare> prevede oltre 200 significati: "andare al Creatore", "andare a nozze", "andare a male", "andare a letto con i polli" ecc.). Per un riferimento generale, si veda Wordnet sviluppato da G. A. Miller presso la Princeton University (http://www.cogsci.princeton.edu/~wn/).

[1] Per approfondire questi aspetti, fra gli altri, si vedano i lavori di Elia [14, 16] e di Silberztein [28]. Al lettore interessato, per ampliare il glossario sulle nozioni fin qui introdotte [1], si consiglia un testo di "Linguistica elementare": per esempio, De Mauro [13].

è possibile tener conto letteralmente di tutto il testo: si pone, quindi, il problema di *estrarre l'informazione significativa* dal corpus, ovvero quella parte di linguaggio che fa la differenza, che contiene gli elementi caratteristici del contenuto o del discorso espresso nel corpus.
Non tutte le parole hanno, naturalmente, la stessa importanza; ma non è la frequenza l'unico elemento a determinare il peso di un termine in un testo. Anche le parole dette una sola volta (i cosiddetti *hapax*) possono essere molto importanti. Molte fra le parole più frequenti sono "parole vuote" (quali per esempio, <e>, <di>, <da>, <il> ecc., dette anche *stop word*), in quanto elementi necessari alla costruzione della frase; oppure sono parole strumentali con funzioni grammaticali e/o sintattiche (<hanno>, <questo>, <perché>,<non>, <tuttavia>), che non sono portatrici di significato autonomo.
Si considerano, al contrario, "parole piene" gli aggettivi, i sostantivi, i verbi e gli avverbi, in quanto termini che hanno un senso in sé (si veda a tal proposito il riquadro sulla parola); le parole più frequenti celano in sé molti usi e, quindi, molti significati (si pensi alla forma <fine> come nome può voler dire *termine*, *obiettivo* o *scopo*, come aggettivo può significare *raffinato* o *sottile*).
È indubbio, dunque, che il riconoscimento grammaticale, con relativo tagging, risolve non poche ambiguità. A tal fine, esistono strumenti di ***lemmatizzazione automatica***, sia grammaticale sia semantica[12].
Sia gli aspetti grammaticali, sia quelli semantici sono spesso risolvibili solo mediante lettura del *contesto locale*, definito da una, due,...*n* parole che precedono o seguono la parola in esame. È dunque evidente che l'analisi di *sequenze di parole* (o *segmenti*) permette di chiarire il significato presente nel testo, di togliere l'ambiguità ai termini ossia di *disambiguare* il testo. Il significato, per esempio, della sequenza <dato di fatto> è univoco, poiché deriva da una locuzione assai comune[13] che toglie l'ambiguità insita nelle parole semplici come <dato> e <fatto>, che in teoria potrebbero essere verbi, piuttosto che nomi. Queste sequenze, riconoscibili come frasi fisse (*multiword expression*), possono essere individuate già nella fase di *normalizzazione* del testo. Altre disambiguazioni sono, di fatto, realizzate con la lemmatizzazione.
Ma il problema di estrarre l'informazione dal testo non è risolta tanto dalla disambiguazione che semmai serve a non fraintendere un significato con un altro, quanto dal selezionare fra le unità di analisi quelle significative, quelle tipiche o caratteristiche dei contenuti di un testo. In generale, ciò avviene selezionando delle *parole chiave*.
È possibile estrarre queste parole in vari modi:
**1.** con un approccio *corpus based*, si può calcolare un indice, noto come *Term Frequency - Inverse Document Frequency* (TFIDF) [26], che si basa su due assunti:
***a.*** tanto più un termine occorre in un documento tanto più è rappresentativo del suo contenuto;
***b.*** tanti più documenti contengono un termine, tanto meno questo è discriminante [27].

[12] Per l'italiano, come software per la lemmatizzazione automatica dei testi, fra gli altri, si veda Lexical Studio, sviluppato da Synthema, http://www.synthema.it/english/documenti/Prodotti_LexicalStudio_i.pdf.

[13] Le strutture più comuni, ritrovabili nei gruppi nominali, sono del tipo <Nome_Prep_Nome>, <Nome_Agg>, <Agg_Nome>.

L'indice TFIDF è costruito, ponendo a rapporto queste due informazioni[14].

**2.** mediante confronti con informazioni *esterne al corpus*: sia attraverso criteri di *categorizzazione semantica* predisposti sulla base di specifici modelli di analisi, sia attraverso *lessici di frequenza* che costituiscano dei "riferimenti" rispetto ai quali il *sovra-uso* o il *sotto-uso* di un termine nel corpus può assumere un carattere di *specificità*. Nel software Taltac[15], fra le risorse statistiche sono disponibili, per esempio, vari lessici di frequenza che consentono di estrarre il *linguaggio peculiare* di un corpus, mediante il calcolo di uno scarto standardizzato fra le occorrenze d'uso nel corpus e quelle nel lessico prescelto[16].

## 3. L'EVOLUZIONE DEI METODI E DELLE TECNICHE DI ANALISI DEI TESTI: DALLA TEXT ANALYSIS AL TEXT MINING

Lo sviluppo delle tecniche di analisi dei testi ha subito profonde evoluzioni negli ultimi cinquant'anni passando da primordiali indagini semiautomatiche orientate allo studio della frequenza delle accezioni di singole parole in grandi raccolte di testi letterari, ad analisi completamente automatiche in grado di decifrare in profondità il senso di una frase all'interno di sterminate raccolte di materiali testuali quali quelle accessibili oggi dal web. Gli strumenti utilizzati per queste analisi dipendono ovviamente dagli obiettivi, ma si fondano essenzialmente su metodi per il trattamento del linguaggio naturale (*Natural Language Processing*) e su tecniche statistiche di tipo multidimensionale.

### 3.1. Analisi delle concordanze

Ogni metodo o tecnica utilizzata per l'analisi automatica dei testi ha in un modo o nell'altro l'obiettivo di fornire qualche "rappresentazione" del testo, tale da consentirne una lettura mirata.

I primi studi quantitativi sui testi si basavano soprattutto sull'*analisi delle concordanze* che - osservando tutti i contesti locali di una parola d'interesse - consente di discernere i diversi usi e significati di un termine (le sue concrete accezioni nel corpus), per poi confrontare e riunire tali conoscenze in un quadro più complessivo che in taluni casi arriva a definire il lessico di un autore (le prime concordanze furono applicate a studi biblici e risalgono a tempi remoti).

La tipologia delle concordanze presenta una casistica molto ampia. Una discriminazione radicale è espressa dal binomio concordanza verbale – concordanza reale: la prima è concordanza di parole, la seconda è concordanza di cose (concetti, temi, argomenti). Un esempio magistrale di analisi basate sulle concordanze è rappresentato dagli studi di R. Busa circa l'opera di S. Tommaso d'Aquino [10]. Nell'*Index Thomisticus*, la sintesi del lessico di Tommaso ha occupato i primi dieci volumi (su 56) per complessive 11.500 pagine. In questi volumi, sono presenti da una parte tutti i testi con ipertesti interni ed esterni, dall'altra il censimento classificato del vocabolario (il mappale panoramico, secondo l'espressione busiana).

### 3.2. Analisi delle corrispondenze

Per passare da un livello di studio "unidimensionale", quale può considerarsi quello dell'analisi delle concordanze, a uno "multidimensionale" si può utilizzare l'*analisi delle corrispondenze*. È una tecnica statistica proposta inizialmente negli anni Sessanta da J. P. Benzécri [3] come metodo induttivo

---

[14] L'indice *TFIDF* è espresso dalla seguente ponderazione:

$w_{t,d} = f_{t,d} \cdot \log N/f_t$

dove $w_{t,d}$ è il peso del termine $t$ nel documento $d$, $f_{t,d}$ la frequenza del termine $t$ nel documento $d$, $N$ è il numero totale di documenti nel corpus, e $f_t$ il numero di documenti contenenti questo termine.

[15] Software per il Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto: http://www.taltac.it

[16] Si definisce *peculiare* quella parte di linguaggio *tipica* del corpus sia perché è *sovra/sotto-utilizzata* rispetto alla "media" espressa dalle frequenze d'uso nel lessico, sia perché è così *originale* del testo oggetto di studio da non essere presente nel linguaggio assunto come riferimento. Per il calcolo delle occorrenze d'uso (indice d'uso), si veda Bolasco [6].

**Obiettivi e strumenti delle tecniche statistiche di analisi multidimensionale dei testi**

Al fine di analizzarne la variabilità linguistica e la struttura, il corpus viene in genere studiato per *frammenti* (spezzoni brevi di testo: proposizioni elementari o enunciati, singoli documenti, risposte, *e-mail* ecc.) o *per parti* (sub-testi o raggruppamenti dei frammenti per attributi: cronologici, tematici, caratteristiche socio-demografiche ecc.). In questa prospettiva, assume interesse la frequenza delle parole nelle parti o nei frammenti. Infatti, uno studio del testo fondato su base quantitativa, consiste sempre nel confronto di diversi *profili* lessicali, ossia di altrettante sub-distribuzioni statistiche generate dall'insieme delle frequenze delle parole in ciascuna parte e/o frammento. In quest'ultimo caso spesso la quantità di occorrenze viene ridotta a semplice "*presenza/assenza*".

Di fatto queste suddivisioni del corpus danno luogo a matrici di tre tipi diversi: una matrice "frammenti × parole", contenente dati booleani 0/1 (dove 1 indica la presenza della parola nel frammento e o la sua assenza); una matrice "parole × parti", contenente le frequenze con cui ogni parola ricorre nella parte (sub-testo); una matrice "parole × parole" che documenta l'associazione (*co-occorrenza*) di coppie di parole nei frammenti del corpus: qui il dato interno alla matrice può registrare la sola esistenza dell'associazione (0/1) o pesarne l'intensità con una misura di relazione.

Secondo l'algebra matriciale, ogni riga o colonna di queste matrici rappresenta un vettore, descrivente il profilo lessicale. Le tecniche utilizzate per l'analisi di tali matrici mirano alla sintesi o riduzione dei dati, attraverso lo studio della variabilità statistica.

In particolare, le *tecniche fattoriali* - attraverso una riduzione del numero di variabili del fenomeno (vettori colonna) - producono delle nuove variabili sintetiche, in grado di ricostruire i principali assi semantici che caratterizzano la variabilità dei contenuti del testo. L'*analisi delle corrispondenze* è la tecnica fattoriale utilizzata nel caso dei dati testuali. Essa visualizza le principali co-occorrenze fra parole presenti nel testo, sulla base della loro vicinanza nei piani cartesiani costituiti da coppie di assi fattoriali, ricostruendo in tal modo delle vere e proprie mappe del contenuto del testo, che forniscono spesso una rappresentazione globale del senso sottostante il discorso.

Le *tecniche di clusterizzazione* e di *segmentazione* mirano, invece, a ridurre la quantità delle unità statistiche (vettori riga), producendone una classificazione multidimensionale, in grado di definire delle tipologie attraverso le quali leggere simultaneamente le caratteristiche d'interesse.

La *cluster analysis*, come famiglia di metodi di raggruppamento (gerarchici e non, scissori o aggregativi), consente di individuare classi di parole o di frammenti di testo, caratterizzati da una forte omogeneità interna, tale da poter ricostruire i principali "mondi lessicali" presenti nel corpus, ossia i differenti "modi di parlare" del fenomeno studiato, contenuto nel testo.

Per maggiori riferimenti, anche ad altri metodi multidimensionali, si rimanda a Bolasco [6].

per l'analisi dei dati linguistici, assai efficace per trattare matrici di dati di ampie dimensioni, risultanti dalla descrizione di profili lessicali d'interesse. Per esempio, il "profilo" di nomi definito dalle associazioni che questi hanno con un insieme di verbi presenti in un corpus assai esteso di testi: questa informazione è raccolta in una matrice di dati che incrocia le co-occorrenze di nomi (in riga) con i verbi (in colonna). L'analisi delle corrispondenze di tale matrice produce una rappresentazione delle associazioni fra nomi e verbi in maniera tale da riprodurre per vicinanza dei punti su un piano cartesiano la similarità fra profili; questa tecnica si rivelò assai utile a ricavare induttivamente alcune regolarità linguistiche sulla base della cosiddetta distanza del chi-quadro[17]. Per maggiori dettagli si veda il riquadro di approfondimento.

Più recentemente, con il crescere della disponibilità dei testi da analizzare e per rispondere all'incremento esponenziale delle fonti quotidianamente da consultare/interrogare per esempio sul web in aziende o istituzioni, in parallelo alle tecniche di Text Analysis, si sono sviluppate procedure di *Text Mining* (TM) per estrarre informazioni da materiali espressi in linguaggio naturale, riassumibili sotto due "logiche": *Information Retrieval* (IR) e *Information Extraction* (IE)[18]. In maniera assai schematica si può dire che l'IR s'interessa al documento nella sua globalità, mentre l'IE seleziona le informazioni specifiche all'interno del documento, che in genere vanno a popolare un database strutturato.

### 3.3. Information Retrieval

Sono ormai molto diffusi software di *recupero di informazioni* in grado di effettuare ricerche su grandi collezioni di testi sulla base di richieste (*query*) formulate come singole parole o come frasi: l'esempio più comune può essere quello dei motori di ricerca sul web.

Impiegando software tradizionali, i singoli documenti d'interesse sono trattati come entità a sé stanti e, in particolare, non vengono prese in considerazione le possibili relazioni tra i documenti. Nei software sviluppati, invece, su *database*, ai singoli documenti vengo-

[17] Questo processo è definito in dettaglio da Benzécri ([3], p.102-105); svariati esempi di applicazione sono in Benzécri [2].

[18] Per una recente panoramica su questi due punti di vista Poibeau [25].

no di solito associati dei *metadati*. In tal modo, è possibile classificare e pesare in misura diversa i risultati della *query* sulla base di queste informazioni aggiuntive.
Questi software permettono di cambiare in tempo reale la visualizzazione delle informazioni in base alle esigenze del momento, utilizzando un approccio ai dati di tipo OLAP (*On-Line Analytical Processing*).
La fase di Information Retrieval si compone essenzialmente di due sottofasi: la *selezione delle fonti e il recupero dei testi* (unitamente alle eventuali informazioni relative ai metadati).
Ai fini del recupero dei documenti è necessario effettuare una scelta sul tipo di analisi delle parole e/o delle frasi.
Questa analisi può essere di tre tipi (o un insieme combinato dei tre): ortografico, semantico e statistico.
- *Ortografico*: riconoscimento delle parole in base alla loro grafia, senza alcun tentativo di correlarle al contesto.
- *Semantico*: associazione della parola al concetto che vuole esprimere. Parole diverse possono essere usate per esprimere concetti simili, pre-definiti in una base di conoscenza.
- *Statistico*: confronto della frequenza d'uso delle parole con una distribuzione di riferimento (lessico di frequenza).

Nella selezione delle fonti si individuano i soli documenti rilevanti, cioè compatibili con i criteri della richiesta[19]. Le fonti possono essere le più diverse: archivi che contengono informazioni espresse in linguaggio naturale, database strutturati che contengono informazioni già sintetizzate (con o senza metadati), immagini di documenti (in questo caso entra in gioco una componente successiva del processo che si occupa della scansione OCR (*Optical Character Recognition*) per trasformare l'immagine in testo).
Nella fase di IR si estraggono dai documenti selezionati quei frammenti di testo che contengono le parole o le frasi che costituiscono i criteri della richiesta. Soprattutto nel caso di frasi, la qualità dell'algoritmo di selezione e di estrazione è cruciale per ottenere buoni risultati. Per esempio è molto importante controllare la co-occorrenza delle parole e valutare la loro vicinanza all'interno del testo.
Individuate le parole (o le frasi), si calcola un peso per ciascun termine (si può usare la frequenza all'interno del documento, o funzioni più complesse, come l'indice TFIDF precedentemente definito).
Esistono diversi metodi per misurare questa rilevanza: vettoriale, probabilistico e booleano.
- Con il metodo *vettoriale* si rappresentano in spazi geometrici i documenti e le richieste che li hanno "generati": in tale spazio la vicinanza tra richiesta e documento misura la rilevanza di quest'ultimo rispetto alla prima.
- Con il metodo *probabilistico* un documento è tanto più rilevante quanto maggiore è il peso delle parole compatibili con la richiesta.
- Con il metodo *booleano* si valuta la presenza/assenza di parole tra documento e richiesta.

Con l'ultimo metodo si può solo dire se un documento è o no rilevante rispetto a una query, mentre con i primi due, oltre a determinare la presenza/assenza di rilevanza tra documento e richiesta, si genera anche una graduatoria di pertinenza, utile per filtrare i documenti.

### 3.4. Information Extraction

Dopo aver recuperato i documenti rilevanti, occorre sintetizzarne il contenuto informativo e renderlo disponibile per ulteriori analisi. Un compito molto impegnativo, le cui tecniche non sono del tutto standardizzate (nell'ambito del text mining).
La rappresentazione standard di un documento è quella di un vettore nello spazio geometrico definito da un numero di componenti pari all'ampiezza del vocabolario del corpus. Ma questo modo di rappresentare i documenti pone problemi di dimensione, perché cresce con l'ampiezza del vocabolario. Sono stati messi a punto diversi modi per ridurre la dimensione dei vettori-documento, tra i quali, per esempio, il considerare solo le parole significative del vocabolario e, quindi, utilizzare vettori-docu-

[19] A volte questa fase non può essere eseguita in modo automatico e viene affidata ad un esperto del settore.

mento di dimensione pari solo al numero di parole chiave.
La rappresentazione vettoriale dei documenti ha, peraltro, il difetto di non cogliere le relazioni tra parole, portando così potenzialmente a una rilevante perdita di informazione nel passaggio dal "*discorso*" alla sua formalizzazione vettoriale. Sono oggi disponibili varie tecniche per evidenziare l'informazione legata a queste relazioni che si basano in sostanza sullo studio delle co-occorrenze di parole nell'ambito della stessa frase[20]. Studiando le co-occorrenze che superano una soglia stabilita (in termini di frequenza), si cerca di derivare delle regole generali di associazione, che permettano, in relazione al contesto di analisi, di identificare delle sequenze significative di parole (non necessariamente adiacenti).
Un ulteriore passo molto importante, previa un'efficiente disambiguazione delle parole, è la *classificazione* dei documenti. Questa viene eseguita a partire dai metadati eventualmente associati ai documenti, e in genere mediante una lista pre-definita di categorie nelle quali far rientrare i documenti basandosi sulla presenza delle parole e/o delle sequenze più significative in essi contenuti. A tal fine, si utilizzano processi semi-automatici, che possono essere addestrati o che comunque sono in grado di migliorare la loro capacità di assegnazione in base alle operazioni precedenti. L'obiettivo perseguito con questi processi – definito da un punto di vista formale – consiste nell'attribuire un valore *vero* o *falso* a ciascuna coppia (documento, categoria) per tutti i documenti da analizzare e tutte le categorie presenti nelle liste di riferimento [27].
Operando sul versante della sintesi del contenuto, si riconducono le parole e le sequenze che caratterizzano i documenti a classi di significato derivate da una base di conoscenza esterna al corpus: in tal modo è possibile *concettualizzare* i documenti, producendone una rilevante riduzione in termini di dimensione, senza però perdere quantità significative di informazione. Questa fase, detta *summarization*, fornisce una rappresentazione astratta dei documenti che enfatizza i temi qualificanti del testo ed elimina gli altri[21]. La summarization è un pre-requisito per il popolamento di un eventuale database di concetti/azioni/parole d'interesse, che è strutturato in maniera più rigida rispetto a un testo espresso in linguaggio naturale.
Dopo aver classificato i documenti, si pone il problema della loro visualizzazione in un grafico sintetico di facile interpretazione. Normalmente questo problema viene risolto in due modi: mediante tecniche di *clustering*, che permettono di spostare l'attenzione dai singoli documenti a gruppi di documenti, in minor numero e quindi più facilmente rappresentabili; oppure mediante analisi di tipo multidimensionale (metodi fattoriali), che consentono la proiezione dei singoli documenti in spazi geometrici ridotti (tipicamente 2-3 dimensioni).

## 4. ESEMPI DI TEXT ANALYSIS

Al fine di ripercorrere alcune tra le fasi classiche della Text analysis, si illustrano in concreto due casi di studio. Il primo riguarda procedure e risultati di un monitoraggio sull'informazione relativa all'*elettrosmog*. Il secondo concerne un'analisi di messaggi *on-line* scambiati fra insegnanti incaricati di diverse funzioni obiettivo (FO). In entrambi i casi, l'obiettivo dell'analisi consiste nel conoscere il contenuto di fondo dei materiali oggetto di studio, al fine di valutare l'atteggiamento, le intenzioni e i diversi punti di vista degli "autori" dei testi (nella fattispecie, giornalisti o insegnanti).

### 4.1. Campi di applicazione

I campi di applicazione della Text analysis e le fonti di materiali testuali sono stati finora i più diversi. Fra questi, i testi tradizionalmente intesi (letterari, tecnico-scientifici o altra saggistica) rappresentano solo una mi-

[20] Anche se si stima che circa il 10-20% delle relazioni significative tra parole sia di tipo inter-frase, cioè a cavallo di frasi diverse.

[21] Si veda a tal proposito lo studio di Mani e Maybury [23].

nima parte delle applicazioni. Fra i tipi di corpus più studiati figurano quelli relativi a: discorsi politici (parlamentari, elettorali, dibattiti) e relazioni periodiche di pubbliche istituzioni (Banca d'Italia, Onu ecc.); rassegne stampa o intere annate di periodici; documenti tecnico-settoriali (archivi documentali, brevetti); collezioni o raccolte di *testi corti*: progetti, abstract, bibliografie, manifesti politici, messaggi pubblicitari, titolazioni di articoli di stampa, agenzie d'informazione. Ma sono assai frequenti anche analisi di testi prodotti a partire da indagini sul campo: indagini con risposte libere alle domande aperte nei questionari, interviste non direttive o semistrutturate, storie di vita o discussioni di gruppo (*focus group, forum in Internet, chat o news group*). Sono stati, anche, analizzati protocolli clinici, biografie, trascrizioni di messaggi non testuali (linguaggi visivi, musicali, gestuali/comportamentali ecc.), nonché "trascrizioni" del linguaggio parlato attraverso il riconoscimento vocale e infine, recentemente, studi su e-mail e sul lessico degli *sms*.
Queste applicazioni interessano psicologi, sociologi, medici, antropologi, storici, semiologi e specialisti della comunicazione.
La maggior parte delle analisi si fonda sull'interpretazione delle variazioni linguistiche con finalità psico/socio-linguistiche e sul riconoscimento del senso di fondo espresso nei testi. Esempi particolari di analisi, fra gli altri, sono alcuni studi sull'autenticità dell'autore di un documento o sulla dinamica del discorso nelle arringhe processuali oppure analisi del linguaggio in condizioni estreme di sopravvivenza quali quelle che si determinano ad alta quota, nelle profondità sottomarine.

## 4.2. Una rassegna stampa sull'elettrosmog

Lo studio era volto a misurare accuratamente nel tempo e nello spazio la presenza di temi e argomenti intorno al modo di trattare il fenomeno dell'inquinamento elettromagnetico sulla stampa quotidiana, a partire da un campione di testate giornalistiche a diffusione nazionale e locale, in un periodo di quattordici mesi, dall'ottobre 1999 al novembre 2000[22].
La rassegna è formata da 685 articoli[23] raccolti per individuare le caratteristiche generali del *linguaggio* presente nella stampa, con l'obiettivo di catturare la *terminologia* utilizzata e cogliere il livello di attenzione verso i vari aspetti del fenomeno e il loro tipo di percezione.
Le fasi di studio[24] hanno comportato: in primo luogo, una analisi generale del *vocabolario* utilizzato, in termini di forme testuali (parole e locuzioni) più frequenti; in secondo luogo, una evidenziazione dei *lemmi* più ricorrenti *per categorie grammaticali*; in terzo luogo, mediante una riduzione al tema[25] delle principali unità lessicali selezionate, l'individuazione del *linguaggio peculiare*, che ha permesso di quantificare le diverse percezioni del fenomeno.
Osservando il vocabolario già riportato in tabella 2, è interessante notare che il termine *elettrosmog* non è la parola-tema per eccellenza, ma è preceduta da antenne/a, telefonia, ripetitori-installazione-impianti, onde e cellulari. Ciò consente di definire subito l'ampio spettro del "tratto semantico" che ruota intorno all'argomento, ancor meglio inquadrabile dalle espressioni più ricorrenti riportate in tabella 3. Al di là del fenomeno in sé, in essa appaiono anche termini riguardanti il *Comune*, il *sindaco*, la *salute*, che costituiscono tracce importanti del rapporto che il fenomeno ha con l'opi-

[22] Rapporto Interno Consorzio Elettra 2000, Centro di Documentazione, http://www.elettra2000.it.

[23] Il corpus è pari ad un dossier di 750 pagine; la sua analisi ha prodotto un "vocabolario" di oltre 20.000 parole diverse, per un totale di 250.000 occorrenze.

[24] Per un maggior dettaglio su queste fasi si rimanda al Report, disponibile presso il Centro di Documentazione del Consorzio Elettra 2000.

[25] Per riduzione tematica s'intende un raggruppamento delle occorrenze di parole o espressioni secondo la loro *radice* riguardante il tema o significato comune. Ciò può concernere solo più flessioni di uno stesso lemma (<*cellular*+> sta per "cellulare" o "cellulari", siano essi aggettivi o sostantivi) o più entrate dello stesso lessema (<*controll*+> corrisponde alla fusione delle occorrenze di "controllo/i", di "controllore/i" e di varie voci e flessioni del verbo "controllare").

| Telefonia | | | |
|---|---|---|---|
| Telefonia mobile | 348 | Gestori di telefonia | 30 |
| Telefonia cellulare | 266 | | |
| **Elettrosmog** | | | |
| Campi elettromagnetici | 292 | Emissioni/radiazioni elettromagnetiche | 81 |
| Onde elettromagnetiche | 274 | Contro l'elettrosmog | 54 |
| Inquinamento elettromagnetico | 231 | Impatto ambientale | 50 |
| **Istituzioni e Legislazione** | | | |
| Il Sindaco | 229 | Ministero dell' Ambiente | 33 |
| Il Comune | 195 | In regola | 40 |
| Amministrazione comunale | 114 | Nulla osta | 32 |
| Consiglio comunale | 105 | Raccolta di firme | 37 |
| Concessione edilizia | 75 | Contro l'installazione | 35 |
| Legge quadro | 52 | All'unanimità | 32 |
| Legge regionale | 36 | | |
| **Salute** | | | |
| La salute | 334 | Tutela della salute | 53 |
| Salute dei cittadini | 77 | Salute pubblica | 67 |
| **Collocazione Impianti** | | | |
| Stazioni radio | 127 | Centro abitato | 91 |
| Radio base | 98 | Centro storico | 39 |
| Stazioni radio base | 76 | In città | 51 |
| Sul tetto | 103 | Territorio comunale | 36 |
| Nuove antenne | 47 | Campo sportivo | 34 |
| Antenna selvaggia | 33 | In prossimità | 36 |
| Alta tensione | 51 | A ridosso | 34 |
| Ad alta frequenza | 33 | Nelle vicinanze | 34 |
| 6 V | 38 | 50 m | 32 |
| Volt per metro | 31 | Pochi metri | 30 |

**TABELLA 3**
*Tematiche più ricorrenti nella Rassegna Stampa*

nione pubblica e con i problemi legati alla salute.
Successivamente si è proceduto all'analisi dei contenuti specifici degli articoli, al fine di cogliere le diversità di approccio al tema dell'elettrosmog delle diverse testate giornalistiche. Questa fase avviene confrontando i "profili lessicali" dei vari giornali mediante l'applicazione di un opportuno test statistico, che misura lo scarto tra la frequenza dei termini di ciascun giornale e la loro frequenza generale nel corpus. In tal modo, si estraggono le parole ed espressioni *specifiche* di ciascun giornale. Questa tecnica fa emergere i vari modi di percezione, i diversi livelli di attenzione e il tipo di "polemiche" sollevate nella stampa, le cui risultanze tematiche generali sono riassunte nella tabella 4.
La presenza dei temi presenti negli articoli si rileva anche attraverso i verbi, che possono essere raccolti nelle voci generali riportate nel riquadro a fianco.
Da quanto esposto finora emerge che esistono profonde diversità di trattazione del fenomeno, in gran parte dipendenti dall'area geografica d'appartenenza della testata, nonché

| | | |
|---|---|---|
| a = *impianti* | 14% | installare, montare, spostare, smantellare, ... |
| b = *opinione pubblica* | 53% | chiedere, spiegare, individuare, verificare, ... |
| c = *rischio* | 33% | evitare, intervenire, bloccare, causare, ... |

| | |
|---|---|
| 1 | **impianti**: antenn+, install+, ripetitor+, impiant+, tralicci+, apparecch+, lavori, elettrodott+, stazion+ radio base, telefonic+, posiziona+, base, antenn per la tele>, posizione, emittent+, alta tensione, antenn selvagg+, cavi |
| 2 | **cittadini**: cittadin+, contro, chied+, richiest+, protest+, comitato, spieg+, abitant+, società, comitati, condomin+, bambini, ricors+, popolazione, quartiere, persone, denunc+, firma+, gente, battaglia, petizion+, associazione, assemblea, inquilin+, guerra, comunicazione, contro l elettrosmog, proprietari, lotta |
| 3 | **prodotti**: telefonin+, cellular+, telefon cellular+, telefonia mobile, concession+, auricolar+, telefonia, telecomunicazion+, radiotelevis+, televisiv+, consumatori, Gsm |
| 4 | **gestori**: Tim, Omnitel, Wind, gestor+, Telecom, milioni, Enel, Blu, mila, Umts, di proprietà, miliardi, gestor telefonia mobil>, Rai, licenze, licenza |
| 5 | **ambiente/elettrosmog**: ambient+, elettrosmog, problem+, onde elettromagn+, camp+ elettromagne>, emission+, radio, inquin+ elettromagn, Arpa+, territorio, inquin+, onde, esposizion+, camp+, emess+, frequenz+, Ambiente, elettric+, stazioni+, concession ediliz+, centro, camp magnetic+, elettromagnetic+, alta, elettronic+, ministero dell Amb>, impatto ambientale, magnetic+ |
| 6 | **salute/rischi**: risch+, salute, pericol+, preoccup+, Asl, dann+, radiazion+, nociv+, provoc+, tutel+, sanitar+, allarm+, alla salute, cautel+, tumor+, sospen+, evit+, a rischio, salute dei cittadi>, conseguenze, Sanità, leucem+, protezione, salute pubblica, tranquill+, cancr+, Oms |
| 7 | **istituzioni**: sindaco+, Comune, legge, assessor+, comunale, autorizzazion+, approv+, consiglier+, Giunta, Tar, Consiglio Comunale, regionale, Amministrazione comun+, presidente, decreto, ordinanz+, Region+, parere, delibera, responsab+, commissione, Governo, amministrazione, comuni, intervento, Comuni, autorità, indagine, risposta, soluzione, Calzolaio, comunali, misure |
| 8 | **controllo/sicurezza**: controll+, norm+, regolament+, x metri, verific+, rilasc+, volt per metro, sul tett+, sicurezza, provvedimento, microtesla |
| 9 | **ricerca**: scientific+, siti, studi, studio, ricerca, risultati, monitora+, esperti, misurazion+, sito, mapp+, attenzione, ricerche, prevenzione, Università, ricercatori |
| 10 | **territorio**: zon+, limiti, vicin+, scuole, residenti, abitaz+, distanz+, case, città, palazz+, a poc distanz, edifici+, abitar+, limite, urbanistic+, metri, aree, ediliz+, centr abitat+, entro, sportiv+, fino a, vicinanz+, limiti previsti, scuola, livelli, luoghi, in città. |

**TABELLA 4**
*Sintesi delle principali radici lessicali della rassegna stampa, raggruppate per temi*

dall'essere un quotidiano a carattere nazionale o regionale.
Dall'analisi si è potuto evincere tra l'altro come testate quali La Stampa, il Corriere della Sera, Il Sole 24 Ore, Italia Oggi, Il Messaggero e La Repubblica pongono un'attenzione maggiore, sia in assoluto sia rispetto agli altri giornali, a una trattazione del fenomeno in termini di tematiche generali sull'*ambiente*, l'*elettrosmog*, la *salute*, la *ricerca*, ma parlano anche significativamente dei prodotti. Al contrario, testate quali Il Tirreno, Il Secolo XIX, Corriere Adriatico e altri giornali regionali incentrano la loro attenzione su problemi locali e particolaristici, legati ai singoli impianti, al territorio e sono sensibili alle azioni dei cittadini e delle istituzioni di governo locali.

### 4.3. Un forum di discussione fra insegnanti per la formazione a distanza

L'analisi della discussione sulla formazione a distanza dell'Istituto Regionale di Ricerca Educativa per il Lazio era mirata a conoscere il *lessico praticato* in oltre 29.000 messaggi scambiati in un anno sul web fra insegnanti in diverse "conferenze" sulle funzioni obiettivo, nei vari forum provinciali predisposti dalla Biblioteca di Documentazione Pedagogica di Firenze (http://www.bdp.it/).
L'insieme dei materiali testuali, assai voluminoso (1,8 milioni di occorrenze, equivalenti a oltre 6.000 pagine di testo) è portatore di moltissime informazioni, che sono state via via estratte. A parte uno studio a sé stante sulla concatenazione dei messaggi, sono ri-

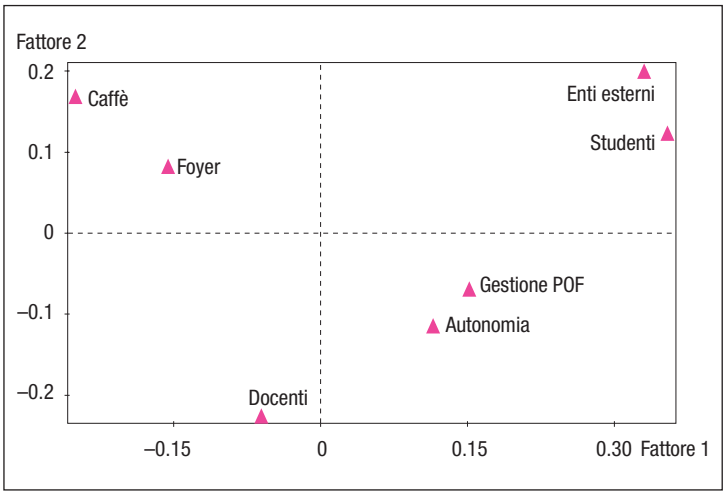



**FIGURA 1**

*Mappa delle 7 conferenze sul sito delle FO*

sultate molto interessanti l'analisi della punteggiatura, l'inizio del messaggio, l'analisi dei verbi e degli aggettivi.

Tutte queste sub-analisi hanno testimoniato l'entusiasmo degli insegnanti nel partecipare alla discussione e nella scoperta di poter comunicare a distanza e saper navigare in Internet, nonché il desiderio di portare la propria esperienza e raccontare eventi con molti particolari[26].

A titolo di esempio, si privilegia in questo articolo l'estrazione dell'informazione limitata ai verbi e i risultati emersi dall'applicazione dell'analisi delle corrispondenze.

Per quanto riguarda l'estrazione dei verbi peculiari, ottenuti confrontando la frequenza dei lemmi nel corpus dei messaggi con quella presente nel linguaggio standard, emergono i toni generali del <*ringraziare, sperare, confrontare, contattare, condividere, gradire, augurare, volere, collaborare, conoscere, piacere, imparare, incontrare, desiderare, coordinare, scusare*>, oltre che i riferimenti all'attività web come <*allegare, navigare, scambiare, inserire, scaricare, visionare, accedere, comunicare*>, che documentano le due principali dimensioni percepite della discussione nei forum.

L'applicazione dell'analisi delle corrispondenze produce invece una mappa della variabilità di linguaggio in funzione dei tipi di conferenze (Figura 1), secondo un *continuum* che dà il *senso* latente del "discorso" sviluppatosi nei forum. L'interpretazione degli assi di un piano fattoriale viene fatta a posteriori a partire dalla disposizione dei punti sul piano cartesiano, o meglio, asse per asse, secondo la graduatoria delle coordinate dei punti sull'asse (per approfondimenti si rimanda a Bolasco [6]). In particolare, nel nostro caso, l'interpretazione del posizionamento delle conferenze in figura 1, si dedurrà dal posizionamento delle parole in figura 2. In quest'ultima figura, seguendo l'asse orizzontale da sinistra a destra si osservano parole coerenti con un *gradiente* crescente di densità di *interessi e argomentazioni* che partendo da un livello minimo del grado di realizzazione, partecipazione, collaborazione nelle attività ("ci sono anch'io", "finalmente", "riuscita" "messaggi", "buon lavoro a tutti" espressioni tipiche delle conferenze "Caffè" e "Foyer", meno ricche di contenuti), passa via via secondo un crescendo del primo fattore nelle cinque conferenze sulle funzioni obiettivo. Dapprima fra le FO "Docenti", poi "Autonomia" e "Gestione del POF" si incontrano "sostegno al lavoro dei docenti", "corso integrativo", "bisogni formativi", "documentazione", "tecnologie", "informatica", "laboratori", "autonomia", "programmazione", fino ad arrivare ai "Progetti con Enti esterni" e ai "Servizi per gli Studenti", che risultano essere le conferenze più "dense" di messaggi argomentati e di attenzioni/sensibilità educative manifestate dagli insegnanti ("rapporti", "genitori", "educazione", "continuità", "dispersione scolastica", "interventi", "progetti", "educazione alla salute", "alternanza scuola lavoro"). Per brevità si lascia al lettore la scoperta di altri contenuti; per approfondimenti si rimanda a Bolasco [7]. Dalla lettura si confermerà l'impressione generale di un crescente grado di interesse nella partecipazione e nella comunicazione delle esperienze, che trova nelle aree di discussione quali il Foyer e il Caffè il suo livello minimo (coloro che sono rimasti "alla finestra", a guardare dall'esterno questo nuovo strumento di comunicazione *on-line*) e nelle aree relative alle

[26] Si rimanda il lettore interessato, all'analisi di dettaglio in Bolasco [7].

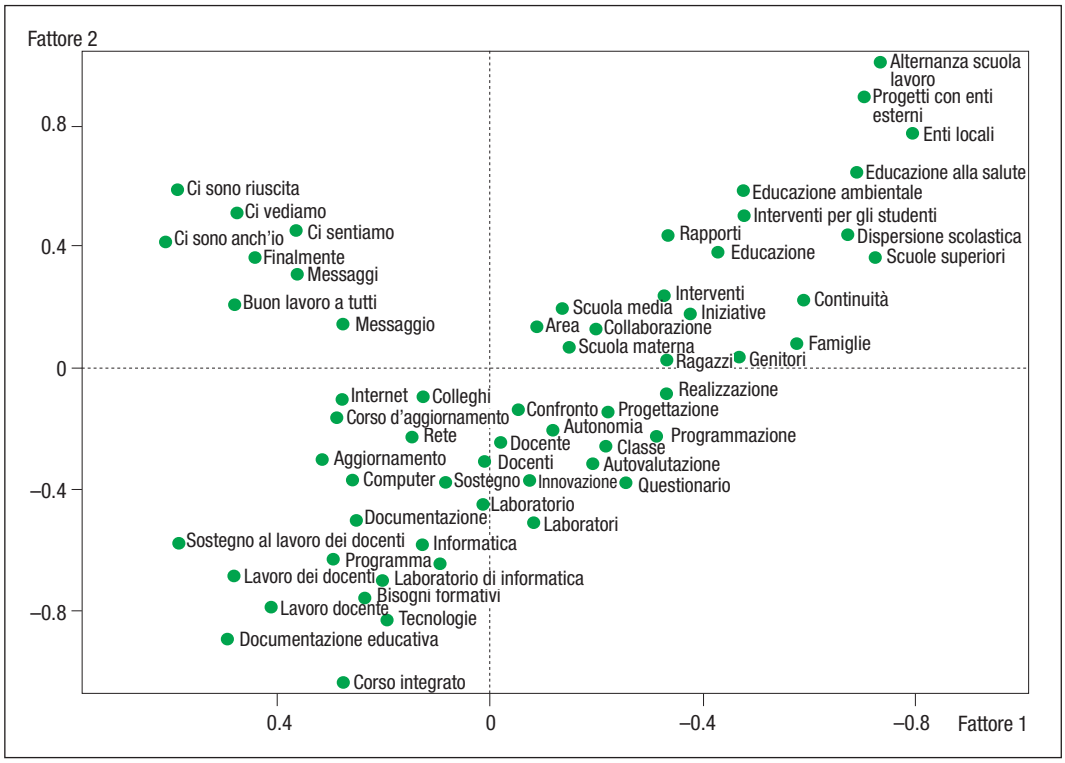



**FIGURA 2**

*Piano fattoriale delle corrispondenze - Mappa delle unità lessicali (parole e sequenze più significative) associate ai forum sulle FO*

FO su Studenti e su Attività con enti esterni il suo massimo (coloro che hanno già *svolto* molte attività e rendono conto delle esperienze già compiute).
Seguendo l'asse verticale, dall'alto verso il basso, si passa dalla dimensione del "progetto" a quella del "corso integrativo", ossia dal generale-ideale (sia Caffè, sia Enti esterni, ma con modalità opposte espresse dal primo asse fattoriale, rispettivamente "da realizzare"/"già realizzato") al particolare-concreto (Docenti). Le parole diverse in ciascun quadrante segnano dunque le differenze fra i linguaggi degli insegnanti.

## 4.4. Ulteriori esempi di applicazione

Da uno *studio sulle encicliche papali* svolto da Bisceglia e Rizzi [4] si rileva come anche solo le alte frequenze siano in grado di selezionare alcuni elementi essenziali dei documenti in esame. L'analisi delle prime 5 occorrenze più utilizzate (*top five*) dagli ultimi quattro pontefici nelle loro encicliche fornisce interessanti elementi che caratterizzano i pontificati (Figura 3). Il termine *fede* è presente solo in Pio XII, i termini *bene* e *sociale* in Giovanni XXIII, il termine *mondo* solo in Paolo VI e *uomo* solo in Giovanni Paolo II. Al contrario, il fatto che *Dio*, *chiesa* e *vita* siano comuni ai 4 papi rende questi termini meno significativi, perché prevedibili.
Nel campo del marketing e degli studi sociopsicologici, un contributo interessante all'analisi semantica nell'estrazione di informazione è dovuto all'*approccio semiometrico* [22]. La semiometria è una tecnica di descrizione dei legami semantici fra parole e si fonda sull'analisi di un insieme selezionato di termini che, al di là del loro significato, evocano ricordi e/o provocano sensazioni gradevoli-sgradevoli[27]. A partire da un campione di individui intervistati che reagisce all'insieme

[27] Si tratta di circa 200 sostantivi, verbi o aggettivi come per esempio: l'assoluto, l'ambizione, l'amicizia, l'angoscia, astuto, il coraggio, il pericolo, la disciplina, risparmiare, efficace, l'eleganza, la famiglia, la gioia, la gloria, la guerra, la giustizia, avventuroso, boheme, concreto, Dio, nobile, sublime, teatro ecc.

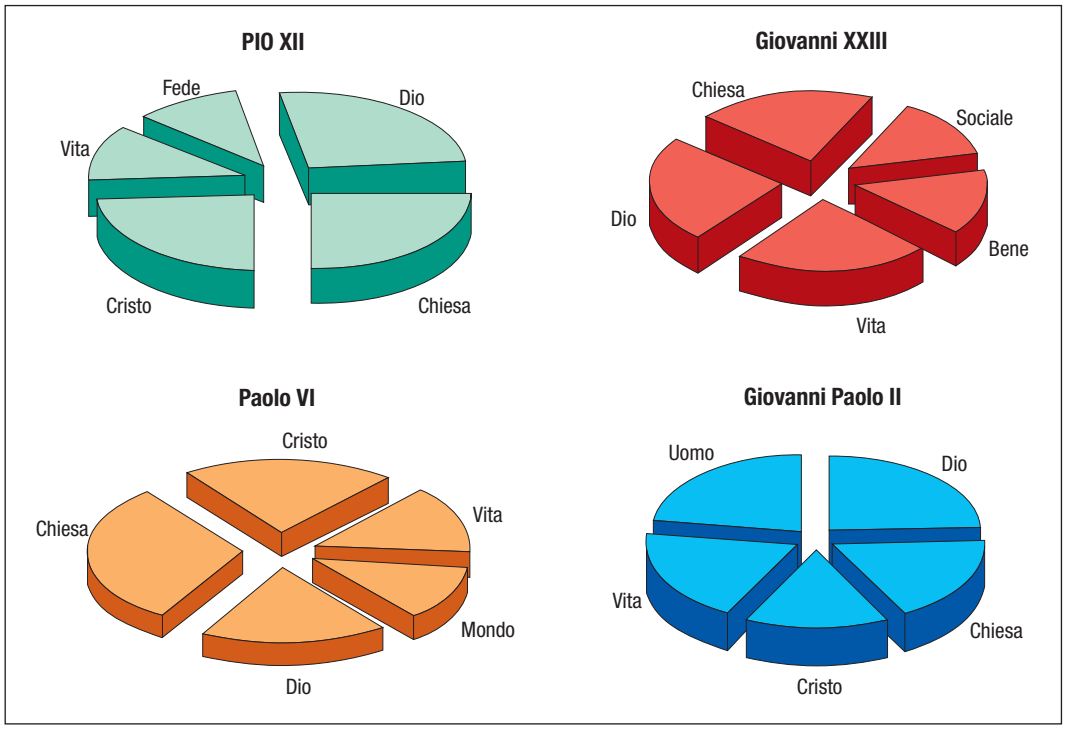



**FIGURA 3**
*Le cinque parole maggiormente utilizzate da questi pontefici in tutte le encicliche da loro promulgate*

di questi stimoli su una scala di sensazioni a 7 livelli da –3 a +3, secondo un crescendo di gradevolezza, si descrivono sistemi di valori e stili di vita. La tecnica di rappresentazione dell'informazione è quella dei piani fattoriali, ottenuti con l'analisi delle corrispondenze, che consentono il posizionamento degli individui su polarizzazioni semantiche quali: dovere/piacere, attaccamento/distacco, spirito/materia e così via.

Per **Technology Watch** (TW) si intende l'attività di monitoraggio della tecnologia, con lo scopo di evidenziare le caratteristiche delle tecnologie esistenti e le loro relazioni, nonché di identificare e descrivere le tecnologie emergenti.
Gli elementi fondanti del TW sono quindi la capacità da un lato di raccogliere tutte le informazioni sulle tecnologie consolidate che potrebbero non essere comunemente note, dall'altro di evidenziare sviluppi tecnologici ancora in uno stato embrionale, cogliendone potenzialità e campi di applicazione e analizzandone le relazioni con le tecnologie già note.

## 5. UN ESEMPIO DI TEXT MINING NEL TECHNOLOGY WATCH

I campi di applicazione del *Text Mining*[28] spaziano dal mining sul web (flussi di navigazione dei siti, comportamenti di visita nel sito), all'analisi delle banche dati sui brevetti (*Patent Analysis*) e più in generale al ***Technology Watch*** dalle azioni di *Customer Relationship Management* (gestione dei *call center, re-routing* di e-mail, analisi di *complaint*, monitoraggio sulla pubblicità dei prodotti, marketing diretto) alla gestione delle Risorse Umane (analisi dei *curriculum vitae* per la selezione di competenze specifiche, monitoraggio dell'Intranet aziendale); dalla classificazione automatica delle risposte in linguaggio naturale nelle indagini Istat (censimento, forze di lavoro, indagini multiscopo) alla categorizzazione dei *document warehouse* aziendali (database di case editoriali, basi documentali di istituzioni pubbliche ecc.).
Un esempio d'applicazione di Text Mining per il Technology Watch si trova all'interno del progetto europeo FANTASIE[29] (*Foreca-*

[28] Per un riferimento generale al TM si veda il sito del progetto europeo NEMIS (*Network of Excellence in Text Mining and its Applications in Statistics*): http://nemis.cti.gr

[29] Si veda il sito: http://www.etsu.com/fantasie/fantasie.html

*sting and Assessment of New Technologies and TrAnsport System and their Impact on the Environment*) sullo sviluppo tecnologico legato ai problemi dei trasporti e del traffico, volto a valutarne la situazione attuale e gli sviluppi a breve e medio termine. Le fonti per le analisi di text mining sono in questo caso interviste, documenti ricavati dalla letteratura e brevetti.

In una prima fase, sono stati intervistati esperti nei campi coinvolti nella ricerca. L'analisi delle interviste ha consentito di creare una prima base di conoscenza, estraendo la terminologia di riferimento in relazione a tecnologie, materiali, mezzi di trasporto e infrastrutture.

In un secondo momento, sulla base di questi riferimenti, è stata effettuata con tecniche di IR la selezione dei documenti rilevanti nella letteratura, sia quella pubblicata su carta che quella disponibile su Internet. Un problema tipico di questo genere di fonte è il ritardo temporale fra lo sviluppo delle ricerche e la loro effettiva pubblicazione, che la rende poco adatta per analizzare tecnologie emergenti. Al contrario, l'analisi su brevetti (*patent analysis*) nell'attività di TW è il metodo più efficace per estrarre conoscenza su argomenti in espansione.

Per questa ragione, la terza fase del progetto Fantasie ha previsto un'analisi di brevetti[30]. A differenza delle interviste e della letteratura, i brevetti sono documenti fortemente strutturati, dove il recupero di informazioni dipende dai campi analizzati (richiedente del brevetto, descrizione delle procedure, materiali brevettati, data ecc.). Sul testo contenuto in questi campi sono state applicate le tecniche di TM a livello morfologico, statistico e semantico per individuare le unità lessicali utili per l'*analisi concettuale*. Quest'ultima, a differenza dell'analisi per parole chiave, permette un *mining* più esteso, in quanto non è vincolata alla presenza di uno specifico termine bensì fa riferimento ad una base di conoscenza (rete semantica, thesaurus ecc.). Dopo aver ultimato la fase di IR e aver selezionato i brevetti pertinenti, sono state effettuate su di essi alcune analisi statistiche di clusterizzazione. A tal fine, sono stati costruiti degli indicatori quali: il conteggio dei brevetti pertinenti rispetto a una query, il numero di citazioni di un particolare brevetto (confrontato con una media di riferimento), il ciclo di vita di un brevetto misurato in termini di durata di una generazione di brevetti.

La combinazione dei risultati di questi tre momenti del progetto di ricerca ha permesso una classificazione dei vari documenti estratti, la cui similarità è servita ad evidenziare i temi più presenti e/o la co-presenza di tecnologie diverse nell'ambito dello stesso argomento, ossia è servita a estrarre l'informazione d'interesse, utile a orientare le future politiche del settore dei trasporti.

Un secondo esempio di TM proviene dal campo biomedico. Cineca (www.cineca.it) ha analizzato circa 400.000 pubblicazioni medico-scientifiche riguardanti il ciclo di vita delle cellule (fonte PubMed: http://www.pubmed.com), con l'obiettivo di individuare automaticamente pattern di parole in grado di selezionare documenti secondo citazioni dirette di nomi di geni o frasi descrittive di concetti altamente correlati con essi.

Nelle fasi di preparazione dei documenti, un particolare rilievo ha assunto l'identificazione dei nomi dei geni (l'analisi grammaticale li identifica come nomi propri), che nella fattispecie costituisce una operazione di IE effettuata utilizzando solo un dizionario con i nomi ufficiali e gli *alias* dei geni.

Sui documenti rappresentati in forma matriciale (indicando la presenza/assenza dei termini di interesse) è stato applicato un algoritmo di clustering i cui risultati vengono utilizzati come base per le successive operazioni di mining. Queste ultime vengono svolte direttamente on-line (in un'area ad accesso riservato), per permettere ai ricercatori di selezionare i documenti di interesse. In risposta a una query per ciascun cluster di documenti viene visualizzata in forma di istogramma la co-presenza dei singoli geni all'interno di ciascun documento.

[30] L'analisi dei brevetti richiede notevoli risorse, poiché le banche dati di brevetti sono consultabili a pagamento, soprattutto per le sezioni contenenti le informazioni più significative.

## 6. CONCLUSIONI

Questa panoramica sulle caratteristiche dell'analisi automatica dei testi ha fornito alcuni scorci sulle concrete possibilità di estrarre informazione d'insieme da un corpus, nella moderna tradizione degli studi di analisi del contenuto o nelle recenti attività di text mining a fini aziendali.

L'ambiguità teorica del linguaggio è fortunatamente assai ridimensionata da un forte effetto "contesto", che in ogni applicazione circoscrive tutte le possibili accezioni di un termine spesso a una sola alternativa o poco più.

Il confronto con lessici di frequenza, attraverso misure di contrasto capaci di valutare il sovra/sotto-utilizzo dei termini, consente l'estrazione del linguaggio peculiare di un corpus. Lo studio delle associazioni di parole completa il riconoscimento del senso da attribuire alle parole chiave di un testo; l'uso di tecniche statistiche multidimensionali ne permette una rappresentazione complessiva, in grado di rivelare anche i principali assi semantici che sono alla base della variabilità linguistica dei testi investigati o il cosiddetto "imprinting" che caratterizza la tipologia del testo (scritto/parlato, formale/informale ecc.).

In sostanza, attraverso la scelta di una appropriata unità di analisi del testo (diversa a seconda degli scopi), è oggi possibile raggiungere un buon livello nella cattura dei significati presenti in un testo, senza necessariamente doverlo leggere o ascoltare.

Questa capacità, fondata essenzialmente sui continui progressi della linguistica computazionale, come si può intuire, è di grande rilevanza. Per esempio, essa aprirà definitivamente la strada a una traduzione automatica, se non "fedelissima", almeno essenziale, che tuttavia non può fare a meno dell'inventariazione di ampie e approfondite risorse linguistiche, quali sono le cosiddette basi di conoscenza (grammatiche locali, reti semantiche, ontologie o altro) e i dizionari elettronici multilingue (non solo di semplici lemmi, ma anche di locuzioni e forme composte).

Le prospettive dei metodi di analisi dei testi (testi ormai quasi tutti disponibili sul web) si indirizzano verso la pratica di un reale multilinguismo, che consenta lo studio simultaneo di informazioni espresse in differenti lingue su uno stesso argomento. Se ne cominciano a vedere alcune concrete applicazioni nell'ambito del *technology watch* e della patent analysis.

## Bibliografia


[1] Beccaria G.L.: *Dizionario di Linguistica*. Einaudi, 1994, Torino.

[2] Benzécri J.P.: *Pratique de l'Analyse des Données, tome 3: Linguistique et Lexicologie*. Dunod, 1981, Paris.

[3] Benzécri J.P.: *Histoire et préhistoire de l'analyse des données*. Dunod, 1982, Paris.

[4] Bisceglia B., Rizzi A.: *Alcune analisi statistiche delle encicliche papali*. Libreria Editrice Vaticana, 2001. Città del Vaticano.

[5] Bolasco S., Lebart L., Salem, A.: JADT 1995 - Analisi statistica dei dati testuali. *CISU*, Roma, Vol. 2, 1995b.

[6] Bolasco S.: *L'analisi multidimensionale dei dati*. Carocci ed., Roma, 1999, p. 358.

[7] Bolasco S.: Analisi testuale dei messaggi nel sito FO, in M. Radiciotti (ed.) *La formazione on-line dei docenti Funzioni Obiettivo. Indagine qualitativa sugli esiti dei forum attivati dalla Biblioteca di Documentazione Pedagogica*. Franco Angeli, Milano, 2001.

[8] Brunet E.: *Le vocabulaire de Jean Giraudoux: structure et évolution*. Ed. Slatkine, 1978, Genève.

[9] Brunet E.: *Le vocabulaire de Victor Hugo*. Slatkine-Champion, 1988, Genève-Paris.

[10] Busa R.: *Index Thomisticus: Sancti Thomae Aquinatis operum omnium Indices et Concordantiae*. Frommann – Holzboog, Stuttgart, Vol. 56, 1974-1980.

[11] Busa R.: *Fondamenti di Informatica Linguistica*. Vita e pensiero, 1987, Milano.

[12] Cipriani R., Bolasco S.: *Ricerca qualitativa e computer*. F. Angeli, 1995, Milano.

[13] De Mauro T.: *Linguistica elementare*. Bari, Editori Laterza, 1998, p. 144.

[14] Elia A.: *Dizionari elettronici e applicazioni informatiche*.In: Bolasco S., et al. (Eds.), 1995, p. 55-66.

[15] Elia A.: *Per una disambiguazione semi-automatica di sintagmi composti: i dizionari elettronici lessico-grammaticali*. In: Cipriani R. e Bolasco S. (Eds.), 1995b.

[16] Elia A.: *Tecnologie dell'informazione e della comunicazione*. In: Gensini S., (ed.) Manuale della comunicazione, Carocci Ed., Roma, 1999, p. 248-257.

[17] Grigolli S., Maltese G., Mancini F.: *Un prototipo di lemmatizzatore automatico per la lingua italiana*, 1992. In: Cipriani R. e Bolasco S. (Eds.), 1995, p. 142-65.

[18] Guiraud P.: *Problèmes et méthodes de la Statistique linguistique*. Presses Universitaires de France, 1960, Paris.

[19] Herdan G.:. *Quantitative Linguistics*, London, Butterworths, 1964. (trad. it.: *Linguistica quantitativa*, Il Mulino, Bologna, 1971).

[20] Lebart L., Piron M., Steiner F.: *La sémiométrie. Essai de statistique structurale*. Dunod, 2003, Paris.

[21] Lebart L., Salem A.: *Statistique textuelle*. Dunod, 1994, Paris.

[22] Lebart L., Salem A., Berry L.: *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht-Boston-London, 1998.

[23] Mani I., Maybury M.T.: *Advances in Automatic Text Summarization*. The MIT Press, 2001, Cambridge (Mass).

[24] Muller Ch.: *Principes et méthodes de statistique lexicale*. Hachzette, 1977, Paris (ristampa: Champion, 1992).

[25] Poibeau T.: *Extraction Automatique d'Information: du texte brut au web semantique*. Hermes - Lavoisier, 2003, Paris.

[26] Salton G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[27] Sebastiani F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, Vol. 34, n. 1, 2002, p. 1-47.

[28] Silberztein M.: *Dictionnaires électroniques et analyse automatique de textes.Le système INTEX*. Masson, 1993, Paris.

[29] Tournier M.: *Bilan critique in AA.VV. En hommage à Ch. Muller: Méthodes quantitatives et informatiques dans l'étude des textes*. Slatkine Champion, Genève-Paris, 1986, p. 885-9.

[30] Zampolli A., Calzolari N.: *Problemi, metodi e prospettive nel trattamento del linguaggio naturale: l'evoluzione del concetto di risorse linguistiche*, 1992. In: Cipriani R. e Bolasco S. (Ed.), 1995, p. 51-65.

Sergio Bolasco, statistico, professore ordinario di Statistica insegna "Metodi esplorativi per l'analisi dei dati" presso la Facoltà di Economia dell'Università degli studi di Roma "La Sapienza"; ha diretto ricerche anche a livello internazionale sulle metodologie per lo studio automatico dei testi e sulle tecniche di *Text Mining*; è autore di un manuale di "Analisi multidimensionale dei dati".
sergio.bolasco@uniroma1.it

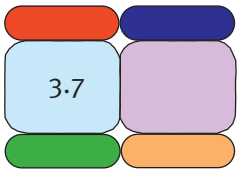
Bruno Bisceglia, ingegnere elettronico, sacerdote della Compagnia di Gesù, ha conseguito il dottorato in Teologia presso la Pontificia Università Gregoriana, da anni si dedica allo studio dei testi mediante microanalisi ermeneutica computerizzata, ed è professore incaricato presso la Facoltà di Ingegneria delle Telecomunicazioni dell'Università del Sannio in Benevento.
bisceglia@unigre.it

Francesco Baiocchi, statistico, tecnico di ricerca nella direzione censimento della popolazione dell'Istat, già consulente del Censis; esperto in sviluppo software, è uno degli autori di Taltac, libreria di programmi per il trattamento automatico lessico-testuale per l'analisi del contenuto.
baiocchi@istat.it

# SISTEMI DI IDENTIFICAZIONE PERSONALE

**Furio Cascetta**
**Marco De Luccia**

In questo articolo vengono illustrate le principali tecniche utilizzate per il riconoscimento automatico delle persone. Per ogni tecnologia trattata (impronte digitali, riconoscimento dell'iride e della retina, riconoscimento vocale ecc.), viene descritto il principio di funzionamento, i principali vantaggi e i limiti operativi. Infine, vengono riportati i tipici campi di applicazione per ogni tipologia di identificazione personale, evidenziandone i possibili sviluppi futuri.




## 1. DEFINIZIONI E CLASSIFICAZIONE

Come punto di partenza è importante introdurre alcune definizioni di base e classificare la varie tecnologie di identificazione automatica - in anglosassone *AIDC technologies* (*Automatic Identification and Data Capture*) - evidenziando le loro funzioni e le specifiche capacità. Come si può osservare dalla figura 1 vi è una notevole varietà di tecnologie AIDC oggi disponibili, suddivisibili in due ampie categorie:

***a.*** **trasporto dati (*data carriers*)**: questa categoria comprende le tecnologie finalizzate alla "raccolta, memorizzazione e trasporto" di dati e informazioni (prevalentemente codificati), su opportuni supporti. Appartengono a questa categoria i metodi ottici (principalmente i codici a barre), i metodi basati sulla memorizzazione magnetica (banda magnetica) e quelli basati sulla memorizzazione elettronica (*RFID-tag, smart card, chip, smart label* ecc.);

***b.*** **riconoscimento di aspetti (*feature extraction*)**: questa categoria contiene a sua volta tre sottogruppi a seconda che il tipo di aspetto "estratto" si riferisca a un'immagine di un oggetto o di una persona (sistemi di visione), oppure sia attribuibile a un'azione dinamica della persona (voce, firma, andatura ecc.), oppure, infine, sia associabile a una proprietà chimico-fisica del materiale costituente l'oggetto (per esempio, i complessi composti chimici responsabili degli odori e delle profumazioni).

In questo articolo verranno trattate le principali tecnologie di identificazione delle persone, basate sul *riconoscimento biometrico* (Figura 2), e in particolare:

- *identificazione personale basata sul riconoscimento biometrico di "aspetti statici"*: cattura ed elaborazione di immagini umane (aspetti anatomici), come per esempio impronte digitali, geometria e "impronta" vascolare della mano, geometria del viso, iride, retina;
- *identificazione personale basata sul riconoscimento biometrico di "aspetti dinamici"*: timbro vocale e modo di parlare (analisi spettrale del campo sonoro), firma dinamica (pressione), digitazione (pressione), andatura (passo).

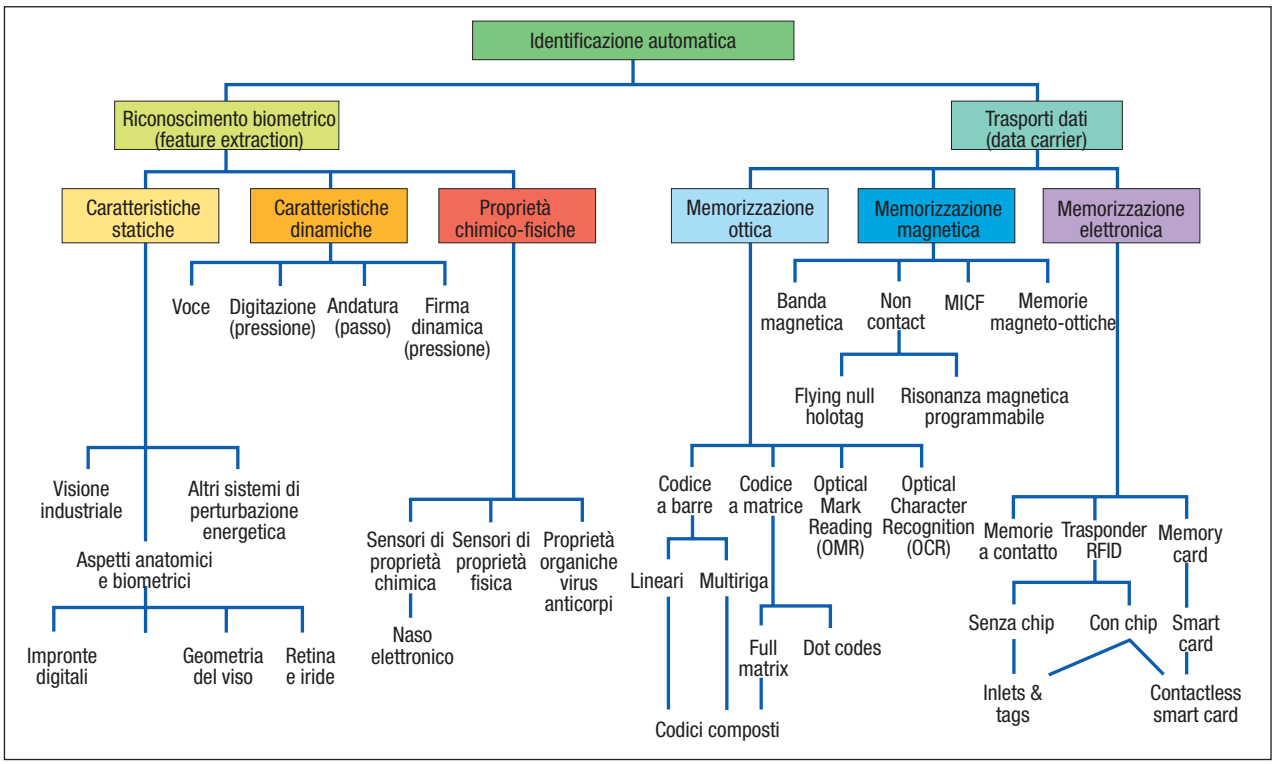



**FIGURA 1**

*Classificazione delle tecnologie di identificazione automatica.* (Fonte: AIDC, Automatic IDentification and Data Capture)

## Storia della biometria

Per migliaia di anni gli uomini hanno istintivamente utilizzato alcune caratteristiche fisiche (come il volto, la voce, il portamento ecc.) per riconoscersi gli uni con gli altri. Circa a metà dell'800, A. Bertillon, capo della sezione identificazione criminali della polizia di Parigi, sviluppò l'idea di usare alcune misure del corpo umano (altezza, lunghezza delle braccia, piedi, dita, ecc.) per identificare i responsabili dei crimini. Verso la fine del XIX secolo, questa idea di partenza fu ulteriormente sviluppata grazie alla scoperta (dovuta agli studi di F. Galton e E. Henry) del carattere distintivo delle impronte digitali, ovverosia del fatto che queste individuano biunivocamente una persona. Subito dopo questa scoperta, le polizie di tutto il mondo cominciarono ad acquisire e memorizzare in appositi archivi le impronte di criminali, detenuti e sospetti. Inizialmente, le impronte erano "registrate" su supporto cartaceo, inchiostrando i polpastrelli dei soggetti in questione e realizzando il "timbro dell'impronta". Subito dopo questa fase, le forze di *intelligence* e di pubblica sicurezza perfezionarono le loro tecniche per il rilievo, sulle scene del crimine, delle impronte digitali lasciate dai protagonisti di azioni delittuose. In questi anni, la polizia comincia a fare sempre più affidamento su tecniche di indagine scientifiche, che si affiancano e quelle tradizionali (logica deduttiva) nelle investigazioni. Segni evidenti di questo nuovo "approccio scientifico" nel condurre le indagini si riscontrano anche in alcuni famosi personaggi della letteratura poliziesca (per tutti, Sherlock Holmes).
La scienza biometrica comincia, quindi, a essere impiegata nelle attività giudiziarie e anticrimine, così come in applicazioni inerenti la sicurezza di un numero sempre crescente di persone. Oggi, in piena èra digitale, un numero elevatissimo di persone utilizza tecniche di riconoscimento biometrico, non solo nel campo della giustizia, ma anche in applicazioni civili e militari. Le previsioni di alcuni analisti di mercato affermano che, entro il 2010, la maggior parte degli abitanti della Terra avrà a che fare, episodicamente o in maniera continua, con le tecniche di riconoscimento biometrico.

## 2. TECNICHE DI IDENTIFICAZIONE BIOMETRICA (BIOMETRIC IDENTIFICATION SYSTEMS)

Queste tecniche "intelligenti" di riconoscimento coinvolgono sistemi esperti, reti neurali, sistemi a logica *fuzzy* e lo sviluppo di sofisticate tecniche di elaborazione elettronica (*computing*). I principali vantaggi di queste tecniche, rispetto a quelle convenzionali, sono connessi alla loro capacità di ricordare e di apprendere.
Gli scienziati da tempo si sono prefissi lo scopo di progettare macchine e sistemi in grado

di emulare alcune abilità umane, tra cui quella dell'identificazione basata su riconoscimento biometrico, ovvero dell'identificazione tramite l'acquisizione e successiva elaborazione di immagini.
Le principali aree di interesse delle tecnologie biometriche sono:
- autenticazione e verifica diretta dell'identità personale (prova dell'effettiva identità dichiarata dal diretto interessato);
- identificazione indiretta di una persona per mezzo delle caratteristiche biometriche disponibili.

Le principali caratteristiche fisiologiche o comportamentali che possono essere utilizzate per l'identificazione personale devono soddisfare i seguenti requisiti essenziali (Figura 3):
- *universalità* (ogni individuo deve avere quella caratteristica);
- *unicità* (non è possibile che due persone condividano la stessa identica caratteristica biometrica);
- *permanenza* (la caratteristica biometria deve rimanere immutata nel tempo);
- "*catturabilità*" (nel senso che la caratteristica biometria deve poter essere misurata quantitativamente).

Il termine "biometrico" (*biometrics*) si addice, dunque, allo studio dei metodi automatici per l'identificazione o l'autorizzazione di persone che utilizzano caratteristiche fisiologiche o comportamentali.
Esempi di tecniche biometriche sono il riconoscimento della geometria della mano, delle impronte digitali, dell'immagine dell'iride, dell'immagine del volto, il modo di parlare, il modo di firmare.
Buoni risultati possono ottenersi anche utilizzando la combinazione di più tecniche di riconoscimento.
Esistono altre tecniche per l'identificazione personale, tra cui il confronto di immagini della retina (*retina image comparison*), confronto della traccia vocale (voice matching), confronto del DNA (*DNA matching*), ma non vengono ancora diffusamente utilizzati in virtù della loro complessità.
Ai fini di una classificazione più generale le tecniche di riconoscimento biometrico si suddividono tra quelle che implicano un rico-

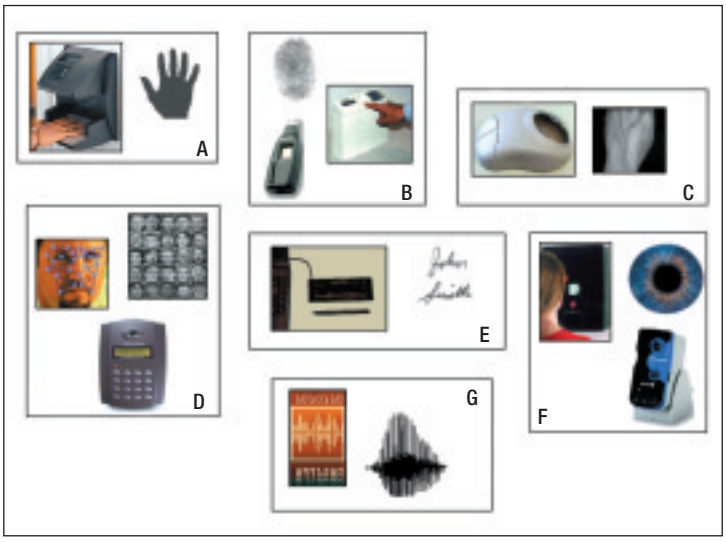



**FIGURA 2**
*Rassegna di alcune tecnologie biometriche*

| Aspetto biometrico | Impronta digitale | Iride | Voce | Geometria facciale | Geometria della mano |
|---|---|---|---|---|---|
| |  |  |  |  |  |
| Limiti alla universalità | Menomazioni o disabilità | Menomazioni o disabilità | Menomazioni o disabilità | Nessuno | Menomazioni o disabilità |
| Unicità | Alta | Alta | Bassa | Bassa | Media |
| Permanenza | Alta | Alta | Bassa | Media | Media |
| Catturabilità | Media | Media | Media | Alta | Alta |

**FIGURA 3**
*Confronto tra alcune tecnologie biometriche in funzione dei requisiti essenziali*

noscimento biometrico di *aspetti statici* (impronte digitali, geometria della mano, del volto, dell'iride ecc.), rispetto a quelle basate sul riconoscimento biometrico di *aspetti dinamici* (voce, firma, portamento ecc.).

## 2.1. Identificazione personale basata sul riconoscimento biometrico di "aspetti statici"

Le principali tecniche "intelligenti" per l'identificazione biometrica tramite il rilievo di aspetti statici sono:

***a.*** riconoscimento delle impronte digitali (*fingerprint recognition*);
***b.*** riconoscimento del volto (*face recognition*);
***c.*** riconoscimento dell'iride e della retina (*iris and retina recognition*);
***d.*** riconoscimento della geometria della mano (*hand recognition*);

### 2.1.1. Riconoscimento delle impronte digitali (fingerprint recognition)

Usato da oltre 100 anni, il riconoscimento dell'impronta digitale è la più antica tecnica di identificazione personale. I primi studi scientifici sulle impronte digitali risalgono già ai primi del Settecento, ma i fondamenti della moderna identificazione biometria sono stati sviluppati da F. Galton e E. Henry verso la fine dell'Ottocento.

Un'impronta digitale è formata da una serie di compositi segmenti curvilinei. Fu proprio Galton, nei suoi studi, a dimostrare il carattere di unicità e di permanenza delle impronte digitali. Gli studi di Henry, invece, condussero alla prima schematizzazione della struttura globale di un'impronta digitale (il celebre "sistema Henry", per la classificazione delle impronte digitali).

Già nei primi anni del XX secolo, le impronte digitali furono accettate come valido strumento per l'identificazione personale. Ovviamente, l'identificazione manuale tramite impronte digitali è un processo lungo, tedioso e costoso. Per cui già nel 1960, da parte della polizia di Parigi e di Londra, si registrano i primi tentativi di studio per la realizzazione di un sistema automatico di identificazione delle impronte digitali.

In epoca più recente, gli studi pionieristici di Galton e Henry sono stati approfonditi e perfezionati. In sintesi, si può affermare che esistono due particolari "aspetti" che caratterizzano un'impronta digitale: i cosiddetti *punti core* e i *punti delta* (Figura 4).

Lo schema a blocchi di un sistema automatico di autenticazione dell'impronta digitale (AFAS: *Automatic Fingerprint Autentication System*) è rappresentato in figura 5. L'*input* al sistema AFAS è l'immagine dell'impronta digitale e l'identità dell'individuo corrispondente; l'*output* è una risposta SI/NO. Il sistema AFAS confronta l'immagine in input con

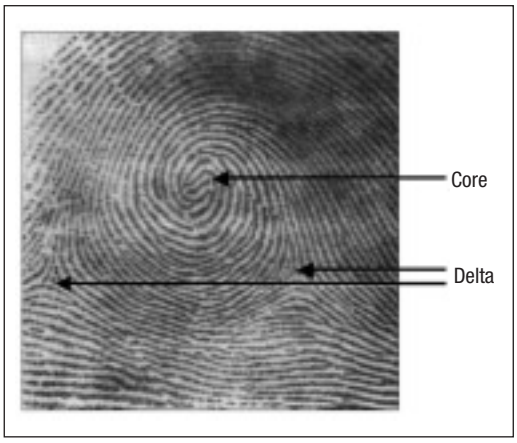



**FIGURA 4**
*Analisi di un'impronta digitale: i punti "core" e i punti "delta"*

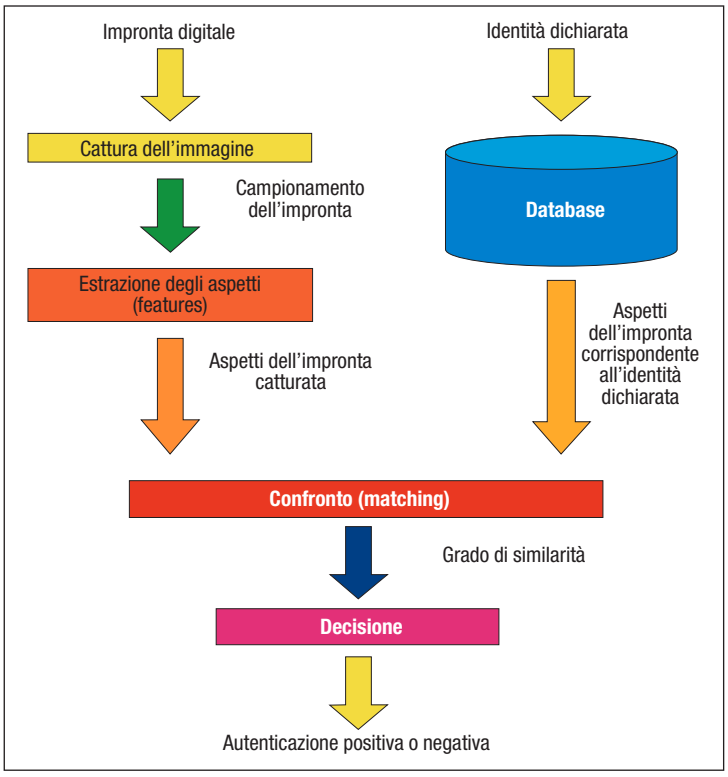



**FIGURA 5**
*Schema a blocchi di un sistema automatico di verifica (autenticazione) dell'impronta digitale (AFAS)*

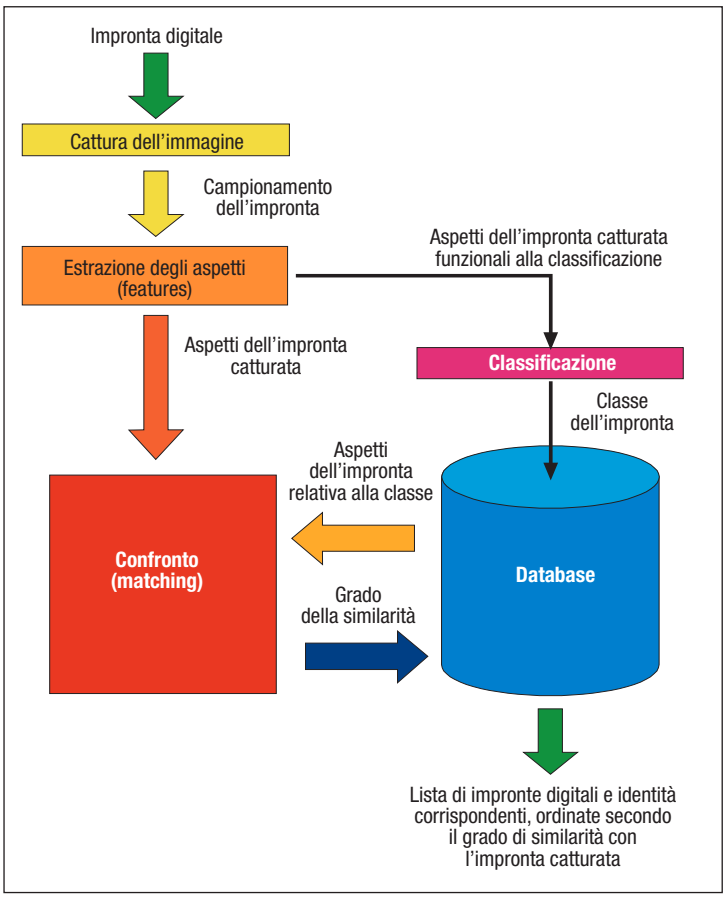



**FIGURA 6**

*Schema a blocchi di un sistema automatico di identificazione dell'impronta digitale (AFIS)*

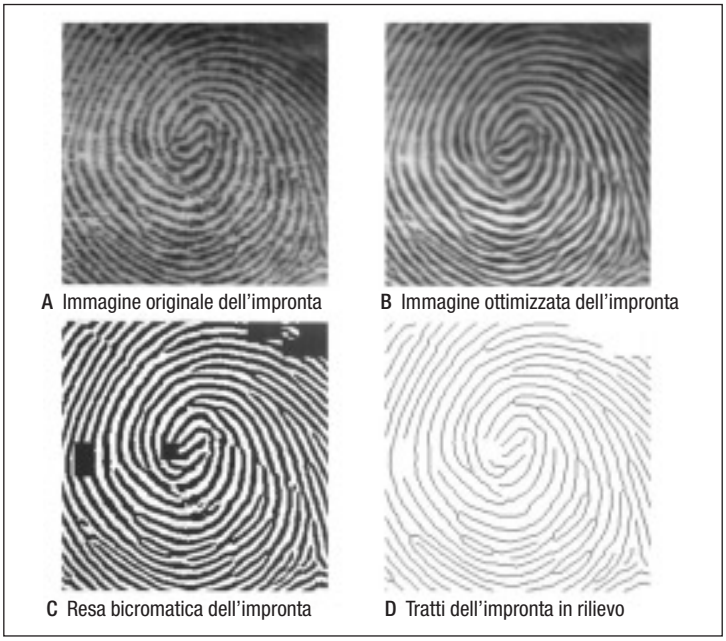



**FIGURA 7**

*Esempio di scansione e di elaborazione elettronica di un'impronta digitale*

un'immagine di riferimento memorizzata nell'archivio (*database*) delle identità.
Al contrario, in un sistema AFIS, *Automatic Fingerprint Identification System* (Figura 6) l'input è costituito unicamente dall'impronta digitale e l'output è rappresentato da una lista di identità di persone di cui si dispone registrata l'immagine dell'impronta (nel *database*) con un "punteggio", per ogni identità, che indica la similitudine tra le due impronte digitali.
Il più antico e conosciuto metodo per "catturare" l'immagine delle impronte digitali è quello di spalmare dell'inchiostro sui polpastrelli del soggetti e di realizzarne l'impronta come se fosse un timbro. L'immagine che ne scaturisce può risultare, ovviamente, molto distorta e, quindi, poco attendibile.
Risultati migliori si possono ottenere con sistemi digitali come, ad esempio, l'acquisizione dell'immagine attraverso una microcamera che realizza una scansione (*scanning*) dell'impronta digitale. Anche in questo caso, però, è possibile ottenere immagini distorte a causa della secchezza della pelle, sudore, sporco o umidità. Tipicamente, l'immagine acquisita è a elevata risoluzione (circa 500 dpi).
Una volta acquisita l'immagine dell'impronta digitale (Figura 7), occorre provvedere alcune complesse elaborazioni elettroniche e informatiche (*fingerprint image processing*).

1. *Riconoscimento degli aspetti*: l'impronta viene rappresentata come un'alternanza di segmenti "solchi" e di "valli", intervallate da discontinuità, dette *minutiae*. Lo stesso Galton definì quattro tipi di *minutiae*, successivamente perfezionate e implementate.
2. *Classificazione delle impronte*: ai fini della classificazione delle impronte digitali, esistono quattro diversi approcci:
   - sintattico (*syntactic approach*);
   - strutturale (*structural approach*);
   - rete neurale (*neural network approach*);
   - statistico (*statistical approach*);
3. *Confronto tra impronte digitali*: il confronto è il processo di comparazione e misura di similitudine tra le geometrie di due immagini di impronte digitali. I principali approcci di confronto sono il confronto puntuale e quello strutturale.

### 2.1.2. Riconoscimento del volto (*face recognition*)

Il riconoscimento del volto è il metodo innato utilizzato dagli uomini per riconoscersi gli uni dagli altri. Le tecniche di riconoscimento del volto, rispetto ad altre tecniche biometriche, presentano il vantaggio di essere non invasive, ovvero di richiedere poca o nulla cooperazione, non essendo soggette a eventuali modifiche di comportamento (volontarie e involontarie) da parte dell'individuo (passivo) sottoposto a riconoscimento.

Grazie alla buona accettabilità da parte degli individui, la tecnica di riconoscimento del volto (*Facial Recognition Technology*) è diventata abbastanza popolare negli USA a partire dalla metà dagli anni '90. Da un punto di vista tecnologico, oltre ai recenti successi dei componenti *hardware* per l'acquisizione delle immagini (microcamere digitali a elevata risoluzione), significativi progressi sono stati ottenuti anche nel campo dello sviluppo dei *software* di riconoscimento.

Le principali tecnologie impiegate nel riconoscimento del volto sono:

- la tecnologia PCA (*Principal Component Analysis*),
- la tecnologia LFA (*Local Feature Analysis*),
- le reti neurali.

In funzione del tipo di applicazione, i sistemi di riconoscimento del volto devono essere progettati e realizzati a seconda del tipo di atteggiamento assunto dall'individuo, che possono essere di tre tipi:

1. *cooperativo*: il soggetto è motivato a utilizzare il sistema per farsi riconoscere e accedere, attraverso appositi varchi (portali, tornelli, porte ecc.), alle aree consentite;
2. *non cooperativo*: se il soggetto è distratto o comunque non si preoccupa né di favorire né di ostacolare il riconoscimento;
3. *ostile o reticente*: quando il soggetto si attiva per evitare il riconoscimento e assume comportamenti evasivi.

Il volto umano è composto da un complesso set di "immagini multidimensionali". Da un punto di vista biometrico, il riconoscimento del volto non è caratterizzato da un'elevata *permanenza*: le molteplici espressioni del volto, l'età, i radicali cambiamenti nel *look* (capelli, barba, baffi ecc.), la presenza di occhiali, sono esempi di caratteri esteriori che possono mutare nel tempo rendendo difficoltoso il riconoscimento facciale. Le caratteristiche "non permanenti" del volto, implicano una notevole complessità di problemi tecnici da risolvere. Ciò nonostante, sono state sviluppate con successo alcune tecniche che consentono di conseguire soddisfacenti e pratici risultati di identificazione personale (*Personal Identification*) a prezzi accessibili.

Oggi le tecniche di riconoscimento del volto vengono utilizzate principalmente in *modalità verifica*, confrontando l'immagine del volto del dichiarante (acquisita in diretta) con quella pre-registrata nel sistema. In modalità identificazione l'impiego è limitato ai piccoli *database*.

### 2.1.3. Riconoscimento dell'iride e della retina (iris and retina recognition)

Un'ulteriore tecnica di identificazione personale (PI) utilizza la caratteristica visibile dell'iride umano. L'iride è la porzione anulare colorata dell'occhio che circonda la pupilla scura (nera) e racchiusa nei tessuti bianchi del bulbo oculare (*sclera*) (Figura 8). Un sistema di riconoscimento dell'iride richiede un apparato di cattura dell'immagine dell'occhio (anche una tradizionale camera CCD: *Charge Coupled Device*, ovvero "dispositivo

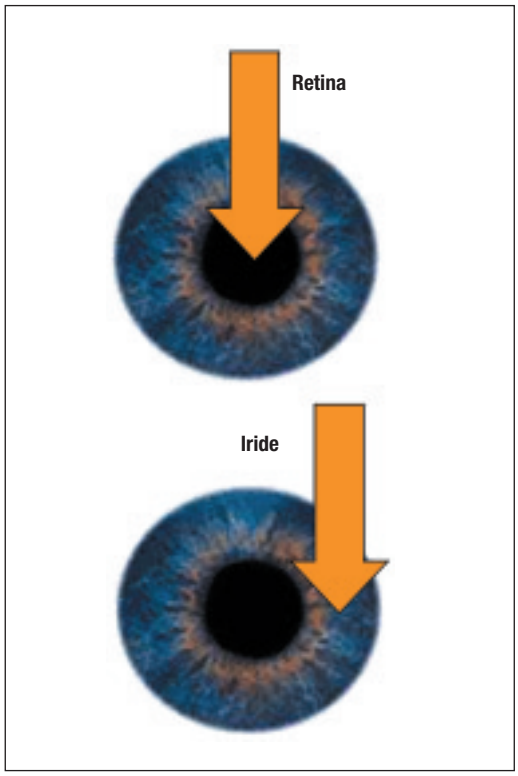



**FIGURA 8**
*Riconoscimento dell'iride e della retina*

ad accoppiamento di carica") e l'utilizzo di appropriati *software* che tramite algoritmi isolano e trasformano la porzione dell'iride in elementi caratteristici dell'identità (detti anche "sagome" o *template*).
L'iride umano è composto da tessuti elastici connettivi che si sviluppano completamente già nell'ottavo mese di gestazione. Il colore dell'iride spesso cambia durante il primo anno di vita, sebbene studi clinici hanno dimostrato che una volta stabilizzato il colore assunto dai tessuti dell'iride si mantiene inalterato per tutta la vita. L'iride è relativamente immune dai disturbi ambientali, ad eccezione della risposta istintiva della pupilla alla luce. Un aspetto molto importante, che lo rende particolarmente adatto per l'identificazione biometrica, è che l'iride di ogni individuo possiede una serie di dettagli e di particolari con spiccati caratteri di unicità.
Il riconoscimento dell'iride è una delle poche tecnologie che ben si adatta a essere utilizzata nella "modalità identificazione"; dotata di buona accuratezza viene impiegata principalmente nelle applicazioni di sicurezza. Richiede una certa collaborazione da parte dell'individuo per la cattura di un'adeguata immagine; non essendoci contatto fisico tale tecnologia è fondamentalmente non invasiva.
Il *riconoscimento biometrico della retina* si basa sull'unicità del suo schema vascolare. Già nel 1930 due oftalmologi scoprirono che ogni occhio umano possiede uno schema vascolare unico e stabile nel tempo (non muta durante tutta la vita dell'individuo).
Il sistema é ancora poco utilizzato: ad oggi, vi è un unico produttore di sistemi di scansione della retina. La retina è localizzata all'interno dell'occhio, nella sua parte posteriore. Uno speciale *scanner* illumina la retina, attraverso la pupilla, con una luce nell'infrarosso (IR) e memorizza le informazioni dalla riflessione del contrasto vascolare.
La scansione della retina viene considerata un'eccellente e accurata tecnica di identificazione personale; grazie alla sua "invulnerabilità" è un sistema molto efficace nei casi in cui è richiesta un'assoluta sicurezza nel controllo degli accessi. La tecnologia non è di facile impiego e richiede sia personale esperto, sia la partecipazione dell'individuo da identificare. Viene considerato un metodo piuttosto invasivo, poiché di solito le persone preferiscono evitare un dispositivo che interagisca con i loro occhi, in quanto lo percepiscono come potenzialmente pericoloso. Ciò rappresenterà un limite all'impiego finché non si riuscirà a realizzare una scansione della retina in maniera più *friendly*.
Il riconoscimento biometrico della retina funziona in maniera soddisfacente sia in modalità verifica (autenticazione), sia in modalità identificazione. Questa tecnica, rispetto alle precedenti è piuttosto costosa. In applicazioni dove è necessaria un'estrema sicurezza viene utilizzata e tollerata, mentre non si addice ad applicazioni che coinvolgono il grande pubblico.

### 2.1.4. Riconoscimento della geometria della mano (hand recognition)

Un sistema di riconoscimento della geometria della mano misura le caratteristiche fisiche (geometriche) della mano (palmo e dita) dell'individuo. La tecnologia principale impiega una telecamera digitale per catturare la *silhouette* dell'immagine della mano, sia il dorso sia il palmo. Alcune misure geometriche (dimensionali, tipo lunghezze, distanze, angoli ecc.) della mano dell'individuo vengono calcolate dal sistema attraverso le immagini acquisite. Il sistema non considera, ovviamente, i dettagli della superficie della pelle (come le impronte digitali).
Sebbene questa tecnologia sia utilizzata con un certo successo da circa 20 anni, è ancora piuttosto dibattuto l'aspetto della "unicità" della geometria della mano: secondo alcuni esperti, infatti, la geometria della mano non è ricca di elementi identificativi univoci così come le impronte digitali o l'iride. Anche l'aspetto della "permanenza" è discusso, poiché molteplici possono essere le cause di insidiose instabilità e cambiamenti nel tempo (età, malattie, incidenti).
Per questi motivi, il riconoscimento della geometria della mano meglio si adatta a esser utilizzato in "modalità verifica" (autenticazione). Considerata un buon compromesso tra prestazioni e facilità d'uso, questa tecnologia viene ritenuta invasiva, poiché richiede un contatto fisico con la mano dell'individuo.

## 2.2 Identificazione personale basata sul riconoscimento biometrico di "aspetti dinamici"

La categoria degli aspetti biometrici di natura dinamica include l'aspetto che, tradizionalmente, è riconosciuto come uno dei caratteri salienti della persona: la voce. Il *riconoscimento della voce* (*voice recogniton*) è, da sempre, una delle forme principali e più naturali di identificazione dell'individuo interlocutore (si pensi alla storia delle comunicazioni a distanza, prevalentemente basate sulla trasmissione della voce). La sua trasposizione nell'ambito dei processi automatici incontra, quindi, il massimo grado di accettabilità da parte degli utenti, superiore anche alla cattura della geometria del volto e nettamente al di sopra delle altre più intrusive tecnologie biometriche.
Tuttavia, il motivo di questa familiarità con i metodi di riconoscimento vocale è anche la causa principale della loro media accuratezza. La voce umana, infatti, è l'unica tra le caratteristiche biometriche a presentare, oltre a una connotazione tipicamente fisiologica, una sensibile influenza comportamentale legata allo stato psicologico dell'individuo, tale da compromettere entro certi limiti il carattere di unicità dell'impronta vocale. Anche elementi di carattere comportamentale propri della voce (quali velocità ed inflessione della parlata) possono comunque contribuire a un processo di riconoscimento vocale.
La metodologia principale finalizzata all'individuazione dell'impronta vocale di una persona consta nell'analisi del contenuto in frequenze delle onde acustiche risultanti dal flusso d'aria generato nei polmoni, propagato attraverso il condotto tracheale e portato in risonanza dalle corde vocali. Se, da un lato, rumore ambientale e sensori microfonici radicalmente diversi possono condizionare drasticamente l'efficienza del sistema di registrazione e verifica dell'impronta vocale, dall'altro va osservato che le metodologie di riconoscimento della voce possono essere facilmente implementate e gestite, in presenza di risorse tecnologiche esistenti nella maggior parte delle strutture informatizzate.
Un ulteriore limite del riconoscimento vocale, che rende questa tecnica biometrica adeguata e conveniente per sistemi di *verifica e autenticazione* di persone in strutture con un numero contenuto di "utenti registrati" nel database, sta nella permanenza del timbro vocale, modificabile nel lungo termine per l'età o degrado fisiologico, nel breve termine per stress, fenomeni influenzali e allergie.
Va citata, infine, la metodologia di *riconoscimento della firma* (*signature recogniton*), che condivide con il riconoscimento vocale la connotazione dinamica e la perturbazione dovuta alla condizione emozionale della persona. Dall'approccio originario, che prevede la stima degli scostamenti degli aspetti geometrici della firma dal modello registrato, si è passati e metodologie evolute e, appunto, "dinamiche" che tengono conto di altre caratteristiche di esecuzione quali la velocità, la traiettoria, l'accelerazione e, infine, la modulazione della pressione durante la scrittura.

## 3. ESEMPI DI APPLICAZIONI DEL RICONOSCIMENTO BIOMETRICO

Come già detto in precedenza, le tecnologie di riconoscimento biometrico possono supportare due differenti logiche: la verifica/autenticazione e l'identificazione.
Nella modalità *verifica/identificazione* il sistema automatico valida l'identità dichiarata da una persona comparando le caratteristiche biometriche catturate (*feature extraction*) con dati e le informazioni biometriche pre-registrate nel database del sistema (*template*). Si parla, in questo caso, di *riconoscimento positivo*.
In un sistema automatico tradizionale per l'autenticazione della persona, l'individuo che desidera essere riconosciuto dichiara la sua identità mediante un codice identificativo personale, numerico o alfanumerico (*PIN, login* o *userID*). Questo codice tipicamente viene immesso manualmente nel sistema, tramite digitazione su tastiera o tramite lettura da un supporto di tipo magnetico o "smart card" (*data carriers*) in possesso del soggetto stesso. L'autenticazione è demandata alla verifica della corrispondenza dell'identità dichiarata con una *password*, o *altro codice di accesso*, immessa nel sistema in un secondo momento e compatibile con il livello di accesso richiesto dall'utente.

**FIGURA 9**

*Controllo degli accessi tramite riconoscimento biometrico*

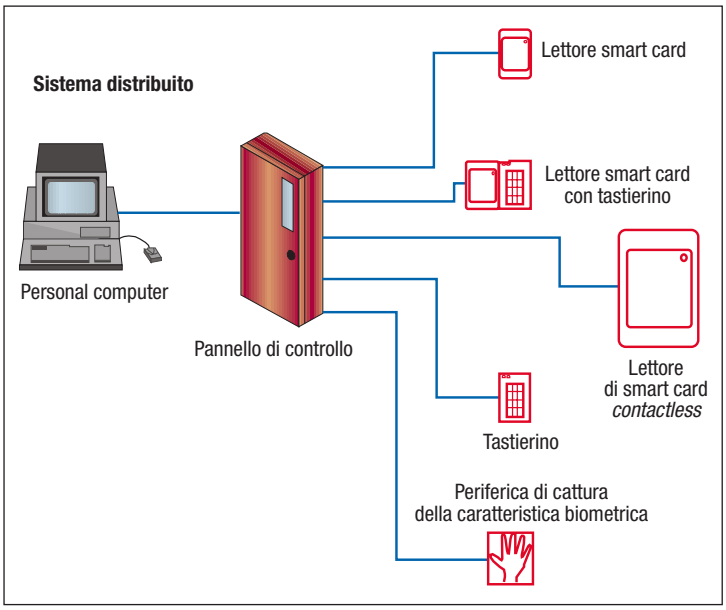



**FIGURA 10**

*Architettura tipo di un sistema di controllo degli accessi tramite identificazione biometrica*

Rientrano nei sistemi tradizionali di accreditamento personale le tessere *bancomat*, le *smart card* GSM con *codice di accesso*, le carte di credito ecc.. In tutti questi esempi, l'identità dichiarata dal titolare viene verificata mediante confronto tra da dati immessi nel POS o telefono cellulare e i corrispondenti dati memorizzati nell'archivio telematico. La trasmissione dati tra la postazione periferica (POS o cellulare) e il database centrale avviene mediante connessione telefonica.

Il punto debole di questi metodi di verifica/autenticazione personale è la loro frodabilità, dovuta al fatto che il codice di accesso o password può essere trafugato o dedotto, e quindi utilizzato fraudolentemente da terzi, evidentemente malintenzionate.

Per limitare e mitigare frodabilità, e quindi per migliorare la sicurezza dei sistemi di verifica dell'identità, si possono integrare le tecnologie tradizionali con quelle del riconoscimento biometrico.

Infatti, il criterio informatore per aumentare la sicurezza negli accessi fisici (di persone) o a sistemi remoti (Internet, Intranet, reti telematiche aziendali ecc.) consiste nel sostituire ai codici di accesso alfanumerici le caratteristiche biometriche (*biometric feature*), strettamente identificative del titolare. Per esempio, aspetti fisici come l'impronta digitale o la conformazione dell'iride connotano la persona in maniera inequivocabile e di difficile contraffazione.

Nel seguito, vengono presentate le principali applicazioni delle tecnologie di identificazione biometrica.

❏ *La sicurezza nel controllo degli accessi fisici e informatici* (banche, tribunali, uffici giudiziari e di polizia, impianti militari, settori strategici di industrie, uffici brevetti, R&D -*Research & Development*- ecc.): è possibile realizzare ingressi con tornelli o *gate* di accesso (porte scorrevoli ad apertura automatica), figura 9, inserendo *badge* magnetici in appositi lettori e validando l'identità tramite l'estrazione di una caratteristica biometrica (impronta digitale, iride, geometria della mano) (Figura 10). Oltre all'accesso fisico, l'identificazione biometrica può essere utilizzata per accreditare personale addetto all'utilizzo dei terminali dei sistemi informativi protetti: in questo caso, le postazioni remote sono attrezzate per l'inserimento di smart card e per l'estrazione e l'elaborazione di aspetti biometrici;

❏ *L'accreditamento a servizi o presso istituzioni* (*firma digitale e firma biometrica*): secondo la legislazione italiana (D.P.R. n.513 del 1997 e relativi Regolamenti Attuativi) la *firma digitale* è un sistema che sigla e attesta l'autenticità di un documento trasmesso per via informatica (Internet, posta elettronica, reti locali, memorie portatili ecc.). La firma digitale sta avendo una certa diffusione per tutto ciò che riguarda i rapporti tra il privato e la Pubblica Amministrazione (per esempio, l'art. 31 del-

la Legge 340/2000 prescrive che tutta la documentazione che le imprese devono inviare alle Camere di Commercio sia elettronica e munita di firma digitale), e si prevede che a breve potranno essere usate anche per le transazioni tra privati. La firma digitale, che in realtà è un software di criptatura, viene rilasciata da apposite società dette *certificatori*, autorizzate dall'Autorità per l'Informatica nella Pubblica Amministrazione. Il certificatore comprova l'identità dell'utente e provvede a creare un *certificato di identità* e due *chiavi* personali (una *privata* e una *pubblica*) inserendole in una smart card che riporta in memoria i dati per l'identificazione. Per attivare la smart card l'utente dovrà digitare un codice segreto. Oltre alla smart card, il certificatore fornisce un lettore da collegare a un PC e il relativo software di firma. Il programma ricava dal testo una serie di caratteri (*impronta*) usando una procedura chiamata *funzione di hash* e, usando la chiave privata, esegue la cifratura dell'impronta. Il destinatario del documento deve aver installato lo stesso software; egli riceve l'impronta cifrata, la chiave pubblica (che può solo decifrare e non criptare) e il documento. Egli applicherà la funzione di hash al documento e confronterà il risultato con l'impronta inviatagli (decifrata usando la chiave pubblica); solo in caso di corrispondenza dei due risultati si certifica la paternità e l'integrità del documento (Figura 11). Si sta pensando, per incrementare la sicurezza di accreditamento telematico, di utilizzare alcune caratteristiche biometriche del titolare (firmatario del documento) registrate nella smart card, per avvalorare l'autenticità dei documenti informatici, e delle transazioni commerciali di una certa importanza, veicolati tramite Internet. In questo ambito, si possono considerare molteplici applicazioni, dalla trasmissione di documenti giudiziari in rete, a garanzia dell'identità del giudice estensore, alle varie forme di *banking* e commercio elettronico (*e-commerce*).

❑ *L'anticontraffazione dei documenti d'identità*: la crescente esigenza di sicurezza ha portato molti Stati, tra cui l'Italia, a realizzare documenti d'identità elettronici. Vari progetti pilota sono in atto sia per quanto riguarda le carte d'identità che i passaporti. Le smart card sono in grado di memorizzare, nell'apposito *chip*, molte più informazioni rispetto ai tradizionali dati anagrafici. In questo caso, le tecnologie dei data carriers possono supportare l'inserimento di impronte digitali o immagini di iride/retina, rendendo molto efficiente e sicuro il riconoscimento del titolare, tradizionalmente affidato alla fotografia che accompagna i dati anagrafici. Al di là degli aspetti puramente tecnici, questi sistemi di controllo delle persone sono, da un lato, particolarmente sicuri, dall'altro pongono dei problemi di carattere etico assai dibattuti, inerenti la "schedatura" dell'individuo e i suoi possibili impieghi contrari al rispetto della *privacy*.

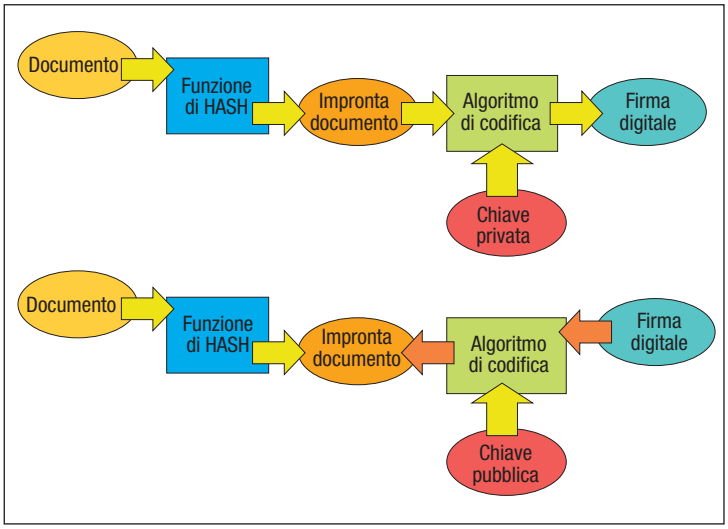



**FIGURA 11**
*Schema logico della "firma digitale"*

A tal proposito va ricordato che nel dicembre 2003 è stato presentato presso l'aeroporto di Fiumicino (Roma) il nuovo prototipo di ***passaporto elettronico***. Si tratta all'apparenza di un comune passaporto, nella cui copertina è stato inserito (non visibile all'esterno) un chip che contiene i dati anagrafici, le impronte digitali e la foto del suo possessore. Scopo principale del passaporto elettronico è quello di evitare le contraffazioni dei documenti, potenzialmente pericolose per la sicurezza del Paese: falsificare un documento cartaceo è indubbiamente più facile che riprodurre un chip con microprocessore. Presentandosi alla frontiera, il viaggiatore dovrà poggiare il passaporto elettronico su un apposito dispositivo di lettura, che innanzitutto verificherà che i dati e la foto riportati in stampa sul documento coincidano con quelli registrati nella memoria del chip. Successivamente, il possessore del passaporto dovrà posare il dito indice della mano destra su un altro lettore: in tal

La risposta alla crescente esigenza di sicurezza internazionale e di gestione delle problematiche dell'immigrazione, da parte delle principali organizzazioni internazionali si è tradotta in un significativo impulso allo sviluppo coordinato di tecnologie per la realizzazione di documenti di viaggio a lettura automatizzata o ***passaporti elettronici*** (*e-passport*). Le diverse realtà nazionali, tra le quali la Comunità Europea, hanno prodotto dei protocolli di sviluppo allo scopo di convogliare le risorse progettuali dei fornitori di tecnologia verso le esigenze valutate prevalenti ai fini della sicurezza nell'autenticazione della persona e dell'anticontraffazione del documento di viaggio.
Le linee guida dello sviluppo di e-passport, comuni per le principali realtà nazionali, sono da individuare nell'adozione di memorizzazione elettronica del dato biometrico su smart card, tecnologia *contactless* di lettura del dato, integrazione di un motore per la crittografia del dato direttamente nella smart card. Gli aspetti tecnologici dettagliati, al di là dei citati elementi fondamentali, oltre che i tempi e le modalità di realizzazione, vengono suggeriti in maniera diversa dai differenti enti e istituzioni. La Comunità Europea, che ha stanziato nel 2003 ingenti fondi per uno studio di settore, ha proposto come indicazione di ricerca per i Governi Nazionali la soluzione costituita da un chip inserito nel passaporto contenente impronte digitali e scansioni retinali. Il Governo Italiano ha recepito le indicazioni comunitarie, portando in fase avanzata il progetto citato nel presente articolo, che ha previsto la realizzazione di un impianto, presso l'aeroporto di Fiumicino di autenticazione automatica mediante la lettura in RF di dati biometrici (l'impronta digitale), memorizzati in un chip contenuto nella copertina cartonata del passaporto, da confrontare con l'impronta catturata in tempo reale presso il *gate* aeroportuale.
Se la scelta di un sistema integrato, costituito da chip e antenna e caratterizzato dalla necessaria flessibilità per essere inserito nel tessuto semirigido della copertina del passaporto, accomuna il Governo Italiano a quello degli Stati Uniti, anche in ambito europeo diverse realtà nazionali hanno optato, forse per differenti esigenze di 'percezione' della lettura elettronica, per l'inserimento nel passaporto tradizionale di una tasca dedicata all'alloggiamento della smart card contenente il dato biometrico.
Nell'area istituzionale delle Nazioni Unite, l'ICAO (*Internationa Civil Aviation Organization*) ha suggerito l'impiego di *chip contactless* inseriti nel supporto cartonato e contenenti, come *feature* biometrico, l'immagine del volto della persona: in merito al formato dell'immagine (del volto come dell'impronta digitale), l'organizzazione internazionale si è espressa in maniera contraria alla compressione del dato in JPEG o JPEG 2000, non in uso nei database biometrici, come pure alla vettorializzazione dello stesso in "*template*", per ovviare alla confusione dovuta alle molteplici tecnologie proprietarie per l'estrazione di "*template*" biometrici e relativa lettura. Sul versante dei produttori di *hardware*, la richiesta di *smart card chip* per documenti di viaggio rappresenta una significativa opportunità di mercato in alternativa ai settore trainanti (i sistemi di telefonia GSM e le carte di credito elettroniche, che però interessano prevalentemente l'area europea, per le differenti esigenze tecnologiche nell'accreditamento del terminale telefonico o del pagamento elettronico espresse negli Stati Uniti). È ovvio, quindi, che i principali produttori di smart card rispondano con interesse alle richieste di tecnologia per la realizzazione del passaporto elettronico. La proposte più concrete e affidabili convergono sull'adozione di chip con 64 kbytes di memoria base (al posto dei più diffusi con 32 kbytes), in grado di impiegare 20 Kbytes per conservare l'immagine del volto, 10 kbytes per una prima impronta digitale, altri 10 kbytes per una seconda e i restanti per i dati alfanumerici relativi alla persona e il *software* di gestione. Soluzioni *high end* prevedono, per la smart card, da configurazioni con 300 kbytes di ROM e 128 kbytes di EEPROM fino a 400 kbytes di "sola" EEPROM. La funzione di verifica crittografica del dato biometrico memorizzato può essere effettuata direttamente dalla smart card, se dotata di opportuno *core* di calcolo dedicato.
Il prodotto integrato di massimo livello tecnologico non può che nascere dalla *partnership* tra il produttore di smart card chip e quello del supporto cartonato con l'alloggiamento per il chip, oltre che con la società di sviluppo del codice di crittografia, tale da garantire una catena produttiva efficiente e soprattutto tracciabile in ogni suo stadio.

modo, viene verificato che le impronte digitali catturate "dal vivo" coincidano con quelle registrate sul chip. L'introduzione del passaporto elettronico richiederà ovviamente risorse economiche e tempo, poiché dovrà coinvolgere tutte le Prefetture che dovranno essere dotate di idonee apparecchiature per la registrazione dei dati sul chip inserito nel passaporto. Anche tutti gli aeroporti internazionali dovranno avere gli appositi dispositivi di lettura. Secondo le previsioni del Ministero degli Esteri, i primi documenti della nuova generazione saranno distribuiti verso la fine del 2004. Da un punto di vista del rispetto della *privacy*, tale sistema con verifica diretta "on-site" (senza l'impiego di un database centrale) tra i dati memorizzati sul chip e quelli catturati presenta minori problemi di accettabilità, in quanto il titolare del passaporto detiene il possesso esclusivo dei propri dati biometrici.
In tutte le suddette applicazioni, l'autenticazione avviene attraverso l'acquisizione in diretta di un aspetto biometrico (impronte digitali, iride, ecc.) e la sua verifica, o in locale (pre-memorizzazione su smart-card) o in remoto, attraverso l'accesso al database centrale.
Accanto alla verifica/autenticazione, va citata anche la procedura di *identificazione*, nella quale il sistema automatico confronta l'aspetto biometrico in ingresso con tutti i template memorizzati nel database, senza alcuna dichiarazione d'identità da parte dell'individuo. Questa modalità implica consultazioni di archivi e database anche di notevoli dimensioni. Le applicazioni sono prevalentemente nel campo della giustizia e della pubblica sicurezza (polizia e *intelligence*): da un rilievo di impronta digitale catturata sulla scena di un crimine, è possibile risalire al potenziale criminale andando a interrogare il database dei soggetti schedati. Tale tecnica di identificazione (detta *negative recognition*) viene utilizzata per restringere il campo dei possibili responsabili di un atto criminoso, da milioni di individui a qualche centinaio. Il sistema informatico, infatti, esclude dalla lista dei so-

spetti tutti gli individui "schedati", la cui impronta è palesemente difforme da quella in oggetto. Le impronte giudicate simili e raggruppabili in classi, concorrono a formare sottoinsiemi ridotti, attribuendo a ogni impronta un punteggio (*score*) di similitudine con quella oggetto della ricerca.

## 4. CONCLUSIONI

L'evoluzione tecnologica dei sistemi di identificazione automatica è in una fase di crescita significativa soprattutto in termini di flessibilità e affidabilità applicativa, lasciando intravedere nella attuale realtà "digitale" una graduale apertura verso campi di applicazione sempre più numerosi, che un tempo richiedevano necessariamente l'intervento e la discrezionalità dell'operatore umano. La strada tecnologica, dunque, può portare in tempi brevi verso obiettivi di maggior sicurezza e semplificazione dei processi in una moltitudine di applicazioni. Tuttavia, al di là dei settori che interessano la sicurezza fisica delle persone, delle comunità e degli Stati, l'elemento moderatore dell'applicabilità dell'identificazione personale è certamente la tutela della *Privacy*. Le recenti reazioni che i consumatori americani hanno espresso, attraverso le loro associazioni, contro gli sviluppatori di tecnologie in grado di tracciare anche soltanto i prodotti preferiti dal singolo individuo nella distribuzione commerciale (mediante l'impiego di *trasponder* in radiofrequenza, ormai soprannominati "spychip"), lasciano intendere come la gestione dell'identità personale e l'accreditamento automatico possano ancor più facilmente essere percepiti come un abuso, quando non strettamente legati alla *security*. La cautela, nell'implementazione dei sistemi di identificazione personale, e la cura del grado di invasività percepita, non solo fisica ma anche nella gestione del dato raccolto, diventano determinanti per le diverse tecnologie di *auto-ID* disponibili, forse più della loro affidabilità ormai consolidata.

## Bibliografia


[1] Ashbourn J.: *Biometrics: Advanced Identity Verification, The Complete Guide*. Springer, London, 2000.

[2] Campbell J.: Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, Vol. 85, n. 9, September 1997, p. 1437-1462.

[3] Daugman J.: High Confidence Visual Recognition of Persons By a Test of Statistical Independence. *IEEE Transactions on Pattern Analysis and Machine Intellegence*, 1993, p. 1148-1161. http://www.labs.bt.com/library/papers/PAMIpaper/PAMIpaper.html. Last accessed: 30 July 2001.

[4] Daugman J.: *Wavelet demodulation codes, statistical independence, and pattern recognition*. Institute of Mathematics and its Applications, Proc. 2nd IMA-IP, 2001, p 244-248.

[5] Jain A., Bolle R., Pankanti S., editors: *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Press, Boston, 1999.

[6] Jain L.C., Halici U., Hayashi I., Lee S.B., Tsutsui S., editors: *Intelligent biometric techniques in fingerprint and face recognition*. CRC Press, Boca Raton – Florida - USA, 1999.

[7] Prabhakar S., Pankanti S., Jain A. K.: *Biometric Recognition: Security and Privacy Concerns*. IEEE Security & Privacy, March-April 2003, p.33-42.

[8] Zhang D.: *Automated Biometrics Technologies and Systems*. Kluwer Academic Publishers, Boston, 2000.



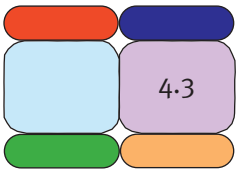
FURIO CASCETTA, professore ordinario presso la Facoltà di Ingegneria della Seconda Università di Napoli. Studioso ed esperto di sistemi di misura, di automazione e controllo, da più di venti anni collabora attivamente con le principali associazioni di categoria del comparto, con i più prestigiosi centri di ricerca nazionali e internazionali del settore e con gli organismi di normazione (sia a livello italiano che europeo).
Dirige la collana editoriale *Misure e Automazione* per l'editore Franco Angeli (Milano).
Collabora a progetti di Alta Formazione, oltre che con l'Università di Napoli, anche con altri Atenei nazionali, tra cui il MIP-Politecnico di Milano, il Politecnico di Bari, l'Università di Palermo, e l'Università Mediterranea di Reggio Calabria.
È autore, o co-autrore, di circa 100 pubblicazioni scientifiche (sia su riviste nazionali che internazionali) e di numerosi libri scientifici , didattici e divulgativi.
fcascett@unina.it

MARCO DE LUCCIA, ingegnere meccanico, da anni collabora con l'area "misure ed automazione" dell'Università di Napoli "Federico II" e della Seconda Università di Napoli. Esperto e appassionato ricercatore nel settore delle nuove tecnologie ICT applicate ai sistemi di misura e telecontrollo. È coautore di articoli tecnici e divulgativi su riviste scientifiche del settore.
marco.deluccia@fastwebnet.it

# INFORMATICA E BIOLOGIA DEI SISTEMI

**Corrado Priami**

In questo contributo si esaminano i più recenti sviluppi informatici in campo biologico. Si mostra come l'evoluzione delle scienze biologiche abbia imposto un cambio di paradigma di riferimento nella bioinformatica: ovvero, passando dai progetti di sequenziamento del genoma alla genomica e proteomica funzionale le tecniche informatiche maggiormente adatte si rifanno alla teoria dei linguaggi di programmazione con particolare riferimento alla concorrenza e alla mobilità del codice.




## 1. INTRODUZIONE

L'uso di tecniche informatiche in domini applicativi legati alla biologia risale ormai agli anni ottanta quando venne coniato il termine **bioinformatica**. Essenzialmente si trattava di memorizzare ed esaminare la grande quantità di dati che veniva prodotta dai biologi. Pertanto, i settori dell'informatica maggiormente interessati furono gli algoritmi, le basi di dati e alcune branche dell'intelligenza artificiale (in particolare, reti neurali) per cercare di estrarre dati significativi e fare predizioni da insiemi spuri di dati prodotti da esperimenti. Le tecniche informatiche maggiormente usate per questi scopi riguardano modelli statici di fenomeni biologici; nessun riferimento viene fatto a possibili evoluzioni funzionali e temporali dei fenomeni. Tuttavia, la definizione di bioinformatica fornita da Hwa Lim che ne coniò il temine è "studio del contenuto informativo e del flusso informativo in sistemi e processi biologici". Nei primi decenni di bioinformatica si è guardato quasi esclusivamente al contenuto informativo (e, quindi, a tecniche statiche) e si è ignorato il flusso informativo.

Per capire le motivazioni che privilegiano la bioinformatica statica a quella dinamica occorre guardare brevemente all'evoluzione che ha avuto la biologia in questi ultimi anni. Il lancio del progetto "genoma umano" ha portato alla scoperta di quantità di dati superiori alle aspettative di chiunque e in breve tempo. A questo punto, conoscendo tutti i geni che compongono il DNA umano si apre una nuova sfida che spesso viene indicata come genoma funzionale (*functional genomics*). Attualmente, si conoscono tutti i "mattoncini" del funzionamento del corpo umano (i geni) e di alcuni si conoscono anche le funzionalità se presi in isolamento, ma ben poco si sa di come i geni e le proteine che questi codificano si comportano in situazioni normali o patologiche. Da qui, la necessità di nuove tecniche per modellare il comportamento di sistemi biologici, e non solo la loro struttura come avveniva fino a pochi anni fa.

La complessità dei sistemi da trattare è tale da non consentire uno studio accurato e

Il termine **"Bioinformatica"** è stato coniato da Hwa Lim alla fine degli anni ottanta per indicare l'applicazione di tecniche informatiche nel dominio applicativo delle scienze della vita. La definizione proposta recita: "lo studio del contenuto informativo e del flusso di informazione nei sistemi e nei processi correlati alla biologia." Tuttavia, questa definizione è unilateralmente legata alla biologia e questo non consente di sviluppare la dignità paritetica che informatica e biologia devono avere in questa area di ricerca al fine di ottenere importanti risultati per entrambe le discipline. A questo proposito una definizione migliore di bioinformatica potrebbe essere "La Bioinformatica è il campo della scienza in cui biologia e informatica si fondono in una unica disciplina per facilitare nuove scoperte biologiche e determinare nuovi paradigmi computazionali sul modello dei sistemi viventi." Questa definizione presa da NCBI al sito www.ncbi.nlm.gov/Education è molto generale, ed evidenzia nettamente la necessità di interazione tra informatici e biologi e, quindi, la necessità di costruire un linguaggio comune alle due discipline per poter interagire e collaborare. Questo può essere fatto solo mediante lo sviluppo di curricula formativi interdisciplinari e mediante l'attivazione di grandi progetti di ricerca. Dal punto di vista dei contenuti, la nuova disciplina deve sicuramente comprendere la definizione di tecniche statistiche e algoritmi necessari a studiare la grande mole di dati che si rende disponibile come risultato degli esperimenti, la definizione di strumenti di scambio e memorizzazione di informazioni accessibili su larga scala, la definizione di metodologie di rappresentazione e simulazione del comportamento di sistemi complessi come le reti geniche o metaboliche o i meccanismi di segnalazione *intra* e *inter*-cellulari. Infine, la bioinformatica dovrebbe assumere il ruolo che ha la matematica per la fisica, e cioè quello di fornire le basi teoriche per i recenti sviluppi biologici nelle aree *omics* (genomics, proteomics, metabolomics ecc.). Questa visione ultima è quella più cara ai biologi teorici che vorrebbero vedere inserita in questa disciplina la loro lunga esperienza sulle teorie evolutive accoppiata con i recenti sviluppi sulla genomica funzionale. Da qui il ruolo primario di modellazione e analisi di sistemi che si vuole affidare alla bioinformatica e che esamineremo più in dettaglio rispetto alle altre possibilità in questo contributo. L'obiettivo ultimo è, quindi, quello di avere teorie predittive del comportamento dei sistemi e anche metodi prescrittivi della loro evoluzione (si veda, a tal proposito il paragrafo ... con le conclusioni del contributo).

La **biologia dei sistemi** è un approccio introdotto recentemente da Leroy Hood e basato sulla teoria dei sistemi per studiare fenomeni biologici. Inizialmente la biologia basava la sua ricerca su un approccio riduzionistico in cui i sistemi venivano scomposti nei loro componenti elementari, si studiavano i singoli componenti per acquisire nuova conoscenza per poi cercare di ricombinarli insieme e avere conoscenza sull'intero sistema. Questo approccio è fallito a causa della enorme complessità dei sistemi biologici e quindi l'incapacità di dominare intellettualmente il processo di ricombinazione. L'idea alla base della biologia dei sistemi è quella di trasformare la biologia in una scienza di scoperta in cui si individua un sistema e se ne studiano le caratteristiche. Anche se il passaggio paradigmatico ai sistemi è interessante e utile di per se, avendo ormai a disposizione l'informazione completa sulle sue potenzialità fornita dai risultati del progetto genoma, l'impatto di un tale approaccio ha portata enorme. Potremmo sintetizzare dicendo che la biologia sta passando dalla produzione della conoscenza all'organizzazione della conoscenza acquisita. Ovviamente dopo aver organizzato il materiale disponibile si dovrà procedere alternando fasi di produzione e fasi di organizzazione come avviene per tutte le scienze sperimentali.
L'obiettivo della biologia dei sistemi si sposa perfettamente con quello della bioinformatica che consideriamo in questo contributo, infatti è prevedere correttamente e modificare il comportamento dei sistemi biologici. Per raggiungere questo obiettivo le strategie della biologia dei sistemi prevedono l'uso di sistematiche perturbazioni genetiche e ambientali dei modelli con un monitoraggio accurato delle risposte globali a questi cambiamenti al livello di geni, proteine, meccanismi di segnalazione e fenotipi. Il monitoraggio deve basarsi non solo su osservazioni qualitative, ma anche quantitative che devono guidare la definizione di strutture dinamiche per la modellazione del comportamento dei sistemi. Su tali modelli si devono poi definire verifiche e controlli iterativi come accade nella definizione di nuovi programmi software per permettere la previsione di nuovi comportamenti.

scientifico se non si fa uso di tecniche strutturate e non ambigue di modellazione e di analisi. Da qui la nascita di una nuova branca di biologia chiamata ***biologia dei sistemi*** (***systems biology***) che ripercorre tappe che altre discipline hanno percorso in passato e fa ricorso alla teoria dei sistemi complessi per rappresentare sistemi biologici. Questa area della biologia è estremamente attiva in questi anni e cerca di lanciare un progetto simile al progetto genoma umano, ma con enfasi sulle funzionalità e interazioni dei componenti basilari del corpo umano.
Tornando all'informatica, negli ultimi anni c'è stata molta attenzione ai sistemi mobili e distribuiti e ciò ha portato alla definizione di semplici calcoli (linguaggi formali dotati di sintassi, definizione della simbologia che utilizzano, e semantica, significato attribuito ai simboli, definite rigorosamente) in grado di rappresentare i possibili comportamenti di tali sistemi in modo non ambiguo. Inoltre, tali calcoli sono dotati di strumenti formali di supporto in grado di effettuare analisi e verifiche di proprietà. Anche in questo campo occorre confrontarsi con l'enorme complessità di sistemi formati da milioni di entità geograficamente disperse in grado di comunicare e cooperare senza avere una completa conoscenza dell'ambiente di esecuzione globale e senza avere completa affidabilità e disponibilità di risorse.
A questo punto, unendo gli sforzi fatti nell'area dei linguaggi di programmazione per modellare i sistemi di calcolo globali (*global computing*) e quelli fatti in biologia per passare a un approccio sistemico nello studio dei fenomeni naturali nasce la controparte dinamica della bioinformatica statica e inizia lo studio

del flusso informativo nei sistemi biologici. In questa breve nota verranno trattati questi aspetti dinamici per tre motivi: sono più nuovi e, dunque, meno conosciuti, consentono di aprire nuove frontiere di ricerca che, invece, sono ormai ben chiare nella bioinformatica statica e possono, infine, consentire avanzamenti nello stato dell'arte sia nella modellazione di sistemi informatici complessi e mobili sia nelle conoscenze biologiche fornendo predizioni di comportamento sulla base di simulazioni e modelli analitici.

## 2. RAPPRESENTAZIONE DI SISTEMI BIOLOGICI

Nella biologia moderna è ormai chiara la necessità di integrare tutti i dati provenienti dalle discipline dette "omics" (*genomics, proteomics, metabolomics* ecc.) per ottenere dei modelli di sistemi complessi che possano essere studiati mediante strumenti automatici. Attualmente, seguendo un approccio guidato da ipostesi i formalismi usati per rappresentare i sistemi sono di due tipi: quelli grafici informali utilizzati solitamente nei *data base* pubblici (come per esempio, EcoCyc, KEGG, aMAZE, TransPath, INETRACT, SPAD si vedano per esempio [5, 8, 13] e le Figure 1 e 2) e quelli rigorosi prevalentemente basati su equazioni differenziali. Se, invece, si utilizza un approccio basato su scoperte, i modelli vengono inferiti mediante tecniche statistiche come le reti bayesiane.

I formalismi grafici spesso hanno il difetto di non essere formali e ancor più spesso di essere ambigui; inoltre, non esiste una notazione standard (già le Figure 1 e 2 differiscono nella simbologia grafica e nella loro semantica – cioè il significato attribuito ai simboli utilizzati). I formalismi matematici non sono composizionali e non si riescono a inferire in modo automatico dalle rappresentazioni grafiche. Queste limitazioni impediscono di ottenere dei modelli che permettano di studiare in modo soddisfacente i sistemi biologici sia

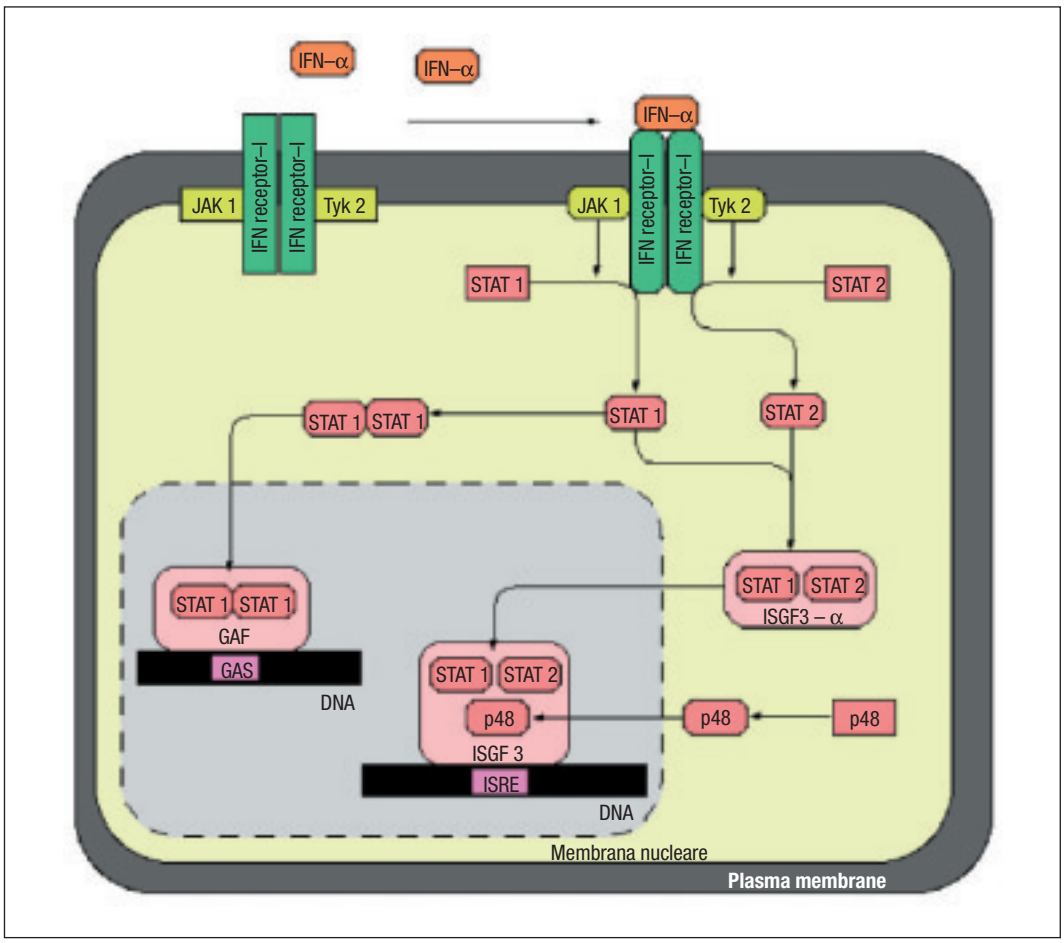



**FIGURA 1**
*Rappresentazione in SPAD*

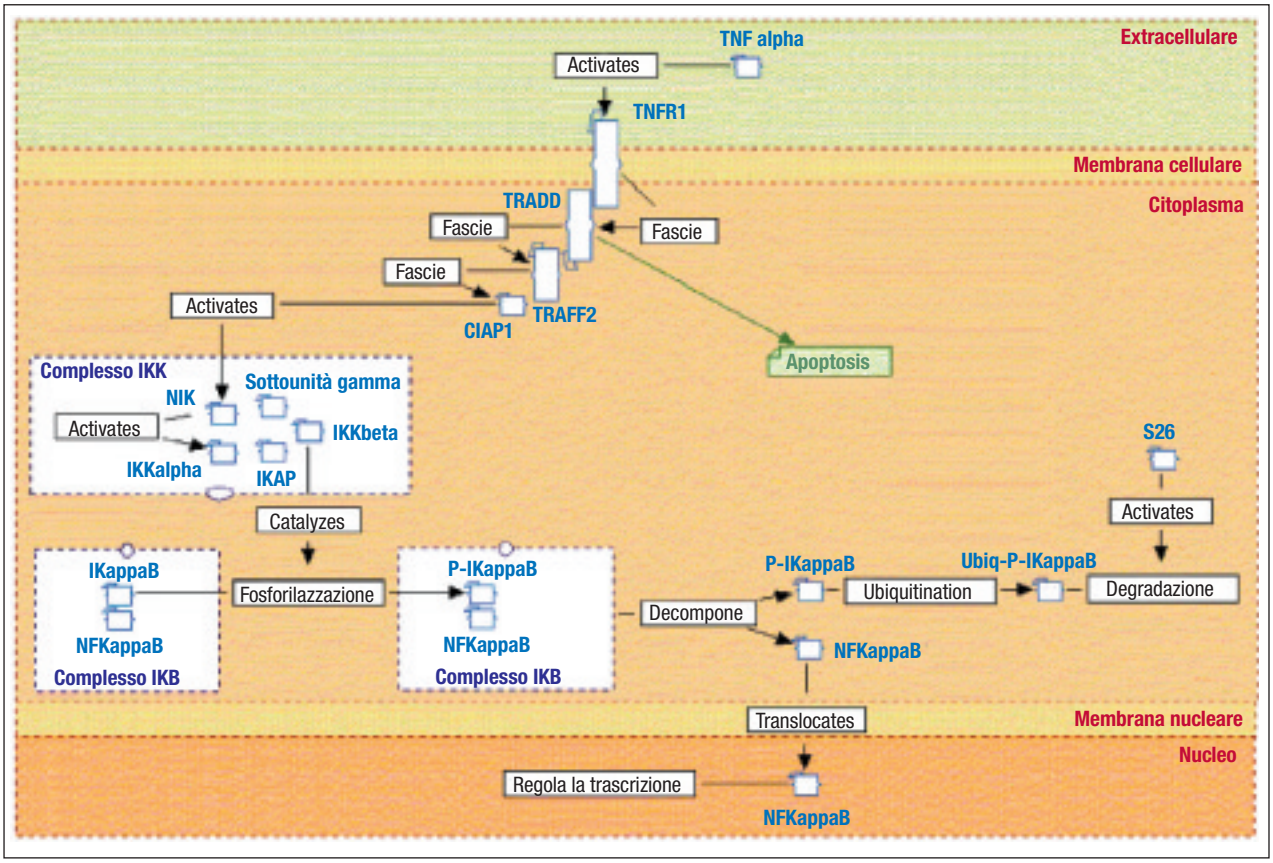



**FIGURA 2**
*Rappresentazione in aMAZE*

in termini di struttura che di funzionalità. Recenti sforzi della comunità biologica mostrano che tali limitazioni possono essere superate se si adottano tecniche di modellazione concettuale che sono da molti anni alla base della teoria dei sistemi, dello sviluppo del *software* o della progettazione di grandi basi di dati. Un aspetto cruciale è la possibilità di combinare nello stesso modello sia aspetti statici che dinamici del sistema. Infine, l'importanza della semantica viene evidenziata per determinare dipendenze e relazioni tra le varie parti del modello in modo non ambiguo. La possibilità di definire una struttura (testo di un programma) e un meccanismo (semantica operazionale) per derivare da questa la descrizione di un comportamento dinamico (sistema di transizione) è una peculiarità della definizione formale dei linguaggi di programmazione. Quindi la teoria dei linguaggi, in particolare di quelli concorrenti e mobili, può fornire un notevole supporto alla biologia dei sistemi dove molti eventi concorrenti che modificano l'evoluzione del sistema nel suo complesso sono sempre presenti.

Le problematiche che si devono risolvere per avere rappresentazioni utili e non ambigue dei sistemi sono la definizione di:

- una rappresentazione grafica standard e facilmente comprensibile ai biologi;
- una rappresentazione formale dei sistemi a cui si possano applicare metodi rigorosi di analisi e simulazione;
- un meccanismo di estrazione automatica delle rappresentazioni formali da quelle grafiche nascondendo i dettagli matematici agli utenti;
- un meccanismo di riflessione dei risultati delle analisi nell'interfaccia grafica.

Un approccio recentissimo per rappresentare sistemi biologici in modo grafico, ma semi-formale è basato su **UML** (***Unified Modeling Language***) che fornisce meccanismi naturali per descrivere i componenti di un sistema e le loro interazioni. Inoltre, la modularità dei diagrammi e delle descrizioni UML ha una corrispondenza immediata con la struttura multilivello dei sistemi naturali (per esempio, un organismo è composto di organi che sono a loro volta composti da

L'acronimo **UML** indica un linguaggio grafico e semi-formale usato per la progettazione di sistemi basati su tecnologie ad oggetti. Esso è sicuramente il linguaggio più diffuso anche in ambiente industriale; è supportato da oltre 70 strumenti automatici di progettazione e da oltre 80 libri descrittivi. La sua diffusione è in continuo aumento e sta coprendo un sempre maggior numero di domini applicativi compresi recentemente i sistemi biologici. Una delle maggiori caratteristiche che lo rende estremamente versatile è la sua estendibilità con meccanismi (profili) previsti già nella definizione del linguaggio. L'utilizzo di profili è stato recentemente adottato da alcuni biologi [12] per definire uno standard di modellazione di sistemi biologici. Inoltre, l'esistenza dello standard XML (*eXtensible Markup Language*) per salvare i modelli UML facilita l'integrazione tra strumenti automatici di tipo diverso. In particolare, molti data base biologici pubblici consentono di esportare informazioni in questo formato.

cellule che sono composte da componenti quali proteine, apparati, nucleo, DNA ecc.). Sono stati già definiti anche alcuni ambienti software per modellare sistemi biologici con UML (si veda www.biouml.org). Lo sforzo più significativo in questa direzione è comunque fornito dalla definizione di un profilo SB-UML specifico per la biologia dei sistemi [12] sottoposto al comitato di standardizzazione OMG.

Utilizzare UML e i suoi meccanismi di estensione per la biologia dei sistemi è una scelta strategica per i seguenti motivi. Il formalismo è grafico e non molto distante da quelli solitamente usati dai biologi ai quali si richiede, quindi, un piccolo sforzo di adeguamento. Anche se UML non ha una semantica formale, è sufficientemente strutturato da consentire la definizione di traduttori automatici in calcoli formali. Importanti istituti come il Pasteur di Parigi lo utilizzano come meccanismo di rappresentazione per il circolo delle informazioni interne e questo favorisce la sua diffusione. Inoltre, essendo uno standard molto diffuso nell'area IT (*Information Technology*) corredato da molti strumenti automatici, dovrebbe essere ridotta la fase di *start-up* per la produzione di strumenti mirati al dominino biologico.

Adesso si descriverà brevemente come le **algebre di processo** possono rappresentare i sistemi biologici. I processi biomolecolari sono reti di proteine che interagiscono, ciascuna composta da molte parti strutturali distinte e indipendenti chiamate "domini". Le interazioni binarie tra domini dipendono dalla complementarietà strutturale e chimica di particolari porzioni delle proteine. L'interazione tra proteine causa a sua volta variazioni biochimiche dei domini che influenzano le future interazioni dei componenti coinvolti. Inoltre, l'interazione tra proteine guida direttamente il funzionamento delle cellule, e le modifiche delle proprietà biochimiche delle proteine sono, quindi, il meccanismo principale che guida molte funzionalità cellulari. Queste caratteristiche corrispondono piuttosto strettamente a quelle dei sistemi distribuiti in cui la topologia di interconnessione delle varie componenti può variare dinamicamente cambiando così le potenziali interazioni future.

Per avere un parallelo più dettagliato tra sistemi biologici e algebre di processo si possono considerare le molecole che interagiscono come processi concorrenti e la complementarietà delle caratteristiche biochimiche come coppie di operazioni complementari (*send* e *receive*) sullo stesso canale di comunicazione. La modifica successiva all'interazione biologica è modellata consentendo la comunicazione di canali che, quindi, alterano la struttura topologica della rete di interconnessione. Infatti, se un certo processo riceve un nuovo nome di canale, da quel momento in poi lo può utilizzare per comunicare con tutti gli altri processi che lo conoscono. Al contrario, se un certo processo consuma il nome di un canale per effettuare su di lui una comunicazione, non potrà poi più comunicare con i processi che conoscono quel canale fino a che non acquisisce nuovamente il nome. Tecnicamente, il comportamento dinamico dei sistemi biologici viene formalmente definito dalla semantica operazionale dei calcoli. In letteratura sono stati proposti recentemente numerosi calcoli per rappresentare sistemi biologici (si ricordano tra questi Biochemical $\pi$-calculus [10], BioAmbients [11], Core Molecular Biology [2], Brane calculus [1]). La descrizione accurata degli aspetti quantitativi che guidano i processi molecolari viene inglobata nel parallelo sopra riportato utilizzando **algebre di processo stocastiche** in cui le transizioni sono governate da distri-

Le **algebre di processo** sono dei semplici calcoli introdotti alla fine degli anni settanta da Tony Hoare e Robin Milner per modellare le peculiarità dei sistemi concorrenti in modo rigoroso. Esse comprendono pochi operatori che compongono azioni elementari indicate con lettere minuscole nel seguito e processi indicati invece con lettere maiuscole: sequenzializzazione di azioni e processi ($a.P$), composizione parallela di processi ($P|Q$), composizione non deterministica di processi ($P + Q$), dichiarazione di nomi nuovi (*new a*), operatore di scelta [$x = y$], ricorsione (rec $X. P$). Le azioni sequenziali possono essere di tre tipi: $a!b$ per spedire il nome $b$ sul canale $a$, $a?x$ per ricevere un dato che rimpiazzerà la variabile targa $x$ sul canale $a$, oppure $t$ per rappresentare un'azione interna del sistema non visibile a un osservatore esterno. Lo scopo principale è quello di definire l'interazione e la cooperazione tra processi concorrenti e mobili.
La semantica intuitiva degli operatori elencati sopra è la seguente. L'azione a è la prima azione atomica che il processo $a.P$ può compiere. La ricezione $a?x$ lega le occorrenze della variabile x nel processo prefisso P. In altre parole, un dato sarà ricevuto sul canale a e sostituirà tutte le occorrenze libere della variabile targa $x$ in $P$. Il prefisso di invio $a!x$ invia il nome $x$ sul canale a senza legare le occorrenze di $x$ in $P$. Nel processo (new $x$)$P$, l'operatore di restrizione new crea un nuovo (unico) nome $x$ il cui raggio di azione è $P$. L'operatore di scelta [$x = y$] è soddisfatto se i due nomi sono uguali e consente l'esecuzione del processo che prefigge. Se la scelta non è soddisfatta l'esecuzione si ferma. Nella composizione parallela $P|Q$ i due processi sono eseguiti indipendentemente e possono comunicare se condividono uno stesso nome di canale. In altre parole $a!x.P|a?y.Q$ può comunicare inviando dalla parte sinistra alla parte destra della composizione il nome $x$ sul canale $a$. Il processo risultante dopo la comunicazione sarà $P|Q\{x/y\}$, dove $\{x/y\}$ rappresenta l'operazione di sostituzione del nome $x$ alle occorrenze libere di $y$ nel processo cui è applicata. La somma rappresenta una scelta non deterministica: $P + Q$ si comporterà in modo mutuamente esclusivo o come $P$ o come $Q$. Infine, rec $X.P$ rappresenta la definizione ricorsiva del processo $P$, cioè la possibilità di ripetere l'esecuzione del processo $P$ tante volte quante si vuole.
La semantica formale di questi calcoli è solitamente fornita in modo operazionale sfruttando l'approccio operazionale introdotto da Gordon Plotkin e basato su assiomi e regole di inferenza. Il rigore formale che ne deriva consente di dimostrare proprietà dei programmi senza perdere in intuizione. Il comportamento dinamico dei sistemi rappresentati viene espresso mediante sistemi di transizione che sono essenzialmente dei grafi etichettati orientati. Gli stati rappresentano le configurazioni del sistema e le transizioni le azioni che il sistema può compiere per cambiare configurazione. Le etichette delle transizioni forniscono informazioni sul tipo di azione che esse rappresentano.

**Algebre di processo stocastiche**
Inizialmente le algebre di processo sono state utilizzate solo per descrivere e studiare aspetti qualitativi di sistemi concorrenti e mobili. L'evoluzione della teoria e le prime applicazioni a casi di studio reali hanno subito mostrato il limite di un approccio qualitativo. Per esempio se si vuole progettare un sistema distribuito di una qualche complessità non si può prescindere dalle prestazioni del sistema sin dai primi passi di progettazione. Questo ha fornito la spinta per estendere la teoria delle algebre di processo con informazioni quantitative vedendo la comparsa in letteratura sia di algebre di processo temporali che probabilistiche. Questo primo passo nel quantitativo non è però sufficiente a risolvere tutti i problemi posti dalla progettazione avanzata di sistemi. Il passo decisivo viene fatto da Jane Hillston quando introduce una variante stocastica di una semplice algebra di processo. L'idea di base è quella di arricchire i prefissi sequenziali delle algebre di processo (si veda il riquadro su algebre di processo) con una distribuzione probabilistica: i nuovi prefissi hanno, quindi, la forma $(a,F).P$ dove $a$ è l'azione standard delle algebre di processo e $F$ è la distribuzione probabilistica continua. A questo punto, il supporto a tempo di esecuzione del calcolo viene reso probabilistico introducendo il concetto di gara tra tutte le azioni che sono abilitate per essere eseguite in una data configurazione. L'idea è che tutte le azioni abilitate tentano di eseguire il loro compito, ma solo la più veloce riesce. Un teorema fondamentale delle distribuzioni continue assicura che la probabilità che due azioni abilitate terminino simultaneamente è zero. Ciò rende non ambiguo il meccanismo di scelta della azioni abilitate. A questo punto, il sistema di transizione viene a coincidere, con piccoli aggiustamenti tecnici, con un processo stocastico che può essere studiato per avere misure quantitative del sistema che rappresenta facendo riferimento a tecniche standard. Il vantaggio di questo approccio rispetto, ad esempio, a reti di code è che il passaggio dalla specifica al processo stocastico avviene automaticamente attraverso la semantica del calcolo e quindi può essere dimostrato corretto una volta per tutte.

buzioni probabilistiche. Il calcolo delle distribuzioni da associare alle transizioni si basa sull'osservazione che l'interazione tra due proteine è governata da una costante di interazione determinata empiricamente in base alle affinità biochimiche dei reagenti e dalla concentrazione. L'unico dei calcoli menzionati che può gestire aspetti stocastici è il $\pi$-calcolo nella sua variante stocastica [9, 10] (si veda, a tal proposito il paragrafo 4 sulla simulazione). Per esempio, il meccanismo di trascrizione regolato da un ciclo a *feedback* positivo astrattamente rappresentato dal diagramma biologico in figura 3 è tradotto dal programma in Biochemical $\pi$-calcolo stocastico di tabella 1.
Si conclude questo paragrafo discutendo brevemente i meccanismi di estrazione e riflessione, ipotizzando di avere una rappresentazione grafica UML-like. In informatica sono stati fatti notevoli sforzi per cercare di ottenere modelli formali basati su algebre di processo a partire da rappresentazioni UML [5]. Queste tecniche possono essere usate per ottenere descrizioni formali di sistemi biologici quando questi sono rappresentati mediante UML e, quindi, applicare le tecniche di analisi tipiche delle algebre di processo.

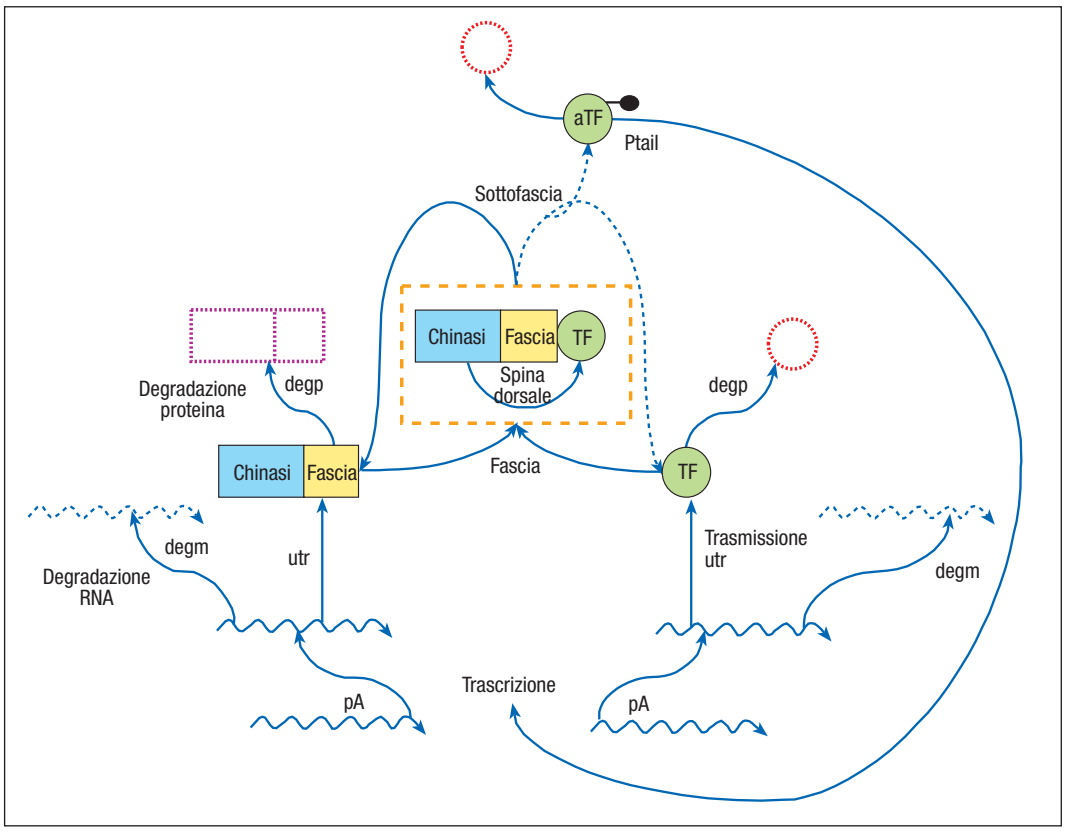



**FIGURA 3**

*Rappresentazione di un meccanismo di trascrizione regolato da un ciclo a feedback positivo*

```
Sys = Gene_A|Gene_TF|Transer|Transl|RNA_Deg|Protein_Deg
Gene_A = (basal(), 4).(Gene_A|RNA_A) + (pA(), 40).(Gene_A|RNA_A)
RNA_A = (utr(), 1).(RNA_A|Protein_A) + (degm(), 1)
Protein_A = (νbb1, bb2, bb3)(Binding_Site|Kinase)
Binding_Site = (bind ⟨bb1, bb2, bb3⟩, 0.1).Bound_Site + (degp(), 0.1).(bb3, ∞)
Bound_Site = (bb1, 10).Binding_Site + (degp(), 0.1).(bb3. ∞).(bb3, ∞)
Kinase = (bb2 ⟨ptail⟩, 10).Kinase + (bb3(), ∞)
Gene_TF = (basal(), 4).(Gene_TF|RNA_TF) + (pA(), 40).(Gene_TF|RNA_TF)
RNA_TF = (utr(), 1).(RNA_TF|Protein_TF) + (degm(), 1)
Protein_TF = (bind (c_bb1, c_bb2, c_bb3), 0.1).Bound_TF + (degp(), 0.1)
Bound_TF = (c_bb1(), 10).Protein_TF + (c_bb3(), ∞) + (c_bb2(tail), 10).
           ((c_bb1(), 10).Active_TF(tail) + (c_bb3(), ∞))
Active_TF(tail) = ( tail, 100).Active_TF(tail) + (degp(), 0.1)
Transer = (basal, 4).Transer + (ptail(), 100).(pA, 40).Transer
Transl = ( utr, 1).Transl
RNA_Deg = (degm, 1).RNA_Deg
Protein_Deg = (degp, 0.1).Protein_Deg
```

**TABELLA 1**

*Rappresentazione in BioSPI del sistema di figura 3*

## 3. ANALISI

Le tecniche di analisi del comportamento dei sistemi specificati mediante algebre di processo si dividono in statiche e dinamiche. Quelle più usate fino ad oggi in ambito biologico sono quelle dinamiche che prevedono la costruzione di un modello del comportamento a partire dalla descrizione (per esempio un sistema di transizione, cioè un grafo orientato in cui i nodi rappresentano gli stati del sistema e le transizioni gli eventi che causano il passaggio di stato – si veda tabella su algebre di processo). Le proprietà che si riescono a studiare con queste tecniche possono essere sia di tipo qualitativo che di tipo quantitativo. Tra le prime si ricordano la causalità tra transizioni o eventi, la località in cui certe transizioni avvengono, la concorrenza di transizioni [3].

Lo studio della relazione di causalità tra transizioni (la prima causa la seconda se è condizione necessaria per la seconda e ne influenza l'esecuzione) consente di determinare su un modello dinamico di una malattia quali sono gli eventi scatenanti e consente anche di tracciare in modo preciso il comportamento di un dato farmaco sui meccanismi di segnalazione della malattia. Da qui il concetto di modello predittivo se attraverso queste analisi si riescono a prevedere nuovi comportamenti biologici validabili attraverso esperimenti di laboratorio.

Anche la località gioca un ruolo essenziale nella modellazione e nell'analisi di sistemi biologici. Infatti, è essenziale sapere la localizzazione di certi componenti per determinare la probabilità o semplicemente la possibilità di una loro interazione. Dato un generico modello comportamentale di un sistema biologico, la località può essere utilizzata per ridurne la dimensione eliminando i comportamenti derivanti da interazioni tra componenti che non sono sufficientemente vicini o che non possono proprio entrare in contatto.

Altro esempio importante di proprietà da considerare è la concorrenza, ossia la possibilità per due o più transizioni di avvenire contemporaneamente. Questo consente di studiare fenomeni come ad esempio il *rolling* dei globuli bianchi in corrispondenza di tessuti infiammati nel suo complesso e non semplicemente studiando il comportamento di un singolo globulo. I risultati che si ottengono nei due casi sono abbastanza difformi e hanno ricadute diverse sull'evolvere dell'infiammazione.

Passando alle proprietà quantitative si ricorda come i sistemi di transizione possano, con piccole manipolazioni, essere interpretati come processi stocastici quando gli archi sono etichettati mediante distribuzioni probabilistiche (tipicamente esponenziali in tempo continuo). Queste tecniche sono state adottate da molti anni nel campo della valutazione delle prestazioni dei sistemi distribuiti, originando quelle che sono chiamate algebre di processo stocastiche.

Analizzare le proprietà di sistemi individuali non è abbastanza. Ulteriori conoscenze sulle funzionalità e possibili evoluzioni di reti molecolari possono essere acquisite confrontando i sistemi relativamente alle condizioni biologiche in cui operano, ai tipi di cellule e organismi che li compongono. La biologia computazionale (disciplina che sviluppa algoritmi efficienti per manipolare grandi quantità di dati, ad esempio al fine di confrontare due o più o sequenze di DNA, per ricostruire sequenze di nucleotidi data una conoscenza frammentaria delle sequenze o per generare alberi evolutivi a partire da un insieme di genotipi) ha ottenuto importanti risultati confrontando le sequenze e le strutture di singole molecole. In modo analogo, si possono usare strumenti messi a disposizione dalla teoria della concorrenza come le equivalenze basate sul concetto di bisimulazione per confrontare il comportamento dinamico di intere reti molecolari. Quello che è possibile, quindi, definire è una misura di omologia dei processi molecolari derivata dallo studio dei modelli. Le ricadute di una tale applicazione sono significative sia in campo informatico che biologico-medico. Dal punto di vista computazionale si può trarre ispirazione per nuove nozioni di equivalenza in quanto la nozione di omologia biologica è molto più complessa di quella di bisimulazione. Sul lato biologico lo studio comparativo di condizioni patologiche (per esempio, confronto del comportamento di un tessuto normale e di un tessuto tumorale) può con-

sentire di tracciare a ritroso importanti passi che stanno alla base dell'attivazione della malattia. Questo è tanto più possibile quanto più si riescono a compenetrare nella definizione delle equivalenze anche le nozioni di causalità e località.
Il problema principale delle tecniche dinamiche sopra descritte è dato dalla dimensione del sistema di transizione che è esponenziale rispetto alla descrizione testuale in algebre di processo. La conseguenza immediata è che data la grande dimensione dei sistemi biologici è impensabile avere algoritmi che possano esaminare in modo esaustivo lo spazio degli stati. Le soluzioni proposte sono prevalentemente orientate a ridurre la complessità computazionale del problema informatico a scapito della precisione dei risultati che si possono ottenere. Si descriverà qui un possibile utilizzo di tecniche di analisi statica e si rimanda al prossimo paragrafo la discussione delle tecniche di simulazione.
Le tecniche di analisi statica sono state introdotte originariamente per effettuare ottimizzazioni di compilatori, ma oggi le loro aree di applicazione sono molto più ampie. L'idea alla base di queste tecniche è la possibilità di estrarre informazioni complesse sul comportamento dinamico di sistemi semplicemente guardando alla loro descrizione testuale: il programma. Questo vuol dire che non è necessario costruire il modello del comportamento dinamico (il sistema di transizione) e, quindi, viene meno il vincolo dato dall'esponenzialità della rappresentazione. Al contrario dell'analisi dinamica, ogni analisi statica deve essere definita in relazione a una particolare proprietà che si vuole studiare e al particolare linguaggio di specifica che si intende usare. Da qui la necessità di avere un formalismo unico per descrivere molti aspetti diversi dei sistemi biologici al fine di limitare il numero di analisi che si devono definire.
Le informazioni che si estraggono mediante analisi statica del testo del programma sono corrette rispetto al comportamento dinamico del programma, ma non è possibile ottenere informazioni esatte. Il compromesso dell'analisi statica per avere algoritmi efficienti è a scapito della precisione delle informazioni ricavate. Da qui il concetto di approssimazione. Ci possono essere sia approssimazioni per eccesso che approssimazioni per difetto. Nel primo caso si ottiene uno spazio delle soluzioni del problema che contiene strettamente le soluzioni esatte e, quindi, si può affermare con certezza solo ciò che non potrà mai accadere. Nel caso di approssimazioni per difetto si ottiene uno spazio delle soluzioni del problema che è strettamente contenuto nello spazio delle soluzioni esatte. In questo caso si può dire con certezza solo ciò che accadrà sicuramente. La situazione inaccettabile per una approssimazione è quando lo spazio delle soluzioni calcolato contiene solo un sottoinsieme delle soluzioni esatte perché in questo caso non si ha alcun controllo sulla correttezza dei risultati ottenuti.
Anche se l'analisi statica è stata introdotta per studiare proprietà completamente diverse, può contribuire in modo significativo allo sviluppo della bioinformatica. Infatti le tecniche di approssimazione individuate consentono di studiare sistemi almeno un ordine di grandezza più grandi di quelli studiati mediante tecniche dinamiche. Le proprietà che possono essere esaminate riguardano la localizzazione di componenti all'interno di strutture più complesse, le loro possibili interazioni e migrazioni (per esempio la traslocazione nel nucleo di una cellula e la conseguente trascrizione), la determinazione di cicli a *feedback* positivo o negativo all'interno di grandi reti di segnalazione [7].

## 4. SIMULAZIONE

Come accennato nella precedente sezione, anche le tecniche di simulazione consentono di evitare la costruzione di un intero modello del comportamento dinamico, eliminando il problema dell'esponenzialità delle rappresentazioni. L'idea alla base della simulazione è quella di eseguire il programma che rappresenta il sistema biologico scegliendo una tra tutte le possibili esecuzioni (in termini di modello dinamico vuol dire scegliere un cammino sul sistema di transizione). Ripetendo un numero molto elevato di volte l'esecuzione si ottiene una descrizione "media" del comportamento dinamico del sistema considerato.
I principali tentativi di modellare il comportamento dinamico dei sistemi biologici so-

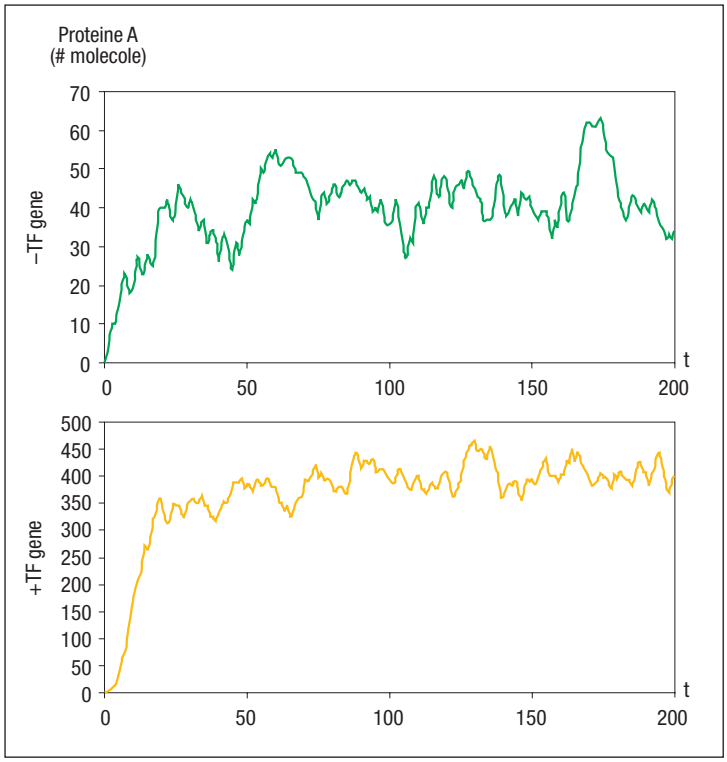



**FIGURA 4**
*Risultati della simulazione mediante BioSPI del sistema rappresentato graficamente in figura 3 e in stocastico p-calcolo in tabella 1*

no basati su equazioni differenziali ordinarie o stocastiche, su metodi di simulazione discreta che possono rifarsi alle tecniche Monte-Carlo, reti bayesiane [4]. Ciascuno degli approcci menzionati è in grado di catturare alcuni degli aspetti specifici dei meccanismi di segnalazione cellulare, ma nessuno è in grado di integrare la dinamica con gli aspetti molecolari e biochimici. Queste limitazioni possono essere superate mediante l'utilizzo di algebre di processo come si è già visto nelle precedenti sezioni.

Il primo ambiente di simulazione basato sulla realizzazione di un supporto a tempo di esecuzione probabilistico per il $\pi$-calcolo, implementando quindi una variante dello stocastico $\pi$i-calcolo [9] è BioSPI [10]. La realizzazione è basata su Flat Concurrent Prolog e supporta completamente comunicazioni e scelte non deterministiche (rese poi probabilistiche dal supporto a tempo di esecuzione che implementa l'algoritmo di Gillespie).

Il sistema realizzato consente di specificare le quantità iniziali dei vari componenti di cui si vogliono studiare le potenziali interazioni e i tassi probabilistici con cui avvengono le comunicazioni sui vari canali che compaiono nella specifica del sistema (si veda per esempio Tabella 1). Il meccanismo di simulazione permette poi di monitorare come le concentrazioni e i prodotti delle interazioni variano al variare del tempo (anche la scala temporale può essere variata scegliendo quella più adeguata al fenomeno considerato). Per esempio il risultato della simulazione del programma di tabella 1 fornisce il risultato riportato in figura 4.

## 5. CONCLUSIONI E SVILUPPI FUTURI

In questo contributo sono stati esaminati gli aspetti della bioinformatica principalmente legati alla descrizione dei comportamenti dinamici dei sistemi biologici complessi. Gli obiettivi scientifici principali che sono stati considerati riguardano:

- la rappresentazione dei sistemi che sia amichevole per i biologi, ma che consenta al tempo stesso di derivare in modo automatico modelli formali per lo studio rigoroso del comportamento biologico;
- l'analisi qualitativa e quantitativa di proprietà dei sistemi e possibile definizione di una omologia di processi molecolari basata su nozioni di equivalenze comportamentali definite i teoria della concorrenza;
- la simulazione del comportamento di sistemi basata su implementazioni di supporti a tempo di esecuzione probabilistici per le algebre di processo;
- la necessità di costituire un linguaggio comune tra informatici e biologi per la forte interdisciplinarietà della bioinformatica.

Le tecniche di analisi e di simulazione non sono in alternativa, ma servono entrambe per avere una visione complessiva del sistema studiato quanto più precisa possibile. Infatti, mentre con le simulazioni si riescono a studiare quantitativamente le variazioni percentuali delle sostanze coinvolte in certi fenomeni biologici esse sono inerentemente limitate nella capacità di fornire risposte generali su proprietà intrinseche di sistemi. Le tecniche di analisi analitica basate sui metodi formali sono, invece, adatte a fornire risposte generali a domande del tipo "un segnale può passare attraverso una certa

molecola sotto certe condizioni? Sotto ogni condizione?" oppure " se modifichiamo il comportamento di una certa molecola, cosa accadrà a un'altra data molecola? O al sistema nel suo complesso?"
Uno dei motivi principali del graduale abbandono dell'approccio riduzionista (si veda il riquadro sulla biologia dei sistemi) nella biologia attuale è dovuto all'impossibilità di coniugare la conoscenza acquisita sulle componenti minimali dei sistemi per ottenere un modello dell'insieme. Ciò è dovuto alla enorme complessità dei sistemi biologici. Un vantaggio che deriva dall'uso delle algebre di processo (e dalla teoria dei linguaggi in generale) è la loro **natura** inerentemente **composizionale** che fornisce regole precise e non ambigue di composizione di oggetti elementari per costruire oggetti più complessi.
In conclusione di questo paragrafo si riportano alcune considerazioni sulle possibili evoluzioni della bioinformatica. L'obiettivo a lungo termine della bioinformatica (o almeno degli aspetti di questa disciplina maggiormente discussi in questo contributo) sono quelli di fornire tecniche sia *predittive* che *prescrittive* del comportamento dei sistemi biologici complessi. Le tecniche predittive sono in grado di prevedere i comportamenti di un sistema e, quindi, sono semplicemente descrittive, mentre le tecniche prescrittive dovrebbero essere in grado di imporre un determinato comportamento (o classe di comportamenti) ai sistemi, essendo, quindi, invasive. Si discutono adesso brevemente le due tipologie di tecniche.
Una tecnica predittiva si basa sulla bontà del modello del sistema biologico e mediante analisi delle proprietà del modello determina possibili evoluzioni. Questa strategia può essere usata dai ricercatori per dimostrare in laboratorio certi comportamenti ancora non noti oppure per prevedere le reazioni di un farmaco in presenza di determinate malattie. Essenziale per ottenere buoni risultati è la fase di validazione dei modelli che si utilizzano e, da qui, la grande attività attuale nella modellazione di sistemi biologici reali mediante algebre di processo al fine di ottenere dal modello almeno tutti i comportamenti noti del sistema considerato mediante applicazione di tecniche di analisi.

Il termine **"composizionalità"** in teoria dei linguaggi indica la possibilità di definire la semantica di un costrutto in termini della semantica dei suoi componenti. Questa è chiaramente una proprietà fondamentale per poter definire in modo finito e chiaro la semantica dei linguaggi di programmazione. Infatti se la definizione della semantica non fosse composizionale dovremmo elencare la semantica di tutti i possibili programmi esprimibili in un dato linguaggio e per ogni linguaggio interessante questi sono infiniti. Parlando di sistemi (biologici) la composizionalità è la possibilità di definire un modello mediante integrazione (composizione) dei modelli dei sotto-sistemi che lo costituiscono. Quando le regole di composizione sono chiare per un certo formalismo e dominio applicativo, la composizionalità è anche una metodologia di progettazione e sviluppo che consente di esaminare e determinare soluzioni per problemi semplici che poi verranno composte per risolvere problemi più complessi.

Le tecniche prescrittive definiscono, invece, algoritmi che vengono eseguiti su *hardware* vivente. Considerando che ogni singola cellula ha approssimativamente 1 MIPS di potenza di calcolo e 1 MEGA di memoria, le potenzialità dei computer viventi sono estremamente interessanti. La possibilità di usare le cellule per eseguire algoritmi avrebbe una ricaduta immensa anche in campo biologico-medico. Per esempio, riprogrammando le cellule che esibiscono comportamenti anomali si potrebbero trovare cure efficaci per tutta la classe delle malattie auto-immuni come la sclerosi multipla oppure per i tumori. Su questa strada si stanno muovendo numerosi importanti gruppi di ricerca cercando di definire un modello completo del funzionamento della cellula. Questo è sicuramente il primo passo per ipotizzare poi metodologie in grado di controllare ed eventualmente modificare il comportamento delle cellule.
Concludendo si può certamente affermare che gli aspetti dinamici dei sistemi biologici e le tecniche informatiche per dominarne la complessità sono un campo di ricerca agli esordi e che probabilmente dominerà la scena bioinformatica dei prossimi anni con attività altamente interdisciplinari. Infatti, per validare i modelli di comportamento che portino a tecniche predittive è assolutamente necessario interagire con biologi e per poter comprendere completamente i risultati delle analisi i biologi devono poter comprendere ciò che gli informatici hanno fatto. Da qui la necessità di creare una co-

munità scientifica internazionale che sia a cavallo tra le discipline delle scienze della vita e quelle del settore dell'informazione. Recentemente la branca della biologia dei sistemi sta cercando di percorrere questa strada reclutando fisici, matematici e informatici all'interno dei loro istituti di ricerca. Analogo sforzo dovrebbe essere fatto nei centri di ricerca bioinformatici che nascono dai dipartimenti di informatica. Occorre tener presente che come sempre quando si creano nuove comunità scientifiche da aggregazioni di persone provenienti da aree disciplinari diverse è essenziale stabilire immediatamente un criterio di pariteticità tra le varie anime del gruppo. Per far questo è necessario individuare con chiarezza quelle che sono le necessità e le aspettative dei biologi e quali quelle degli informatici assumendo che nessuna delle due scienze è al servizio dell'altra. Su questa base paritaria sarà poi possibile attivare progetti di grande respiro e creare una robusta comunità scientifica.

## Bibliografia


[1] Cardelli L.: *Brane calculus*. Rapporto tecnico Microsoft Research Cambridge, 2003.

[2] Danos V., Laneve C.: *Graphs for Core molecular biology*. Proceedings of CMSB03, LNCS 2602, Springer-Verlag, 2003.

[3] Degano P., Priami C.: Noninterleaving semantics of mobile processes. *Theoretical Computer Science*, Vol. 216, n. 1-2, 1999, p. 237-270.

[4] De Jong H.: Modeling and Simulation of Genetic Regulatory Systems. *A Literature Review. Journal of Computational Biology*, Vol. 9, n. 1, 2002, p. 67-103.

[5] Karp P.D., Krummenacker M., Paley S., Wagg J.: Integrated pathway/genome databases and their role in drug discovery. *Trends in Biotechnology*, Vol. 17, n. 7, 1999, p. 275-281.

[6] Koremblat K., Priami C.: *Towards extracting _-calculus from UML sequence and state diagrams*. In Compositional verification of UML models 03, apparirà su ENTCS 2003.

[7] Nielson F., Nielson H.R., Priami C., Scuch da Rosa D.: *Control Flow Analysis of BioAmbients*. Bioconcur 2003, Apparirà su ENTCS.

[8] Ogata H., Goto S., K. Sato, Fujibuchi W., Bono H., Kanehisa Kegg M.: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, Vol. 27, n. 1, 2000, p. 29–34.

[9] Priami C.: Stochastic $\pi$-calculus. *The Computer Journal*, Vol. 38, n. 6, 1995, p. 578-589.

[10] Priami C., Regev, Shapiro E., Silverman W.: Application of a stochastic passing-name calculus to representation and simulation of molecular processes. *Information Processing Letters*, Vol. 80, 2001, p. 25-31.

[11] Regev A., Panina E.M., Silverman W., Cardelli L., Shapiro E.: *BioAmbients: an abstraction for biological compartments*. Apparirà su Theoretical Computer Science.

[12] Roux-Rouquie M., Caritey N., Gaubert L., Rosenthal-Sabroux C.: *Using the Unified Modeling Language (UML) to guide Systemic description of Biological Processes and Systems*. Apparirà su BioSystems.

[13] Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhauser R., Pruss M., Schacherer F., Thiele S., Urbach S.: The transfac system on gene expression regulation. *Nucleic Acids Research*, Vol. 29, n. 1, 2001, p. 281–283.



Corrado Priami ha ricevuto laurea e dottorato di ricerca in informatica all'Università di Pisa, è stato ricercatore all'Ecole Normale Superieur di Parigi e poi all'Università di Verona dove è diventato professore associato di Informatica. Attualmente è professore straordinario di informatica all'Università di Trento, Dipartimento di Informatica e Tlc, dove guida il gruppo di bioinformatica ed è responsabile dei corsi di laurea e laurea specialistica in Informatica. Rappresenta l'Università di Trento nel consiglio di amministrazione della Trento School of Management ed è delegato di ateneo per i progetti europei nel sesto programma quadro. I suoi interessi di ricerca coprono, oltre la bioinformatica, i sistemi distribuiti e mobili e i linguaggi di programmazione in senso lato. Coordina un progetto europeo nell'area del global computing e un progetto nazionale sulle tematiche riportate nell'articolo. Partecipa, inoltre, a numerosi altri progetti sia nazionali che internazionali e ha numerose collaborazioni con industrie nel settore delle comunicazioni mobili e delle biotecnologie. Opera come revisore di progetti europei e nazionali nelle sue aree di interesse. Ha pubblicato oltre 80 articoli su riviste e convegni internazionali ed è autore e curatore di alcuni libri sulle tematiche sopra esposte. Ha partecipato e partecipa, anche come presidente, a numerosi comitati di programma di eventi internazionali prevalentemente nell'area della bioinformatica. Co-dirige una scuola di dottorato aperta a giovani ricercatori informatici e biologi per la costruzione di un gruppo scientifico interdisciplinare.
priami@dit.unitn.it

# DENTRO LA SCATOLA

## Rubrica a cura di

**Fabio A. Schreiber**

Il Consiglio Scientifico della rivista ha pensato di attuare un'iniziativa culturalmente utile presentando in ogni numero di Mondo Digitale un argomento fondante per l'Informatica e le sue applicazioni; in tal modo, anche il lettore curioso, ma frettoloso, potrà rendersi conto di che cosa sta "dentro la scatola". È infatti diffusa la sensazione che lo sviluppo formidabile assunto dal settore e di conseguenza il grande numero di persone di diverse estrazioni culturali che - a vario titolo - si occupano dei calcolatori elettronici e del loro mondo, abbiano nascosto dietro una cortina di nebbia i concetti basilari che lo hanno reso possibile. La realizzazione degli articoli è affidata ad autori che uniscono una grande autorevolezza scientifica e professionale a una notevole capacità divulgativa. Il primo di essi, pubblicato in questo numero, esce a firma del Prof. Luigi Dadda, uno dei Padri Fondatori dell'Informatica italiana e tutt'ora attivo ricercatore presso il Politecnico di Milano. Il Prof. Dadda è stato anche uno dei fondatori dell'AICA e per lunghi anni direttore responsabile di "Rivista di Informatica", il suo organo ufficiale.

# Fondamenti dell'aritmetica digitale: i codici numerici

**Luigi Dadda**

## 1. Introduzione

Questo primo articolo sui fondamenti dell'aritmetica digitale tratterà dei modi di rappresentazione dei numeri nella forma adatta ai calcolatori digitali. È facile che sorga nel lettore la curiosità sulla storia dei numeri stessi. Per soddisfare, almeno in parte, tale possibile desiderio, sono riportati nella bibliografia gli indirizzi di una seria scelta di siti Web.
Se si fa riferimento specificamente alla rappresentazione dei numeri all'interno dei calcolatori un punto fondamentale di tale storia é segnato dalla adozione generalizzata del sistema binario, per la più facile realizzabilità dei circuiti logici e di memoria [10].

## 2. Sistemi numerici posizionali

### 2.1. Numeri interi

Un numero intero $X$ può essere scritto come una stringa di simboli (*cifre* o *digits*):

$$X = x_{n-1}\, x_{n-2} \ldots\ldots x_1\, x_0$$

scelti da un insieme $(0, 1, \ldots, b^{-1})$; $b$ è la base o radice. Ciascuna cifra, $x_i$, ha un "peso" $b^i$, cosicchè:

$$X = x_{n-1} \cdot b^{n-1} + x_{n-2} \cdot b^{n-2} + \ldots\ldots\ldots + x_1 \cdot b^1 + x_0 \cdot b_0 \qquad (1)$$

Poiché il valore o peso di una cifra $x_i$ dipende dalla sua posizione nel numero, i sistemi numerici con base sono anche detti "posizionali". I valori della base $b$ più comunemente utilizzati nei calcolatori elettronici sono: 10 (base decimale), 2 (base binaria), 8 (base ottale), 16 (base esadecinale). Per $b = 2$ per e $b = 8$ le cifre sono normalmente rappresentate con gli stessi simboli usati nel sistema decimale (rispettivamente 0, 1 e 0, 1, 2, 3, 4, 5, 6, 7); per $b = 16$ si adottano i simboli 0,....,9 con l'aggiunta di altri sei: A, B, C, D, E, F.
Si noti che, poiché uno stesso simbolo può essere usato per rappresentare numeri in basi diverse, se si usano basi diverse è necessario indicare la base di ogni numero. Per esempio: $145_{10}$, $101_2$, $176_8$, $1D6_{16}$.
È, infatti, importante notare che il numero 101 potrebbe non essere in base 2 ma in una qualsiasi delle altre basi sopra citate (dieci, otto, sedici) perché i simboli 0 e 1 appartengono a tutte e quattro le basi. La combinazione 101 potrebbe, quindi, rappresentare valori del tutto diversi:

$101_2 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 4 + 0 + 1 = 5_{10}$
$101_{10} = 1 \cdot 10^2 + 0 \cdot 10^1 + 1 \cdot 10^0 = 100 + 0 + 1 = 101_{10}$
$101_8 = 1 \cdot 8^2 + 0 \cdot 8^1 + 1 \cdot 8^0 = 64 + 0 + 1 = 65_{10}$
$101_{16} = 1 \cdot 16^2 + 0 \cdot 16^1 + 1 \cdot 16^0 = 256 + 1 = 257_{10}$

Tutti i calcoli sopra riportati sono stati eseguiti con numeri nella base dieci. Essi sono anche esempi di conversione di numeri interi nelle basi 2, 8, 16 nella rappresentazione equivalente in base dieci.

### 2.2. Numeri frazionari

Per rappresentare numeri frazionari si usa la notazione posizionale con pesi costituiti da potenze negative intere della base:

$$X = 0 \cdot x_{-1} x_{-2} \ldots x_{-n} = x_{-1} \cdot b^{-1} + x_{-2} \cdot b^{-2} + \ldots + x_{-n} \cdot b^{-n}$$

Un numero binario frazionario sarà facilmente convertibile nell'equivalente decimale disponendo dei valori:

$b^{-1} = 0.5_{10}$, $b^{-2} = 0.25_{10}$, $b^{-3} = 0.125_{10}$, $b^{-4} = 0.0625_{10}$, $b^{-5} = 0.03125_{10}$ ecc.

Per esempio:

$0.11001_2 = 0.5 + 0.25 + 0 + 0 + 0.03125 = 0.78125_{10}$

### 2.3. Numeri misti

Sono costituiti da parte intera e parte frazionaria. La loro conversione si ottiene convertendo separatamente le rispettive parti intere e quelle frazionarie.
La conversione in base 2 dei numeri in base 8 e in base 16 si può molto facilmente eseguire esprimendo ciascuna cifra ottale (esadecimale) con il gruppo di 3 (4) bit, che rappresentano la cifra nel sistema binario per esempio:

$325.47_8 = 011\ 010\ 101.100\ 111_2$
$325.47_{16} = 0011\ 0010\ 0101.0100\ 0111_2$
$3B8.45D_{16} = 0011\ 1011\ 0100.0100\ 0101\ 1101_2$

## 3. Conversione di numeri decimali negli equivalenti binari

### 3.1. Decimali interi

Un qualsiasi numero binario intero può esprimersi nella seguente forma:

$$X = \{[(x_{n-1} \cdot 2 + x_{n-2})\ 2 + x_{n-3}]\ 2 + \ldots\ldots\ldots + x_1\}\ 2 + x_0$$

Per esempio:

$$X = 110101_2 = 1 \cdot 2^5 + 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = \{[((1 \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 1] \cdot 2 + 0\} \cdot 2 + 1$$

Da ciò può derivarsi la seguente *regola per la conversione* di un intero decimale nella forma binaria:
**1.** Si divide per due l'intero: il resto vale $x_0$, la *prima* cifra binaria (di peso 1).
**2.** Si divide il quoziente per due: il resto vale $x_1$, la *seconda* cifra binaria.
**3.** Si ripete il passo precedente, ottenendo via via le altre cifre della rappresentazione, fino ad ottenere un quoziente nullo.

*Esempio*:

$X = 54_{10}$
54:2 = 27 con resto 0 = $x_0$
27:2 = 13 con resto 1 = $x_1$
13:2 = 6 con resto1 = $x_2$
6:2 = 3 con resto 0 = $x_3$
3:2 = 1 con resto 1 = $x_4$
1:2 = 0 con resto 1 = $x_5$
e quindi: $X = 110110_2$

### 3.2. Decimali frazionari

Con procedimento analogo al precedente, si ottiene la seguente regola di conversione:
**1.** Si moltiplica per due il frazionario dato: la parte intera del prodotto, che può essere solo *0* oppure *1*, vale $x_{-1}$, la cifra binaria più significativa (di peso $2^{-1}$).
**2.** Si applica alla parte frazionaria del precedente raddoppio lo stesso precedente procedimento, fino ad ottenere una parte frazionaria nulla o ad individuare un numero periodico (si osservi che, in questo caso, un numero decimale frazionario con un numero finito di cifre significative può generare un numero binario frazionario con un numero di cifre infinito).

*Esempi*:

$X = 0.75_{10}$
$0.75 \cdot 2 = 1.50$; parte intera 1 = $x_{-1}$
$0.50 \cdot 2 = 1.00$; parte intera 1 = $x_{-2}$
0.00 e quindi: $X = 0.11_2$

$X = 0.1_{10}$

Si ottiene:

$X = 0.0001100110011001100110011...._2 = 0.\underline{0011}_2$ cioè un numero periodico.

## 4. Numeri binari negativi

Il modo più facile per distinguere i numeri positivi da quelli negativi consiste nell'aggiungere a essi un apposito simbolo, il segno. Per esso basta un bit e una convenzione, per esempio 0 per +, 1 per –: è il metodo di rappresentazione detto del *segno e grandezza* (*sign-and-magnitude*) e in esso occorre disporre di due algoritmi distinti per la somma e per la sottrazione.
Un altro metodo di rappresentazione dei numeri negativi è, invece, basato sui *complementi*. Nel seguito, essi vengono esemplificati facendo riferimento a numeri binari di 4 bit (Tabella 1).
La tabella 1 contiene nella prima colonna le 16 configurazioni di 4 bit. La seconda colonna contiene i valori equivalenti, in base 10, delle predette configurazioni secondo la rappresentazione detta dei *complementi* a 2, *o*, *più in generale, complementi alla base*: il bit all'estrema sinistra ha il ruolo del segno. La terza colonna della tabella 1 mostra i valori delle configurazioni con primo bit di valore 1 nel caso di rappresentazione con i *complementi ad* 1, *o*, *più in generale, complementi alla* (*base* – 1). Il bit più a sinistra rappresenta ancora il segno e vi è anche uno "0" negativo.
L'utilità di ricorrere ai complementi deriva da due tipi di considerazioni: il primo è la facilità di ottenere il complemento a 1 o a 2 di un numero dato mediante dispositivi elettronici; il secondo sta nel fatto che, con l'utilizzo della rappresentazione in complemento dei numeri negativi, le operazioni di somma e sottrazione possono essere eseguite con lo stesso circuito. Tali argomenti verranno trattati in successivi articoli, dedicati alle operazioni aritmetiche e alle macchine per realizzarle.
Qui di seguito si vuol dare solamente un breve esempio di operazioni di somma/sottrazione che utilizzano i concetti esposti; si tenga presente che l'algoritmo di somma è analogo a quello usato nel caso decimale e che, in binario, valgono le seguenti regole: 0 + 0 = 0, 0 + 1 = 1, 1 + 1 = 0 con un riporto di 1. Nella rappresentazione in complemento a 2 il riporto viene trascurato:

| | Complementi a 2 o complementi alla base | Complementi ad 1 o complementi alla (base –1) |
|---|---|---|
| 0111 | +7 | |
| 0110 | +6 | |
| 0101 | +5 | |
| 0100 | +4 | |
| 0011 | +3 | |
| 0010 | +2 | |
| 0001 | +1 | |
| 0000 | 0 | |
| 1111 | –1 | –0 |
| 1110 | –2 | –1 |
| 1101 | –3 | -2 |
| 1100 | –4 | –3 |
| 1011 | –5 | –4 |
| 1010 | –6 | –5 |
| 1001 | –7 | –6 |
| 1000 | –8 | –7 |

**TABELLA 1**
*Esemplificazione di numeri negativi*

| | | | |
|---|---|---|---|
| 0111 (7) | 1101 (13) | 0101 (+5) | 1010 (–6) |
| 0010 (2) | 0011 (3) | 1110 (–2) | 0010 (+2) |
| 1001 (9) | 10000 (16) | (1)0011 (+3) | 1100 (–4) |
| somme di numeri positivi | | somme di numeri in complemento a 2 | |

## 5. Numeri decimali con cifre codificate in binario

I numeri binari sono certamente i più utilizzati nei calcoli scientifico-tecnici, in quanto le operazioni aritmetiche sono eseguibili con la massima velocità quando gli operandi sono rappresentati in tale forma . I dati di ingresso e i risultati sono, tuttavia, quasi sempre forniti o voluti in forma decimale, peraltro facilmente ottenibile con i metodi prima descritti.
Molti calcolatori, tuttavia, sono dedicati a calcoli di tipo amministrativo. Si possono sempre utilizzare allo scopo numeri binari, ma vi sono esigenze che richiedono di eseguire i calcoli direttamente nella forma decimale. Basti pensare al problema degli arrotondamenti, che devono soddisfare precisi requisiti anche legali, difficili da realizzare su numeri convertiti in binario. Si può ricordare l'esempio mostrato in preceden-

za sulla conversione in binario di $0.1_{10}$, che genera un numero binario di lunghezza infinita e che perciò richiede un inevitabile arrotondamento dato che tutti i calcolatori rappresentano ovviamente i numeri utilizzando un numero finito di bit.
Una rappresentazione usata a questo scopo è quella costituita dal cosiddetto codice BCD/8421, rappresentato nella prima colonna della tabella 2. BCD sta per *binary coded decimals*, ossia cifre decimali codificate in binario, e 8, 4, 2, 1 sono i pesi associati nell'ordine ai 4 bit. La somma dei pesi delle cifre binarie pari a 1 individua la cifra rappresentata. Per esempio, 0110 individua la cifra 6 in quanto la sommatoria dei pesi fornisce per l'appunto $0 \cdot 8 + 1 \cdot 4 + 1 \cdot 2 + 0 \cdot 1 = 6$.
Nella tabella 2 sono riportate altre tre rappresentazioni binarie delle dieci cifre decimali, che godono di utili proprietà sulle quali si tornerà in seguito.

## 6. Numeri rappresentati in virgola mobile

Il modo più semplice per rappresentare i numeri reali consiste nel rappresentarne parte intera e parte frazionaria separate dal punto (l'equivalente anglosassone della virgola, di fatto uno standard nel mondo dei calcolatori). Tale rappresentazione non è però adatta nei calcoli tecnico-scientifici, che possono mettere in gioco contemporaneamente numeri piccolissimi e numeri grandissimi. Infatti, a causa del numero finito di bit utilizzati per rappresentare i numeri all'interno dei calcolatori, si potrebbero perdere molte o tutte le cifre significative per i numeri molto piccoli (oppure, per i numeri molto grandi, si limiterebbe il valore massimo rappresentabile).
La soluzione di tale problema è offerta dalla rappresentazione in virgola mobile (*floating point, FP*).Con essa, il numero decimale 765.432 può essere rappresentato con $7.65432 \times 10^2$; l'esadecimale –001D4B4E con $-1.D4B4E \times 16^{-2}$, il binario 101.111 con $1.01111 \times 2^2$.
Un numero in virgola mobile è, quindi, individuato da: segno, base (spesso implicita), esponente (associato alla base), frazione o mantissa (l'insieme delle cifre significative).
Per i numeri FP. binari lo IEEE (*Institute of Electrical and Electronics Engineers*) ha definito, con lo Standard 754, una rappresentazione a 32 bit così costituita:

*s eeeeeeee fffffffffffffffffffffff*

cioè con 1 bit per il segno (s), 8 bit per l'esponente in base = 2 (le *e*), 23 bit per la frazione (le *f*) che si sottintende essere del valore 1.*fffffffffffffffffffffff* (il bit di valore 1 nella parte intera non viene fisicamente memorizzato). In realtà, l'esponente è rappresentato sommando l'esponente vero (che può essere positivo o negativo) alla costante (*bias*) : $0111\ 1111_2 = 127_{10}$, allo scopo di semplificare il confronto tra numeri.
Con la suddetta notazione possono essere rappresentati numeri nel campo $10^{-44.85} - 10^{38.53}$.
È stato definito anche lo standard per numeri FP binari con 64 bit (per maggiori ragguagli vedasi [4, 8]).

## 7. Altri modi e casi di codificazione numerica

Sono in uso (per esempio per rappresentare date e tempi) numeri posizionali con pesi che non sono potenze intere di una base.
Un sistema concettualmente importante di rappresentazioni dei numeri binari si basa sull'uso dei "residui", cioè dei resti della divisione del numero dato per un, opportuna serie di divisori prefissati. Anche questo argomento verrà trattato in un successivo articolo.
Quanto finora detto riguarda comunque i numeri in senso stretto, ma la codificazione binaria è richiesta anche per le informazioni che si

| | BCD 8421 | Biquinario | 2 di 5 | Gray |
|---|---|---|---|---|
| 0 | 0000 | 01 00001 | 00011 | 0010 |
| 1 | 0001 | 01 00010 | 00101 | 0110 |
| 2 | 0010 | 01 00100 | 00110 | 0111 |
| 3 | 0011 | 01 01000 | 01001 | 0101 |
| 4 | 0100 | 01 10000 | 01010 | 0100 |
| 5 | 0101 | 10 00001 | 01100 | 1100 |
| 6 | 0110 | 10 00010 | 10001 | 1101 |
| 7 | 0111 | 10 00100 | 10010 | 1111 |
| 8 | 1000 | 10 01000 | 10100 | 1110 |
| 9 | 1001 | 10 10000 | 11000 | 1010 |

**TABELLA 2**
*Due rappresentazioni binarie delle cifre decimali*

esprimono tramite gli alfabeti delle varie lingue (codificazioni "alfanumeriche").
Sono, inoltre, necessari codici binari speciali per rappresentare altre forme di espressione, come il suono e le figure.
Un aspetto generale della informazione codificata è quello della sua integrità, non solo durante la sua elaborazione ma anche per quanto riguarda la trasmissione a distanza e la memorizzazione. La corruzione dell'informazione può manifestarsi con la perdita di uno o più bit o con la loro alterazione. La teoria dell'informazione ha sviluppato metodi che permettono la rivelazione di errori e anche la loro correzione (entro certi limiti). Furono proposti e usati alcuni codici destinati alle informazioni puramente numeriche, ma presto il problema venne affrontato per l'informazione in generale, anche nella forma alfanumerica. I codici "biquinario" e "due di cinque" della tabella 2 offrono, proprio per la ridondanza che li caratterizza, una sia pur modesta capacità di rivelazione di errori: il primo perché tutte le configurazioni sono costituite da due gruppi di bit e in ciascun gruppo un solo bit ha il valore 1; il secondo perché tutte le configurazioni hanno 2 e 2 soli bit con valore 1.

## Bibliografia


[1] Ancient Egiptian Mathematics
http://home.clara.net/beaumont/egypt/maths/

[2] Chinese Numbers
http://www.mandarintools.com/numbers.html
(con programma in linea per la conversione automatica).

[3] Evolution of Arabic (Roman) Numerals from India
http://www.gosai.com/chaitanya/saranagati/html/vishnu_mjs/math/math_4.html

[4] Hollasch, S.: IEEE Standard, 754-Floating Point Numbers
http://research.microsoft.com/~hollasch/cgindex/coding/ieeefloat.html

[5] Kailash Srivastava, Numbersystem, some clarification
http://manaskriti.com/InteractInn/10119801.html

[6] IEEE-754 Floating-Point Conversion
http://babbage.cs.qc.edu/courses/cs341/IEEE-754.html

[7] Indian numerals
http://www-gap.dcs.st-and.ac.uk/~history/HistTopics/Indian_numerals.html

[8] Maya Mathematics
http://www.michielb.nl/maya/math.html
(con programma in linea per la conversione automatica).

[9] Melville, Ancient and Classical Mathematics
http://it.stlawu.edu/%7Edmelvill/323/index.html

[10] von Neumann, John: The Principles of Large-Scale Computing Machines. Reprinted in: *Ann. Hist. Comp.*, Vol. 3, n. 3, 1946, p. 263-273.

[11] The Arabic numeral system
http://www-gap.dcs.st-and.ac.uk/~history/HistTopics/Arabic_numerals.html



Luigi Dadda laureato in Ingegneria Elettrotecnica. Titolare della cattedra di Elettrotecnica dal 1962. È stato Rettore del Politecnico di Milano dal 1 Novembre 1972 al 31 Ottobre 1984. Nel 1954 ha vinto una borsa di Studio della National Sciences Foundation per svolgere ricerche presso il California Institute of Technology di Los Angeles, dove ha partecipato alla realizzazione di un calcolatore elettronico che nel 1954 è stato installato al Politecnico di Milano. La sua attività di ricerca si é indirizzata inizialmente nel campo dei modelli e dei calcolatori analogici, poi in quello dei calcolatori elettronici dove, in particolare, si é occupato delle unità aritmetiche proponendo soluzioni originali per i moltiplicatori paralleli. Si é in seguito dedicato allo studio delle reti di Petri, proponendone l'impiego per il progetto di sistemi di controllo di grande complessità. Più recentemente ha esteso le ricerche sull'area della elaborazione di segnale con lo studio di nuovi sistemi di convolutori. Oltre all'attività scientifica e didattica, si é anche dedicato alla guida e al coordinamento della ricerca sia in sede nazionale che internazionale. Il Centro di Calcolo e il Laboratorio di Calcolatori Elettronici del Dipartimento di Elettronica del Politecnico di Milano da lui diretto, ha svolto ricerche avanzate sui sistemi di calcolo, sull'architettura di microcalcolatori, sui linguaggi di programmazione, sulle banche di dati e sulle reti di calcolatori. Nel 1980/82 ha presieduto la Commissione per la Scienza e la Tecnologia presso la Presidenza del Consiglio dei Ministri. È fondatore e direttore della "Rivista di Informatica".
luigi.dadda@polimi.it

# ICT E DIRITTO

## Rubrica a cura di

**Antonio Piva e David D'Agostini**

Scopo di questa rubrica è di illustrare al lettore, in brevi articoli, le tematiche giuridiche più significative del settore ICT: dalla tutela del *domain name* al *copyright* nella rete, dalle licenze software alla *privacy* nell'era digitale. Ogni numero tratterà un argomento, inquadrandolo nel contesto normativo e focalizzandone gli aspetti di informatica giuridica.

# La tutela dei dati personali nell'era digitale: il Codice sulla privacy tra vecchi e nuovi adempimenti

## 1. Introduzione

L'interesse e la sensibilità per la tutela della *privacy*, soprattutto a seguito dello sviluppo e della diffusione degli strumenti informatici e telematici, sono notevolmente aumentati nel corso dell'ultimo decennio, determinando di conseguenza una produttiva attività del Parlamento.

Gli sforzi del legislatore hanno portato, da ultimo, all'emanazione del "Codice in materia di protezione dei dati personali", contenuto nel decreto legislativo 30 giugno 2003 n. 196[1].

A partire dal 1° gennaio 2004, data dell'entrata in vigore del nuovo codice, sono state abrogate numerose fonti normative in materia di tutela della privacy tra cui la legge 31 dicembre 1996, n. 675 (aggiornata ben 10 volte in soli 5 anni) e il decreto del presidente della Repubblica 28 luglio 1999, n. 318 (altrimenti detto "Regolamento sulle misure minime di sicurezza").

Il nuovo testo unico ha l'indiscutibile merito di coordinare le molteplici disposizioni già vigenti, apportando ulteriori integrazioni e modifiche anche in recepimento della direttiva comunitaria 2002/58/CE inerente le comunicazioni elettroniche e compiendo un passo fondamentale, da un lato, per facilitare la conoscenza della normativa e dall'altro per garantire una più ampia e responsabile applicazione della stessa, soprattutto in presenza di trattamento informatizzato dei dati.

Il codice unitario, nel segno della continuità con le scelte legislative passate, riprende la terminologia dei testi precedenti, peraltro ulteriormente affinata sulla scorta dell'esperienza maturata negli ultimi anni, aggiungendo nuove definizioni di matrice tecnica, rese necessarie dall'evoluzione tecnologica degli ultimi anni (per esempio, *comunicazione elettronica, autenticazione informatica* ecc.).

Il "Codice in materia di protezione dei dati personali" si compone di tre parti (si veda, a tal proposito, il riquadro 1), inoltre, viene integrato dall'allegato B contente il disciplinare tecnico in materia di misure minime di sicurezza, dove si trovano, per i trattamenti con strumenti elettronici, le disposizioni e le modalità di autentificazione informatica, i sistemi di autorizzazione

**Riquadro 1: Suddivisione del Codice**

*Prima parte* (Disposizioni generali): disciplina sostanziale applicabile a tutti i trattamenti di dati personali, i diritti dell'interessato, gli adempimenti e la sicurezza dei dati e dei sistemi.

*Seconda parte* (Disposizioni relative a specifici settori): norme relative a specifici trattamenti per esempio, gli ambiti giudiziario, bancario e assicurativo, sanitario, dell'istruzione, del lavoro, del giornalismo ecc..

*Terza parte* (Tutela dell'interessato e sanzioni): articoli inerenti la difesa dei diritti dell'interessato e il sistema sanzionatorio.

**Allegato A**: codici di deontologia.

**Allegato B**: disciplinare tecnico in materia di misure minime di sicurezza.

**Allegato C**: trattamenti non occasionali effettuati in ambito giudiziario o per fini di polizia.

[1] Gazzetta Ufficiale 29 luglio 2003, n.174, S.O. Utile riferimento: *www.garanteprivacy.it*

e sicurezza e, inoltre, la regolamentazione del documento programmatico.

## 2. I SOGGETTI E GLI ADEMPIMENTI

Le figure principali individuate dal nuovo testo unico, peraltro già identificate dalla precedente normativa, sono l'*interessato* (ossia il soggetto al quale si riferiscono i dati personali), il *titolare* che decide le finalità e le modalità del trattamento di dati personali, il *responsabile* (individuato facoltativamente dal titolare tra soggetti dotati di esperienza in materia), gli *incaricati*, vale a dire le persone fisiche autorizzate a compiere operazioni di trattamento attenendosi alle istruzioni impartite dal titolare o dal responsabile, in relazione alla tipologia di dati trattati (si veda, a tal proposito, il riquadro 2).

**Riquadro 2: I dati personali**

*Identificativi*: permettono l'identificazione diretta dell'interessato.
*Anonimi*: non possono essere associati a un interessato identificato o identificabile.
*Giudiziari*: idonei a rivelare provvedimenti di natura penale.
*Sensibili*: idonei a rivelare l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni o organizzazioni a carattere religioso, filosofico, politico o sindacale, nonché lo stato di salute e la vita sessuale.
*Quasi sensibili*: il loro trattamento presenta rischi specifici per i diritti e le libertà fondamentali, nonché per la dignità dell'interessato.

Nella prassi il responsabile dei sistemi informativi aziendali viene spesso nominato quale responsabile del trattamento dei dati, costituendo, inoltre, un punto di riferimento per amministratori dei sistemi e incaricati alla custodia delle *password*. Pertanto, è essenziale che questi professionisti prendano conoscenza dei dettami riguardanti i trattamenti informatizzati, contenuti nel menzionato disciplinare tecnico in materia di misure minime di sicurezza, i cui temi vengono illustrati nel prossimo paragrafo.
Oltre agli atti di nomina del responsabile e degli incaricati (che devono essere effettuati per iscritto), vengono posti a carico del titolare del trattamento una serie di adempimenti: in primo luogo, ai sensi dell'art. 13 (che ricalca l'art. 10 L. 675/96), l'interessato deve essere preventivamente *informato*, oralmente o per iscritto, in merito alle finalità e modalità del trattamento, all'obbligatorietà o meno del conferimento dei dati e alle conseguenze di un eventuale rifiuto; egli dovrà, inoltre, sapere quali soggetti potranno venire a conoscenza delle informazioni raccolte e quale la logica viene applicata in caso di trattamento effettuato con l'ausilio di strumenti elettronici (si pensi alla comunicazione via *e-mail* o alla diffusione dei dati anche a mezzo di Internet).
L'interessato ha sempre il diritto di ottenere, oltre alle indicazioni presenti nell'informativa, la conferma dell'esistenza di dati che lo riguardano, l'aggiornamento, la modifica, la cancellazione, l'anonimizzazione dei dati medesimi, potendo opporsi per motivi legittimi al trattamento (come nel caso di invio di materiale pubblicitario non richiesto via Internet, tramite posta elettronica, pratica comunemente detta *spamming*).
In secondo luogo, l'art. 23 (sulla falsa riga dell'art. 11 L. 675/96) conferma che il trattamento è ammesso solo se l'interessato abbia prestato in maniera valida il proprio consenso. Il *consenso*, che deve essere espresso in forma specifica per ogni trattamento chiaramente individuato, va documentato per iscritto, e comunque preceduto dall'informativa. Qualora siano trattati dati sensibili, oltre al consenso manifestato in forma scritta da parte dell'interessato, è necessaria l'autorizzazione del Garante il quale può prescrivere misure e accorgimenti anche tecnici a garanzia dell'interessato stesso, che il titolare è tenuto ad adottare.
Le disposizioni sopra elencate devono essere rispettate anche in caso di trattamento elettronico delle informazioni e sistemi informativi automatizzati (si pensi alla raccolta dati per telefono da parte di un *call center*), eventualmente impiegando gli strumenti innovativi resi disponibili dalle moderne tecnologie, vale a dire anche, per esempio, tramite e-mail o *web*.
In particolare, nell'ipotesi non infrequente di trattamento dei dati mediante un sito Internet, l'informativa deve essere posizionata in calce alla pagina in cui vengono raccolte le informazioni personali (ovvero, in un'apposita finestra *pop-up*, o in altra maniera equivalente) e il consenso, se richiesto, può essere espresso tramite la compilazione di un *form* e successivamente registrato nelle memorie del

sistema, in ottemperanza all'obbligo di documentazione.
Quanto al trattamento dei dati sensibili, per il quale è richiesta la forma scritta, il consenso dovrà essere manifestato in maniera conforme alle normative vigenti in materia di firme elettroniche.

## 3. Le misure di sicurezza

Gli adempimenti inerenti la sicurezza dei dati e dei sistemi informatici sono finalizzati a ridurre al minimo i rischi di distruzione o perdita dei medesimi, ovvero di accesso non autorizzato o di trattamento non consentito, in relazione al tipo di dati trattati.
Il già menzionato Allegato B (Disciplinare tecnico in materia di misure minime di sicurezza) prevede, tra le modalità tecniche da adottare in caso di trattamento con l'ausilio di strumenti elettronici, il ricorso a sistemi di autenticazione informatica e di autorizzazione degli incaricati, mediante codice di identificazione personale (*username*) e parola chiave riservata (*password*), ovvero tramite dispositivi di autenticazione in possesso a uso esclusivo dell'incaricato (per esempio, *token*, *smart card*).
A tale scopo viene suggerito anche l'impiego delle tecnologie biometriche che trovano sempre maggiore applicazione per la loro elevata capacità di verificare in maniera sicura l'identità di un soggetto riconoscendone l'impronta digitale, l'iride, il timbro vocale e perfino i tratti somatici del volto.
I dati personali, inoltre, devono essere protetti contro il rischio di intrusione (per esempio, mediante *firewall* e programmi denominati *Intrusion Detection System*, IDS) e dall'azione di virus tramite idonei strumenti elettronici da aggiornare con cadenza almeno semestrale; vengono anche previsti l'aggiornamento periodico dei programmi finalizzati a prevenire la vulnerabilità dei sistemi e a correggerne i difetti (per esempio, *Patch* e nuove versioni) da effettuarsi almeno annualmente e il salvataggio periodico dei dati (*back-up*) con frequenza almeno settimanale.
Tra le misure di sicurezza, soprattutto nel caso di trattamento di dati sensibili, viene indicato anche l'utilizzo della cifratura, il che lascia pensare all'impiego su larga scala della crittografia basata sull'infrastruttura a chiave pubblica (PKI, *Public Key Infrastructure*).
Un tassello fondamentale nel mosaico della sicurezza informatica è costituito dal documento programmatico, da adottarsi entro il 31 marzo di ogni anno (ma per il 2004 il termine scadrà il 30 giugno!), obbligatoriamente nel caso di trattamento di dati sensibili[2].
Si tratta di un manuale, redatto anche in collaborazione con il responsabile, sulle politiche per la sicurezza delle informazioni e dei sistemi e sulle procedure e modalità di registrazione delle attività degli incaricati.
In particolare, il documento programmatico sulla sicurezza deve contenere l'elenco dei trattamenti, la distribuzione dei compiti e delle responsabilità, l'analisi dei rischi sui dati, le misure di sicurezza da adottare per l'integrità e disponibilità dei dati, la protezione delle aree e dei locali, la modalità di ripristino (si pensi al *disaster recovery*), il piano di formazione del personale incaricato al trattamento e inoltre, per i dati personali idonei a rilevare lo stato di salute e la vita sessuale, le modalità di cifratura e di separazione dei dati dalle altre informazioni personali dell'interessato.
Un'assoluta novità sul tema è l'obbligo di riferire in merito all'avvenuta redazione o aggiornamento del documento programmatico nella relazione accompagnatoria del bilancio d'esercizio. Con questa disposizione viene riconosciuto alla tutela dei dati personali un rango di importanza primaria nell'ambito della gestione aziendale di una società.

## 4. Le sanzioni e le responsabilità

La mancata ottemperanza alle disposizioni indicate nel nuovo codice della privacy comporta conseguenze sotto diversi profili.
Innanzitutto, possono configurarsi violazioni punite con sanzioni amministrative fino a 60.000 euro o addirittura sussistere illeciti pe-

[2] [N.d.A.] Vivamente consigliata anche in presenza di meri dati comuni, in quanto questa documentazione servirà da supporto al piano organizzativo interno nell'adempiere alle disposizioni legislative in materia di sicurezza, nonché come elemento probatorio nelle eventuali procedure ispettive: ovvero, in contenziosi di natura giudiziaria.

**Riquadro 3: Le sanzioni**

*Omessa o inidonea informativa all'interessato*: sanzione amministrativa da 3.000 a 18.000 euro (da 5.000 a 30.000 euro se si tratta di dati sensibili) aumentabile fino al triplo.
*Omessa o incompleta notificazione al Garante*: sanzione amministrativa da 10.000 a 60.000 euro.
*Omessa informazione o esibizione al Garante*: sanzione amministrativa da 4.000 a 24.000 euro.
*Omessa adozione delle misure minime di sicurezza*: arresto fino a 2 anni o ammenda da 10.000 a 50.000 euro.
*Falsità nelle dichiarazioni e notificazioni al Garante*: reclusione da 6 mesi a 3 anni.
*Inosservanza dei provvedimenti del Garante*: reclusione da 3 mesi a 2 anni.
*Trattamento illecito di dati*: reclusione da 6 a 18 mesi (da 6 a 24 mesi se consiste in comunicazione o diffusione, ovvero da 1 a 3 anni se dal fatto deriva nocumento).

nalmente rilevanti per i quali sono previste sanzioni detentive fino a 3 anni (si veda a tal proposito il riquadro 3).
In secondo luogo, alla responsabilità amministrativa e/o penale si affianca quella civile: il danno causato dal trattamento di dati personali deve essere risarcito, a meno che non si dimostri di aver adottato tutte le misure idonee a evitarlo, prova estremamente difficile da fornire in concreto. La giurisprudenza, pronunciandosi sul punto, ha già riconosciuto tale diritto, spesso liquidando poste di danno molto elevate; si ricorda che in materia di privacy può essere oggetto di risarcimento anche il danno non patrimoniale (si pensi al danno morale causato a una persona) in base all'art. 15 e all'art. 2059 c.c..
Concludendo, notoriamente l'ordinamento italiano non ammette l'ignoranza e a maggior ragione, come nel caso della privacy, dopo l'introduzione di uno specifico codice; pertanto, l'adeguamento alla normativa in esame da parte dei titolari del trattamento non può essere ulteriormente dilazionata.
Le responsabilità in gioco e l'inasprimento delle sanzioni, in particolare quelle pecuniarie, devono far riflettere chi già non l'abbia fatto, su tale necessità, soprattutto ora che l'Ufficio del Garante, dopo anni di rodaggio e grazie al recente aumento del proprio organico, sembra deciso ad applicare la legge senza deroghe di sorta.

Antonio Piva laureato in Scienze dell'Informazione, Presidente, per il Friuli - Venezia Giulia, dell'ALSI (*Associazione Nazionale Laureati in Scienze dell'Informazione ed Informatica*) e direttore responsabile della Rivista di Informatica Giuridica.
Docente a contratto di Informatica giuridica all'Università di Udine.
Consulente sistemi informatici, valutatore di sistemi di qualità ISO9000 e ispettore AICA per ECDL base e advanced.
E-mail: antonio_piva@libero.it

David D'Agostini avvocato, ha conseguito il master in informatica giuridica e diritto delle nuove tecnologie, fornisce consulenza e assistenza giudiziale e stragiudiziale in materia di software, privacy e sicurezza, contratti informatici, e-commerce, nomi a dominio, computer crime, firma digitale. Ha rapporti di partnership con società del settore ITC nel Triveneto.
Collabora all'attività di ricerca scientifica dell'Università di Udine e di associazioni culturali.
E-mail: david.dagostini@adriacom.it