1 Aspetti preliminari

1.1 Internet: breve riepilogo storico sulla "*rete delle reti*"

Sebbene il "fenomeno" Internet sia esploso a partire dagli anni novanta, i fondamenti delle sue origini risalgono ad una trentina d'anni prima, in stretta connessione con l'episodio che ha dato inizio alla storia contemporanea ufficiale, ovvero con la conquista dello spazio.

Per approfondimenti si rimanda a http://www.laterza.it; http://www.dariobonacina.net/.

1.1.1 La genesi della "*rete*"

Nel 1957 l'Unione Sovietica vinse il primo atto della competizione per la conquista dello spazio con la messa in orbita dello *Sputnik*. Tale evento colpì duramente gli Stati Uniti d'America non solo perché mise in dubbio il loro primato tecnologico, ma anche perché fece seriamente vacillare la sicurezza che vantavano in campo militare. Come immediata risposta l'amministrazione Eisenhower istituì, presso il Pentagono, l'*Advanced Research Projects Agency* (*ARPA*), con lo scopo di stimolare, anche finanziariamente, la ricerca militare nel settore delle tecnologie e delle comunicazioni.

Nel 1961 si concluse con successo anche la missione Yuri Gagarin, il primo uomo inviato nello spazio: un altro punto messo a segno dall'Unione Sovietica, che confermava il proprio primato in ambito aerospaziale.

Gli Stati Uniti stabilirono allora di investire in modo consistente nella ricerca e costituirono la *NASA* (*National Aeronautics and Space Aministration*), cui il governo trasferì la competenza di gestire i programmi spaziali e il cui frutto furono le missioni «Apollo».

L'*ARPA* dovette perciò rivolgersi ad un nuovo ambito di studio: avendo a disposizione costosi e sofisticati (per quel periodo!) elaboratori elettronici, decise di avviare un progetto per abilitare quelle macchine a comunicare tra loro e a trasferire dati. Nel 1969 ottenne il primo risultato concreto del progetto, chiamato *ARPAnet*. Nonostante il periodo storico-politico dell'epoca (si era da poco usciti dalla "guerra fredda") possa aver indotto a pensare che l'obiettivo primario di *ARPAnet* fosse garantire la sicurezza dei dati in caso di guerra nucleare (equivoco avallato anche dal fatto che l'agenzia risiedeva presso il Pentagono), pare assodato che lo scopo perseguito fosse limitato ad ottimizzare lo sfruttamento delle risorse informatiche nel campo della ricerca: essa fu di fatto lo strumento che permise di realizzare la condivisione dei sistemi informativi tra i poli universitari statunitensi.

Il nucleo originario di "utenti" coinvolti, collegati attraverso il Network Control Protocol (NCP), era costituito da: Università di Los Angeles (UCLA), Università di Santa Barbara (UCSB), Università dello Utah e *Stanford Research Institute* (SRI). L'Università della California fu la prima ad essere dotata di un IMP (*Interface Message Processor*), computer dedicato alla gestione del traffico dati (delle dimensioni di un frigorifero!), la cui memoria centrale vantava "ben" 12 KB (si consideri che la *sim card* di un comune telefono cellulare ne supporta 64!). Circa un mese dopo fu installato presso lo *Stanford Research Institute*, il secondo IMP. Fu così realizzato il primo tratto della rete, costituito da due nodi connessi con una linea dedicata a 50 Kbps. Pochi mesi più tardi vennero connessi e resi operativi anche i nodi delle Università di Santa Barbara e dello Utah.

1.1.2 Sviluppo delle principali applicazioni

Per garantire l'efficienza della comunicazione tra *host* era necessario stabilire un insieme di regole, che dovevano essere condivise dai computer interconnessi. Tali regole, denominate "protocolli", vennero rapidamente approntate e fu stilato un resoconto delle relative specifiche denominato *Network Control Protocol* (NCP). Il primo applicativo, progettato per il trasferimento di file, fu il cosiddetto *File Transfer Protocol* (FTP), tuttora largamente utilizzato.

L'applicazione che ebbe forse maggior successo e che influenzò in modo decisivo l'evoluzione della *rete* fu però la posta elettronica (*e-mail*). Frutto dell'iniziativa di un ingegnere, Ray Tomlinson, a cui si deve l'impiego del carattere "@" per separare il nome dell'*utente* da quello del *server*, la proposta fu accolta subito favorevolmente e le sue procedure applicative vennero presto integrate nel protocollo FTP.

Intanto la rete *ARPAnet*, come veniva ormai ufficialmente chiamata, cominciava a crescere rapidamente: nel 1971 era formata da 15, nodi con 23 *host*, e ne usufruivano alcune centinaia di utenti.

L'introduzione della posta elettronica incentivò l'istaurarsi di una sorta di comunità telematica costituita da giovani ricercatori di informatica (risale al 1975 la creazione del primo gruppo di discussione). Da qui si svilupperà in seguito quel fenomeno dello sviluppo condiviso di software gratuiti che ha generato il movimento «*Open Source*» (per un approfondimento si veda l'Appendice 1).

Nel 1972 il progetto era avviato con successo e gli sviluppi si delineavano ricchi di prospettive, si decise quindi di darne una dimostrazione pubblica in occasione della *International Conference on Computer Communications*. L'esito fu estremamente positivo: da qui prese inizio una collaborazione internazionale tra ricercatori dei settori informatico e delle telecomunicazioni, che ebbe tra gli obiettivi primari la produzione di nuove applicazioni capaci di consentire il dialogo tra sistemi sviluppati su piattaforme differenti. L'impegno fu diretto verso l'individuazione di un protocollo innovativo di interfacciamento tra *host*: il risultato fu l'elaborazione del *Transmission Control Protocol* (TCP), procedura in grado di garantire la comunicazione di pacchetti indipendentemente dalla struttura hardware. Esso comportò l'introduzione del *gateway*, una macchina deputata a svolgere funzioni di raccordo tra reti diverse. I risultati di questo lavoro furono pubblicati nel 1974 in un articolo dal titolo "*A Protocol for Packet Network Internetworking*", in cui comparve per la prima volta il termine «Internet» (http://www.laterza.it/internet/leggi/internet2004/online/07_temi_10.htm).

La DARPA (nuova denominazione dell'ARPA, al cui identificativo fu anteposto il termine *Defense*) investì sul progetto e gli studi che ne seguirono portarono ad un'ulteriore implementazione del protocollo TCP, che fu scomposto in due parti: il TCP propriamente detto con funzione di regolare la creazione ed il controllo gestione dei pacchetti da trasferire e l'IP (*Internet Protocol*), che ne governa l'invio alla corretta destinazione.

Il TCP/IP, che presenta l'imprescindibile vantaggio di essere svincolato dalle caratteristiche fisiche delle macchine connesse, costituisce da allora il protocollo standard di trasmissione via Internet: ogni computer collegato alla *rete* è tuttora dotato di un indirizzo IP.

Alla fine degli anni '70 *ARPAnet* era ancora composta solamente da quindici nodi, ma negli Stati Uniti esistevano ormai numerosi altri dipartimenti di informatica la cui esclusione dal circuito comportava di fatto un depauperamento delle potenzialità della ricerca. Per consentire loro di fornire il proprio contributo allo sviluppo del progetto e per espandere le possibilità di comunicazione tra i ricercatori, la *National Science Foundation* (NSF), ente

governativo preposto alla sponsorizzazione della ricerca di base, decise di finanziare la costituzione di reti più economiche. Nel 1981 era operativo un sistema di allacciamento tra i dipartimenti di informatica dei vari poli universitari statunitensi denominato *CSnet* (*Computer Science Network*).
Altre due tappe fondamentali nell'evoluzione del fenomeno Internet furono: la creazione di un nuovo protocollo per la gestione della posta elettronica, denominato *Simple Mail Transfer Protocol* (SMTP) e la messa a punto di un sistema innovativo di identificazione dei nodi della *rete*, il *Domain Name System* (DNS), che permette di associare ad essi un vero e proprio nome, sicuramente più "intuitivo" rispetto al codice numerico costituente l'indirizzo IP.

1.1.3 La propagazione della *rete*

Visto il rapido diffondersi degli allacciamenti, la Defence Communication Agency, che aveva in gestione *ARPAnet*, decise per motivi di sicurezza di suddividere la rete in due settori. In questo modo sarebbe stato possibile separare una frazione dedicata unicamente alla gestione delle questioni militari (chiamata *MILnet*), che doveva rimanere ad accesso riservato, da una seconda porzione dedicata alla ricerca scientifica, che poteva rimanere priva di vincoli di accessibilità (la quale conservò il vecchio nome di *ARPAnet*). La completa "liberalizzazione" di tale sistema fu supportata dalla creazione di un nuovo organismo di gestione tecnica l'*Internet Activities Board* (IAB), che comprende tra i suoi dipartimenti l'*Internet Engineering Task Force* (IETF), cui fu affidato il compito specifico di definire gli standard della *rete*, compito che mantiene ancora oggi.

Nel frattempo anche altri paesi avevano iniziato a costituire reti di ricerca, fondate sul TCP/IP (le cui specifiche erano disponibili gratuitamente e liberamente utilizzabili), e quindi abilitate all'interazione con quelle nordamericane. Già verso la fine degli anni '60, infatti, altri centri di elaborazione avevano cominciato ad impiegare la tecnologia *ARPAnet* di *packet switching* per collegare i propri sistemi. In Canada, ad esempio, venne selezionata una singola università per ogni provincia e venne creata una *backbone* che attraversava il paese da est ad ovest, nota col nome di "*CAnet*".

Avendo creato i presupposti concettuali per un utenza illimitata, si rese ben presto necessario investire sull'adeguamento delle capacità strutturali della rete, che doveva essere abilitata a sostenere un traffico in continua crescita. La NSF prese in carico la realizzazione di una nuova rete ad alta velocità, cui diede il nome di *NSFnet*, che congiungeva i principali centri di calcolo del paese con una linea dedicata a 56 Kbps. Venne consentito libero accesso a tutte quelle università disposte ad accollarsi il carico di spese necessarie alla creazione delle relative infrastrutture in loco. L'adesione fu pressoché totale: in breve il numero di *host* di Internet aumentò vertiginosamente; si rese quindi necessario adeguare l'efficienza della linea portandola a 1,544 Mbps. La nuova rete fu così in grado di fornire servizi decisamente superiori a quelli della DARPA, che finì perciò col trasferire sulla *NSFnet* tutti i propri siti: l'obsoleta *ARPAnet* venne definitivamente chiusa nel 1989. In quell'anno cadde anche il muro di Berlino, decretando la conclusione definitiva della separazione tra "est" e "ovest" e la liberalizzazione delle politiche dei Paesi dell'Europa orientale.

Nei primi anni '90 le regole di adesione a *NSFnet* vennero modificate per consentire l'ingresso anche ad organizzazioni commerciali, mentre l'utenza, non più confinata all'ambito della ricerca scientifica e tecnologica, cresceva a ritmi sempre più sostenuti (il numero di *host* superava già le 100 mila unità!). Questo rese necessario sviluppare da un lato alcune misure di

sicurezza (il primo virus risale al 1988), dall'altro applicazioni più *user friendly*: un primo passo fu la strutturazione gerarchica delle informazioni secondo uno schema ad albero *server/clients*.

1.1.4 Le applicazioni che hanno reso Internet alla portata di tutti

Nel frattempo una svolta decisiva veniva dall'Europa. Presso il CERN (Centre de Européen de Recherche Nucléaire) di Ginevra un team, diretto da Tim Berners Lee, sviluppò un'applicazione per condividere attraverso la *rete* le informazioni, organizzate secondo una struttura, quella ipertestuale, capace di renderle consultabili in modo immediato ed intuitivo. Risale al novembre del 1990 un documento che descrive in modo dettagliato il protocollo "*HTTP*" e il concetto di "*browser*" e di "*server*", e che ufficializzava il nome "*World Wide Web* (WWW)" [3], scelto da Berners Lee per la sua invenzione.

Nel 1993 fu realizzato Mosaic, il primo *browser* grafico. La semplicità di installazione e di utilizzo, che non richiedeva più una competenza informatica approfondita, resero possibile l'approdo al World Wide Web a milioni di utenti.

Il binomio Mosaic/WWW, proposto in un momento così favorevole (Internet aveva ormai raggiunto i due milioni di *host* e la banda della *NSFnet* era stata portata a 44,736 Mbps) ebbe un successo eclatante. L'ideatore di Mosaic venne convinto a sfruttare commercialmente l'ampio consenso di pubblico della sua creazione, ma per evitare di pagare *royalties* decise di riscrivere completamente un nuovo programma. Venne così elaborata la prima versione (*beta*) di *Netscape Navigator*, le cui caratteristiche innovatrici gli permisero di soppiantare rapidamente il suo predecessore.

Entro breve si assistette anche alla proliferazione di connessioni Internet allestite da gestori privati, prime tra tutti le multinazionali delle telecomunicazioni: il controllo tecnico della *rete* rimaneva alla *Internet Society* (organizzazione *no profit* fondata nel 1992), ma il carico degli investimenti era divenuto ormai troppo oneroso per essere sostenuto dai soli istituti di ricerca.

Alla prima versione di *Netscape* (1994) ne seguirono altre, mentre nel 1995 la Microsoft immetteva nel mercato il suo antagonista per eccellenza: *Internet Explorer*.

Alla fine del 1996 c'erano già più di 3,5 milioni di *host* registrati e circa 60 milioni di utenti. Dalla metà degli anni '90 in poi, le sorti della la *rete*, ormai ampiamente svincolata dall'ambito della ricerca scientifica, rimangono caratterizzate sempre più significativamente dallo sviluppo commerciale.

A conclusione di questo excursus storico sull'evoluzione di Internet, si vuole porre in evidenza come la rapidità e l'efficienza del suo progresso siano stati il frutto della cooperazione intellettuale e della libera circolazione delle idee tra i gruppi di ricerca, i quali condividevano intuizioni, progetti, soluzioni e tecnologie, considerandoli un unico patrimonio collettivo. Qualunque progresso, mantenuto al riparo da condizionamenti politici ed

[3] Un sistema di computer WWW fa da *server* per le pagine. Queste pagine, composte di testo e immagini, ma anche suoni, clips e video, sono redatte usando l'*Hyper Text Markup Language* (HTML). Qualunque elemento che compone una pagina Web può comprendere una connessione ad un'altra pagina. La cosa interessante è che un solo documento HTML può essere collegato a pagine WWW che risiedono all'interno di altri computer, in ogni parte del mondo, grazie ai *server* Web, i quali sono in grado di comunicare grazie alla condivisione di un protocollo applicativo comune l'*Hyper Text Transfer Protocol* (HTTP).

economici, veniva subito reso fruibile per l'intera comunità scientifica, poiché dalla condivisione seguiva un arricchimento che andava a vantaggio di tutti.
Questa ideologia, che rimane alla base dei prodotti *Open Source*, è il motore del progresso scientifico e tecnologico; deve pertanto essere strenuamente difesa nell'interesse pubblico globale.

1.2 L'approccio scientifico allo studio dell'*ambiente*

L'*ambiente*, inteso nella sua accezione più ampia (ovvero comprendente atmosfera, geosfera, idrosfera, biosfera e la totalità delle loro interazioni) è un sistema estremamente articolato e perciò stesso fisicamente irriproducibile, non solo per la sua vastità spaziale, ma anche per la complessità fenomenologica che lo caratterizza. L'enorme numero di variabili in gioco rende ciascun evento sostanzialmente irripetibile (anche per la "*Natura*" stessa!) ovvero nessun fenomeno ambientale consente, di fatto, repliche esattamente identiche: troppi fattori in gioco per riprodurli fedelmente o perché si ripresentino con le stesse modalità.

La ricerca scientifica in questo campo deve quindi necessariamente limitarsi, una volta stabilite le scale spaziale e temporale di interesse, ad identificare le dinamiche che regolano gli eventi e, per valutare l'entità ed i possibili impatti dei fenomeni oggetto di studio, accontentarsi di riprodurli attraverso simulazioni. Quella delle scienze ambientali si configura perciò come una particolare categoria delle scienze della natura, la quale deve operare tramite osservazioni e misure effettuate direttamente in campo, applicando inevitabilmente un approccio di studio multidisciplinare.

Il metodo scientifico (che per assunto distingue la scienza dalle opinioni) definisce come "sperimentale" una conoscenza resa obiettiva dalle verifiche, intese come strumento capace di conferire oggettività al sapere individuale e consentire la sua trasmissione a terzi. Nel caso delle scienze "classiche" (fisica, chimica, biologia, …) tale strumento è costituito da modelli fisici (generalmente riproduzioni empiriche condotte in laboratorio), mentre per le scienze ambientali consiste necessariamente in modelli formali (algoritmi)[4].

Ne consegue che "simulare un evento ambientale" significa fornire uno o più algoritmi che collegano le coordinate dello spazio delle fasi (indicatori dell'evento oggetto di studio) con le variabili dello spazio geometrico e con la variabile tempo, in modo da riuscire a calcolare la traiettoria che descrive l'evento punto per punto ed istante per istante (Marani, 1999). Quali che siano gli algoritmi prescelti come i più funzionali allo scopo che ci si è prefissi, una cosa è certa: il modello che se ne ricava è tanto più perfettibile (e quindi maggiormente affidabile in termini applicativi) quanto più ampio e variegato è il pool di dati su cui poter sperimentare i test di verifica, ovvero in base ai quali poterlo "validare". Questo non significa affatto che quanti più tipi di parametri vengono misurati e quante più variabili vengono inserite nel modello, migliore sarà il risultato. Viceversa, è di fondamentale importanza prestare attenzione a non cadere nella ridondanza, che appesantisce inutilmente i calcoli di elaborazione diminuendo l'efficienza senza per contro incrementare l'efficacia: più "semplice" ed essenziale si riesce a rendere il modello, tanto più risulterà funzionale allo

[4] Il termine "modello" viene talvolta usato con il significato di "simulatore", talaltra con quello di "descrittore". Entrambi questi significati intervengono nel pensiero scientifico: il primo per intendere uno strumento di verifica ed il secondo uno strumento tassonomico. Gli algoritmi, come pure gli apparati di laboratorio, sono i più svariati e si affidano alla fantasia dei ricercatori. Non è questo il luogo per classificarli né tanto meno catalogarli, ma è importante tener presente la differenza fra modelli concettuali (modelli qualitativi che inquadrano i fenomeni) ed i simulatori (modelli operativi che permettono di riprodurre, formalmente o fisicamente, gli eventi).

scopo (a tale proposito si ricorda il principio del «*rasoio di Occam*»[5], criterio minimo ma essenziale con cui è opportuno proceda la scienza).

La possibilità di individuare parametri chiave, capaci di rappresentare in maniera significativa i fenomeni, costituisce una condizione necessaria allo studio delle problematiche ambientali, soddisfacendo a quel criterio di efficienza appena citato.

Le misure eseguite sull'ambiente sono particolarmente preziose per il fatto di essere uniche (ciò che viene rilevato è un dato strettamente legato ad una regione di spazio circoscritta e ad un intervallo di tempo definito) ed irripetibili (le combinazioni possibili delle variabili in gioco sono pressoché illimitate). Per questo è particolarmente importante creare archivi che raccolgano non solo dati isolati, ma sopratutto le successioni di informazioni acquisite in diversi punti dello spazio durante uno stesso arco temporale (scene), in uno stesso luogo durante intervalli di tempo consecutivi (serie temporali o sequenze), oppure seguendo entrambe queste procedure (episodi) (Marani, 1999).

Se il monitoraggio, condotto attraverso la misura automatizzata di parametri indicatori presso stazioni dislocate in punti "strategici" del territorio, consente di tenere sotto controllo le fluttuazioni attuali e recenti dei fenomeni ad essi correlati, i dati raccolti in passato, per quanto approssimativi e poco precisi, se ben documentati, ovvero dotati di un ampio corredo di metadati, sono indispensabili per formulare ipotesi plausibili riguardo gli andamenti temporali dei processi ambientali.

Sono proprio i *metadati*, insieme delle notizie "accessorie" ai dati, costituenti una sorta di loro curriculum identificativo, che conferiscono ad essi non solo una connotazione scientifica, ma anche un vero e proprio significato. La grandezza numerica o l'aggettivo che quantificano o qualificano, rispettivamente, un'osservazione non hanno alcun valore intrinseco, ma ne assumono uno (inequivocabile ed incontrovertibile) solamente in relazione al contesto di acquisizione, ovvero al "come", "dove", "quando" e "perché" il rilevamento è stato compiuto. Perciò ogni considerazione sulla qualità di una qualsiasi informazione è demandata alla qualità dei suoi metadati, che forniscono d'altro canto anche la base su cui poter eseguire operazioni di confronto.

È indubbio che con queste premesse la mole di informazioni da considerare assume dimensioni a dir poco enormi e, fino a poco tempo fa, poteva sembrare assurda la pretesa di volerle non solo conservare, ma anche gestire agevolmente. Tuttavia, ad oggi, lo sviluppo delle nuove tecnologie ha raggiunto un tale livello di sofisticazione da consentire tanto l'archiviazione e la condivisione, quanto l'elaborazione ed il trasferimento di contenuti informativi da un capo all'altro del mondo con estrema agilità. Di fatto, però, problemi politici, di *privacy* e di proprietà stanno ritardando la completa liberalizzazione delle banche di dati ambientali a seguito di un controverso dibattito, sulla opportunità e sulle eventuali modalità di gestione, sorto fra quanti ne auspicano la totale gratuità e quanti ne sostengono la completa riservatezza. Fortunatamente sull'argomento la legislazione nazionale e quella internazionale si stanno muovendo nella direzione della circolazione libera di tutti i dati ambientali di produzione pubblica (si vedano i già citati «D.L. 24 febbraio1997 n. 39» e «L. 16 marzo 2001 n. 108»).

[5] Critedio del *Rasoio di Occam* (Occam's Rasor): «Se per un dato fenomeno sono possibili più spiegazioni, allora è da preferire la spiegazione più semplice».

1.3 Archivi Digitali

La realizzazione di modelli matematici richiede il maggior numero possibile di dati ed informazioni, sia nella fase di definizione, sia per la successiva messa a punto e verifica. Anche un cospicuo patrimonio informativo può però risultare inutilizzabile se i dati stessi non sono raccolti ed organizzati in forme e modi tali da renderne facile l'accesso, la manipolazione, la divulgazione e, prima ancora, se non ne è resa manifesta l'esistenza.

Un archivio è una raccolta di documenti organizzata secondo un ordine tale da agevolarne la consultazione. Lo scopo è conservare il materiale all'interno di una struttura che ne preservi l'integrità e catalogarlo in modo da documentarne l'esistenza e la collocazione, scegliendo un criterio opportuno che fornirà successivamente le linee guida per la ricerca e il recupero. La "variante" informatica di questo tipo di struttura è la "base di dati", o "database", dove gli elementi informativi sono organizzati in tabelle e registrati su supporto digitale. L'informatica ha permesso la realizzazione di strumenti estremamente potenti, in termini sia di capienza sia di versatilità, per l'ordinamento e la gestione dei dati e le tecnologie, in continua evoluzione, prospettano sempre nuovi progressi.

Gli aspetti caratterizzanti ed interessanti (in quanto ricchi di implicazioni) dell'archiviazione digitale, sono sostanzialmente quattro: i) qualunque tipo di informazione che possa essere trasformata in un documento digitale, ovvero tradotta in un insieme di bit, può essere archiviata; ii) lo spazio "fisico" occorrente per contenere le informazioni è estremamente ridotto; iii) la possibilità di riprodurre documenti esattamente identici all'originale, oppure convertiti in formati che rispondano alle esigenze del fruitore, è illimitata; iv) le tecnologie multimediali e la possibilità di strutturare i documenti in forma di ipertesto consentono una consultazione interattiva.

Un'ulteriore "agevolazione" deriva dalle risorse messe a disposizione da Internet, che è in grado di fornire l'accesso ad archivi remoti, permettendo di ottenere a distanza copie della documentazione ricercata.

Rapidità di accesso, facilità nella consultazione, vasta documentazione disponibile in uno spazio ristretto e riproducibilità illimitata sono quindi i vantaggi più cospicui dell'archiviazione digitale su quella tradizionale, senza trascurare che quest'ultima non è in grado di tutelare gli oggetti più fragili e/o consultati con maggior frequenza dall'inevitabile usura.

Per sovrintendere alla gestione degli archivi digitali sono stati sviluppati programmi specifici, detti *DataBase Management System*, che regolano le procedure di catalogazione e di consultazione dei dati.

1.3.1 Database e loro sistemi di gestione

Un archivio informatizzato è costituito essenzialmente da blocchi di dati (*record*) organizzati in unità d'informazione elementari ed indivisibili (*campi*). Le funzioni principali che un database è chiamato a svolgere sono:

- consentire l'accesso ai dati attraverso uno schema concettuale, anziché attraverso uno schema fisico;
- permettere la condivisione e l'integrazione dei dati fra applicazioni differenti;
- consentire e controllare l'accesso condiviso ai dati da parte di più utenti;

- esercitare un controllo sulla congruenza dei dati;
- garantire la sicurezza e l'integrità dei dati.

Tali requisiti sono irrinunciabili quando si utilizzano applicazioni che si servono di Internet come infrastruttura, poiché richiedono sistemi con elevati gradi di fruibilità ed affidabilità (http://www.html.it/sql/sql_02.htm).

Un database è quindi una collezione di dati strutturati che viene organizzata ed amministrata da un software specifico, il *DataBase Management System* (DBMS). Un DBMS svolge essenzialmente funzioni di "intermediatore" tra i dati e chi ne fa uso, attraverso procedure applicative deputate alla gestione coordinata delle informazioni a differenti livelli. Tali procedure permettono di accedere ai contenuti senza bisogno di conoscerne nei dettagli la struttura e la reale allocazione, ma anche di modificare alcune aree di competenza abilitate, assicurando comunque un controllo centralizzato ed integrato del trattamento dati.

L'utente, e gran parte delle applicazioni stesse, non necessitano di raggiungere i dati nella forma in cui sono effettivamente memorizzati, ma ne leggono direttamente una riproduzione logica, il che garantisce un'elevata indipendenza del tipo di memorizzazione dalle applicazioni di interrogazione. Chi amministra il database può scegliere di cambiare la struttura dei dati o anche di impiegare un diverso DBMS senza che ciò venga ad incidere sulle capacità operative delle applicazioni, purché venga mantenuta la rappresentazione logica dei dati, chiamata "*schema del database*". È questa la forma di più basso livello organizzativo cui un utente può accedere, ovvero il livello minimo con cui gli utilizzatori possono interagire.

Nell'architettura tipica di un DBMS si distinguono tre livelli: livello interno, livello logico e livello esterno:

- il livello interno è relativo al modo in cui i dati sono organizzati nelle strutture fisiche di memorizzazione;
- il livello logico è relativo alla descrizione dei dati e delle relazioni tra essi;
- il livello esterno riguarda il modo in cui i diversi utenti interagiscono con la base di dati.

Si distinguono inoltre tre tipi di linguaggi:

- il linguaggio di definizione (*Data Definition Language*), con cui si definiscono gli schemi esterni, lo schema interno e lo schema logico;
- il linguaggio di controllo (*Data Control Language*), che consente di esprimere comandi per estrarre i dati contenuti nel data base;
- il linguaggio di manipolazione (*Data Manipulation Language*), che consente di inserire, cancellare e modificare i dati.

Il linguaggio di definizione viene usato al momento di stabilire un nuovo schema ed ha lo scopo di descrivere i tipi di dati di interesse e le relazioni che intercorrono tra di essi. Per questo motivo non viene utilizzato da tutti gli utenti, ma solo da una specifica figura della organizzazione, il cosiddetto amministratore della base di dati. Il linguaggio di interrogazione ed il linguaggio di manipolazione vengono utilizzati dai vari utenti per l'uso effettivo dei dati di loro interesse (Callegari, 1994).

Sono stati ideati diversi modelli di organizzazione ed archiviazione dei dati e, conseguentemente, differenti tipi di DBMS, che vengono comunemente classificati secondo il

tipo di *schema del database* caratterizzante il livello logico e raggruppati secondo le seguenti categorie:

Database gerarchici: i dati risultano organizzati secondo insiemi connessi tramite relazioni di appartenenza, seguendo una *struttura ad albero* in cui un insieme può comprendere al proprio interno numerosi insiemi, ma può a sua volta ricadere all'interno di un unico altro insieme. Ne consegue che talvolta due record, appartenenti ad alberi diversi, devono contenere una stessa informazione. Ciò causa problemi di ridondanza: il database richiede quindi controlli periodici per verificare la *consistenza* dei dati.

Database reticolari: il modello reticolare è un'estensione di quello gerarchico; anche in questo caso gli insiemi di dati sono legati da relazioni di appartenenza, ma tali relazioni non hanno carattere esclusivo: ogni insieme può essere compreso in differenti altri insiemi, che ne condividono la "proprietà". Nel modello reticolare i record sono legati tra loro con strutture ad anello. Tale configurazione, libera dai rigidi vincoli gerarchici, permette di evitare i problemi di ridondanza ma genera una complessità crescente in modo proporzionale all'aumentare dei dati. Ciò rende talvolta più conveniente realizzare da capo un nuovo database piuttosto che implementare quello preesistente.

Database relazionali: sono fondati sulla base del cosiddetto *modello relazionale*, la cui struttura principale, la *relazione* appunto, è costituita da una tabella bidimensionale composta da righe e colonne. Ciascuno degli oggetti da memorizzare nel database è rappresentato da una riga, detta *tupla*, mentre le caratteristiche che definiscono l'oggetto sono individuate dalle colonne della *relazione*, chiamate *attributi*. Quindi gli oggetti con caratteristiche comuni, ovvero descritti dallo stesso insieme di attributi, apparterranno ad una stessa *relazione* (ovvero potranno essere inseriti nella medesima tabella, di cui occuperanno ciascuno una riga differente).

Database ad oggetti (object-oriented): lo schema di un database ad oggetti è rappresentato da un insieme di classi, che definiscono le caratteristiche ed il comportamento degli elementi che popoleranno il database. La principale differenza con i modelli descritti finora è la non passività dei dati. Infatti, mentre per i casi precedenti le operazioni che devono essere effettuate sui dati vengono demandate alle applicazioni di interrogazione, in un database object-oriented gli oggetti memorizzati custodiscono sia i dati sia le operazioni attuabili su di essi. In un certo qual modo è come se i dati venissero "istruiti" su come comportarsi, in modo tale che non sia necessario disporre dell'ausilio di applicazioni esterne per la loro interrogazione.

Attualmente i più diffusi sono i database relazionali (ampiamente affermatisi già negli anni '80), il cui successo è dovuto, oltre che alle proprietà intrinseche (si basano su un modello, quello relazionale, con solide basi teoriche) ed alla praticità (forniscono sistemi semplici ed efficienti di rappresentazione e manipolazione dei dati), al successo conseguito da un linguaggio di interrogazione standard, lo *Structured Query Language* (SQL), che permette, almeno potenzialmente, di sviluppare applicazioni indipendenti dal particolare tipo di DBMS relazionale adottato.

I database ad oggetti costituiscono invece l'avanguardia della ricerca. Contraddistinti da una notevole flessibilità, derivante da una serie di proprietà specifiche (*Object Identity,*

Incapsulazione, Polimorfismo, Completezza Computazionale, Estendibilità, …) capaci di conferire al sistema abilità particolari, risultano notevolmente adatti ad applicazioni che richiedono di operare su entità complesse, frutto dell'aggregazione di dati differenti.

Tuttavia, l'assenza di uniformità nei modelli di strutturazione degli oggetti (benché siano state definite, come già per il modello relazionale, un insieme di regole caratterizzanti[6]) e la mancanza di un linguaggio di interrogazione standard, danno spazio ad ogni fornitore di elaborare una propria visione autografa, generalmente (e volutamente!) del tutto incompatibile con le altre presenti sul mercato. Recentemente sono stati messi in commercio alcuni database, definiti *object-relational*, che cercano di introdurre nel modello relazionale le caratteristiche di estendibilità proprie dei database object-oriented (http://www.html.it/sql/).

1.3.2 Il modello relazionale

La struttura su cui si fonda il modello relazionale è, come si è detto, una tabella bidimensionale, denominata *relazione*, costituita da un certo numero di righe (*tuple*) e di colonne (*attributi*), che intersecandosi individuano delle caselle (*campi*). Ogni *relazione* rappresenta una categoria di oggetti, detta *entità*; ciascuna *tupla* della relazione conterrà informazioni specifiche, relative ad ogni singolo oggetto appartenente a quella categoria, che viene chiamato *istanza dell'entità*, mentre la tabella compilata viene denominata *istanza di relazione*.

Le *tuple* di una relazione costituiscono un insieme nel senso matematico del termine, cioè una collezione non ordinata di elementi differenti appartenenti ad una medesima categoria. Per distinguere tali elementi l'uno dall'altro viene individuato, o arbitrariamente assegnato, un particolare *attributo* che svolge il ruolo di *chiave primaria*. Tale *attributo* deve presentare un valore sempre definito e differente per ciascuna *tupla*, in modo da consentire di identificarla univocamente. La proprietà delle *relazioni* che stabilisce l'impossibilità per la *chiave primaria* di assumere valore nullo, ovvero che l'*attributo* corrispondente rimanga indeterminato, viene definita *entity integrity* (http://www.html.it/sql/).

Sebbene spesso sia individuabile più di un *attribut*o adatto a svolgere la funzione di *chiave primaria*, generalmente si preferisce assegnare tale ruolo a codici numerici, che vengono appositamente introdotti nella *relazione* come *attributi* fittizi.

Gli *attributi* di una *relazione* sono identificati da un'intestazione e sono delimitati da un *dominio*, il quale ne definisce l'ambito di validità, stabilendo la tipologia e l'estensione caratterizzanti gli elementi costitutivi. Tra le funzioni del DBMS è verificare che gli *attributi* delle *relazioni* siano popolati esclusivamente con valori ammessi dai rispettivi *domini*. Una proprietà caratteristica dei database relazionali consiste nella necessità che i *domini* siano "atomici", ovvero i valori costitutivi degli *attributi* non possono essere ripartiti tra sottodomini di complessità inferiore rispetto a quelli individuati; in altre parole non sono ammessi *attributi* multivalore.

Un'altra importante proprietà dei database relazionali è la *referential integrity* e riguarda le cosiddette *chiavi esterne*. Due relazioni possono infatti essere connesse attraverso la

[6] Nel 1985 è stato pubblicato un "manifesto" dei database relazionali, in cui vengono definite senza ambiguità le caratteristiche che un DBMS deve soddisfare per poter essere chiamato *relazionale*.
Analogamente, nel 1989 sono stati definiti con precisione i requisiti di un DBMS che aspiri al titolo di OODBS. Tali requisiti vengono suddivisi in obbligatori, opzionali, ed aperti (ovvero suscettibili di scelta da parte del produttore). http://www.eptacom.net/pubblicazioni/pub_it/dati.sidebar1.html

condivisione di un *attributo*, caratterizzato da uno stesso *dominio*: se tale *attributo* svolge il ruolo di *chiave primaria* per una di esse, costituirà una *chiave esterna* per l'altra. La *referential integrity* verifica che i valori inseriti nel campo della *chiave esterna* siano compatibili con quelli della *chiave primaria*, ovvero controlla che tali valori, salvo non siano nulli, corrispondano a valori effettivamente presenti nella relazione in cui tale *attributo* funge da *chiave primaria*. In altri termini se si aggiunge una *tupla,* immettendo nella colonna delle *chiavi esterne* un valore non previsto tra quelli già assegnati ai campi della *chiave primaria*, il sistema segnala un errore.

I linguaggi di interrogazione per il modello relazionale si basano su due approcci fondamentali: l'*algebra relazionale* e il *calcolo relazionale*.

Nei linguaggi basati sull'algebra relazionale le interrogazioni sono espressioni composte da operatori algebrici applicati ad una o più relazioni, che danno come risultato nuove relazioni; nei linguaggi basati sul calcolo relazionale le interrogazioni sono espresse per mezzo di predicati che descrivono le proprietà caratterizzanti le relazioni.

Ogni sistema RDBMS fa uso di un proprio "dialetto", nonostante sia definito uno standard ANSI (American National Standard Institute) e ISO (International Standards Organization).

Uno dei linguaggi basati prevalentemente sul calcolo relazionale, il già citato SQL, fornisce all'utente un'interfaccia particolarmente agevole per la formulazione delle interrogazioni.

SQL è stato sviluppato dall'IBM a metà degli anni settanta per la manipolazione logica dei dati in un database relazionale e, a distanza di tempo, è divenuto, grazie alla sua semplicità, il modo standard per accedere ai dati in un database relazionale, risultando assai più simile ad uno strumento di comunicazione che ad un linguaggio di programmazione. È un linguaggio di tipo non procedurale, che consente di operare sui dati tramite frasi composte con "parole chiave" prese dal linguaggio corrente.

Come qualunque altro linguaggio, presenta regole sintattiche ed una grammatica che devono essere rispettate: la prima parola di qualsiasi istruzione è il verbo, che indica al database l'operazione che l'utente desidera compiere; le parole successive al verbo indicano i nomi delle colonne contenenti le informazioni (parametri) che l'utente intende visualizzare; dopo l'elenco dei parametri va inserita la "parola riservata" «FROM», seguita dal nome della tabella in cui si trovano i dati richiesti: questa parte della sintassi indica al database dove cercare le informazioni desiderate. A questa prima parte del "discorso" possono seguire ulteriori "clausole" per specificare le operazioni che si intendono applicare alle informazioni selezionate.

Le istruzioni di base dei tre linguaggi sono:

- Data Definition Language (DDL):
  1 CREATE (crea tabelle, indici, *viste* nel database);
  2 ALTER (modifica la definizione di una tabella);
  3 DROP (cancella tabelle, indici, *viste* dal database);
- Data Control Language (DCL):
  1 GRANT (attiva privilegi su tabelle o *viste*);
  2 REVOKE (revoca privilegi su tabelle o *viste*);
- Data Manipulation Language (DML):

1 SELECT (estrae le informazioni dal database);

2 INSERT (aggiunge nuovi dati in una tabella o *vista*);

3 UPDATE (modifica i valori esistenti in una tabella o *vista*);

4 DELETE (cancella da una tabella o *vista*).

Attraverso questi comandi, le cui istruzioni vengono scomposte dal DBMS in una serie di operazioni relazionali, l'amministratore è in grado di interagire con il database per strutturarlo e regolarne le modalità di utilizzo (http://www.polarnet.cnr.it/CORSI/LAMP/lamp.pdf).

1.3.3 Archivi di dati territoriali: tecnologie SIT/GIS

Tra le forme di organizzazione dei dati hanno assunto notevole importanza le cosiddette banche dati territoriali o *Sistemi Informativi Territoriali* (*SIT*, o *GIS* dall'inglese *Geographical Information System*), *software* progettati per digitalizzare, immagazzinare, manipolare, analizzare e visualizzare o stampare dati georeferenziati, cioè direttamente associati ad una precisa localizzazione geografica.

Lo sviluppo dei SIT/GIS trae origine dalla innovazione tecnica iniziata con l'avvento delle tecnologie *CAD* (*Computer Aided Design*), associate al telerilevamento ed alla aerofotogrammetria, che hanno prodotto un notevole progresso nei settori della cartografia, della pianificazione e della statistica territoriale, avvalendosi degli strumenti di acquisizione automatica di dati. I primi esempi di applicazione delle tecnologie SIT/GIS hanno riguardato la gestione delle risorse naturali, ma a questo settore se ne sono rapidamente aggiunti svariati altri: dai sistemi di navigazione di aerei, navi e automobili, alla gestione delle reti tecnologiche ed informatiche, alla pianificazione territoriale, ai sistemi di monitoraggio e di salvaguardia ambientale. Ampi archivi SIT/GIS sono oggi alla base della gestione amministrativa del territorio a qualsiasi livello (urbano, catastale, comunale, provinciale, regionale, …) e dei rispettivi servizi di protezione e prevenzione. Essi sono inoltre utilizzati in archeologia, scienze sociali, medicina ed in numerosi altri settori lontani dalle applicazioni per le quali furono inizialmente sviluppati (Callegari,1994).

Figura 2: Schema strutturale di un sistema SIT/GIS.
tratto da: http://www.proeco.it/gis/gis_cosa_dwn.htm

La tecnologia SIT/GIS amplifica notevolmente le possibilità di utilizzo della tradizionale cartografia consentendo non solo di informatizzare elementi geografici (come ad esempio: regioni, province, parcelle catastali, case, strade, fiumi, laghi, o qualsiasi altro oggetto geograficamente definibile e graficamente collocabile su una carta geografica) e di stamparli a qualsiasi scala, ma anche di associare a tali elementi svariati tipi di informazioni o dati, sia di tipo spaziale ("dati spaziali", quali: superficie, perimetro, latitudine e longitudine, quota, ecc.) che di altro genere ("dati alfanumerici") consentendo di associare univocamente, a precisi punti dello spazio, informazioni che ne descrivono proprietà e caratteristiche.

Figura 3: Esempio di struttura a stati separati (livelli) in un sistema CAD-GIS
tratto da: http://www.proeco.it/gis/gis_cosa_dwn.htm

Dal punto di vista delle capacità funzionali, un SIT/GIS è permette di svolgere tre fondamentali operazioni:

- input dei dati: acquisizione, memorizzazione, aggiornamento, *editing*;
- analisi dei dati: manipolazione e applicazione di metodologie numeriche di vario tipo (numeriche, statistiche, grafiche, etc.); in questa fase l'informazione contenuta nei dati da implicita diventa esplicita;
- output dei dati: restituzione delle elaborazioni svolte nelle fasi di input e analisi in forma grafica (carte geografiche), alfanumerica (tabelle, rapporti, etc.) o digitale (files di dati).

Queste operazioni costituiscono uno dei due sotto-modelli (quello applicativo) di una classificazione dei SIT/GIS, che si basa su parametri concettuali ben definiti (quali struttura, completezza, dettaglio e coerenza interna), il cui rispetto è essenziale per assicurare solide fondamenta al lavoro stesso di creazione di un prodotto di questo tipo. L'altro sotto-modello è quello astratto, secondo il quale il territorio viene ad essere rappresentato nei suoi attributi: posizionale (dato dalle coordinate spaziali: x, y e z), tematico (dato dai caratteri associati alle entità registrate) e temporale. (http://www.proeco.it/gis/gis_cosa_dwn.htm)

I SIT/GIS rispetto alla tradizionale cartografia hanno l'enorme vantaggio che, una volta collegati gli elementi geografici della rappresentazione digitale agli archivi associati, è possibile ricavare numerose indicazioni mediante operazioni di interrogazione (*query*), capaci di estrarre dai dati informazioni implicitamente presenti, ma difficilmente ricavabili in modo immediato. Ciò viene ottenuto essenzialmente attraverso le abilità modellistiche di procedure statistiche e della ricerca operativa.
In generale un sistema SIT/GIS può essere "interrogato" riguardo a:

- posizione (es: “Quali elementi territoriali si riscontrano in corrispondenza di determinate coordinate geografiche?”, “Quale distanza intercorre tra due definiti elementi geografici”, “Dove è ubicato l’elemento che presenta determinati attributi?” ….). Tra le query di posizione ricadono anche le così dette *analisi di* “*buffering*” che consistono nella creazione di una *zona di rispetto* attorno a elementi con particolari caratteristiche (es: in caso di vincolo idrogeologico, è possibile effettuare query di questo tipo lungo un percorso fluviale e visualizzare tutti quei manufatti che rientrano all’interno dell’area di rispetto del fiume).
- tempo e spazio (es: “Quali sono quegli elementi spaziali la cui estensione è maggiore di 2 ettari?”, “Qual è il percorso più breve tra due punti stabiliti?” ...),
- andamento o *trend* (es: “Quali elementi del territorio hanno subito modifiche in un determinato arco di tempo?” ...);

Inoltre è possibile:

- porre interrogazioni riguardanti i dati collegati agli oggetti (es. “Quale regione ha avuto il maggior numero di nascite in un determinato anno?”);
- porre interrogazioni complesse, cioè composte da due o più interrogazioni: (es: “Quali, tra le città costiere collocate al di sotto di una determinata latitudine ed aventi un porto ittico, hanno avuto un pescato superiore a tot tonnellate nell’anno...?”).

Tutte le informazioni inserite nel sistema possono anche essere “incrociate” tra loro per ottenere nuove informazioni ovvero nuovi *tematismi*. Questa operazione viene chiamata *overlay*. Un *tematismo* (o tema) è il risultato di un’interrogazione che seleziona ed evidenzia dati o rapporti tra dati. Ad esempio consideriamo un carattere, che potremmo chiamare “mare”, rappresentato nella carta digitale come aggregato di poligoni adiacenti che ricoprono un’area corrispondente alla distribuzione geografica della superficie marina di una determinata zona; se a tali poligoni sono associati i valori di una qualche proprietà, ipotizziamo sia la profondità del fondale, è anche possibile raggrupparli secondo differenti “classi di profondità” e creare così una “mappa tematica”, dove tutte le diverse categorie compaiono distinte per colore (ad esempio con tonalità di blu differenti, più scure al crescere della profondità). Più temi, ottenuti dalla elaborazione di dati ascrivibili a proprietà differenti, possono poi essere richiamati simultaneamente per ricavare informazioni aggregate.

Il recente sviluppo di apparati di misura automatizzati (reti di rilevamento e sistemi di sensori remoti), impiegati per il monitoraggio ambientale, è responsabile di un aumento vertiginoso, in ampiezza e numero, delle raccolte di dati archiviati ed amministrati tramite SIT/GIS. Si tratta di procedure altamente funzionali per le operazioni di controllo, le valutazioni di impatto e la gestione delle risorse territoriali, ma difficilmente esportabili all’esterno delle strutture in cui vengono realizzate a causa della notevole quantità di memoria richiesta da tali sistemi per immagazzinare i dati, per compiere le elaborazioni e per sostenere le elevate dimensioni dei file vettoriali, il che le rende ancora ingestibili via Internet.