

Strutturazione dell'informazione e integrazione della conoscenza

ANNA ROVELLA, GIOVANNI MARRÈ

This article reports the results of a survey and an experiment conducted in collaboration between ItConsult and the Laboratorio di Documentazione at the University of Calabria in order to determine the effects of the integration of a document management system in a knowledge Management Platform. The work aims at the extraction of information from unstructured documents and their content management.

Keywords: Knowledge – Information – BMP – Document Management System

Il rapporto tra informazione e conoscenza è – da sempre – oggetto di studio e di riflessione in un contesto multidisciplinare che spazia dall'epistemologia all'informatica, all'economia e alle scienze cognitive, per citarne solo alcuni. La molteplicità di approcci con cui tale relazione è stata affrontata ha determinato la produzione di differenti definizioni finalizzate a delimitare i confini concettuali tra l'uno e l'altro termine [1] e ha, in maniera diversa, influenzato le realizzazioni pratiche conseguenti. Se nel sentire comune informazione e conoscenza sono, non di rado, utilizzate come sinonimi, la differenza che le separa sotto il profilo semantico è stata, per contro, formalizzata in modo evidente. «Quando si cerca di caratterizzare l'attività intellettuale i termini che vengono in mente sono numerosi. A priori, possono essere presi in considerazione tre livelli di analisi:

1. la conoscenza, potremmo dire la comprensione, è il livello superiore, quello della comprensione dei sistemi;
2. il sapere, che è una conoscenza operativa, un *savoir-faire*, un'attitudine senza dubbio, anche un saper essere un saper vivere, ecc.;
3. l'informazione, che è ciò attraverso cui la conoscenza e il sapere possono scambiarsi, ma anche, e sempre più essere prodotti» [2].

In questa tripartizione gerarchica dei tre livelli di acquisizione cognitiva l'informazione si qualifica come momento di iniziale riduzione dell'incertezza e di *input* del processo mentre la conoscenza diventa anche il momento di ripensamento critico e di acquisizione al patrimonio individuale e collettivo di nuovi descrittori della realtà esperienziale, utili alla definizione di strategie decisionali, organizzative e operative. Si delinea, quindi, un processo concettualmente strutturato all'interno del quale i diversi momenti postulano concretizzazioni pratiche definite e strettamente conseguenti. A

fronte di tutto ciò, aziende e pubbliche amministrazioni evidenziano sostanziali difficoltà nella realizzazione di processi di integrazione tra i diversi elementi della catena della conoscenza, tanto da spingere quasi tutte le realtà imprenditoriali del settore a misurarsi con lo specifico problema coniando anche un apposito acronimo: CMIS (*Content Management Interoperability Services*) [3]. Archivi cartacei ed elettronici, sistemi informatici disomogenei, dati strutturati e documenti tradiscono il risultato di una gestione del capitale cognitivo priva di un disegno organico di sviluppo, tesa a privilegiare, di volta in volta, aspetti particolari e segmentati, se non la singola applicazione tecnologica. L'informazione, chiusa all'interno di circuiti ristretti e limitativi, imbocca percorsi ripetitivi e stagnanti che ne inibiscono fortemente la potenziale azione di diffusione e condivisione. Molto spesso i dati, quali elemento primario e grezzo, vengono immagazzinati in maniera quasi compulsiva, senza un necessario riferimento a processi di sistematizzazione e le informazioni derivanti, spesso eccessivamente frammentate, lungi dall'agevolare il processo decisionale ne generano un sostanziale rallentamento. «Informazione e ignoranza, scelta, previsione e incertezza, sono tutte intimamente correlate (...) Al confine della completa conoscenza e della completa ignoranza, sembra intuitivamente ragionevole parlare di gradi di incertezza. Più vasta è la scelta, più esteso è l'insieme delle alternative che si aprono davanti a noi, più incerti noi siamo circa come procedere e di maggiore informazione abbiamo bisogno per prendere la nostra decisione» [4]. L'accumulo di dati e di informazioni non gestite, e la mancata evoluzione ed integrazione tra informazioni e conoscenza, producono disorientamento e attivano un'azione di rallentamento o addirittura paralizzano le capacità decisionali e strategiche delle organizzazioni in una paradossale confutazione dell'assunto di Shannon: pur in presenza di una bassa entropia positiva non aumenta il supporto alla decisione [5].

La conoscenza rifugge dalla frammentazione: essa consegue e consolida il suo valore mettendo a sistema le informazioni e strutturandone organicamente le unità costitutive mediante un'accurata operazione di definizione dei concetti cui fa seguito l'integrazione di contenuti e di forma nel processo di comunicazione che si occupa di socializzarle. L'informazione, così elaborata, è in grado di offrire una nuova prospettiva di interpretazione di eventi e oggetti, lasciando cogliere significati in precedenza nascosti e evidenziando relazioni inattese. Ed è proprio in questo fluido compenetrarsi e comparteciparsi che essa diventa un fattore di mediazione, elemento necessario a produrre e costruire dinamicamente nuova conoscenza attraverso la ristrutturazione e l'integrazione di nuove valenze. Tuttavia, senza voler qui ulteriormente approfondire la complessa disquisizione terminologica e le problematiche connesse alla formalizzazione ed utilizzazione della conoscenza, vorremmo limitare la nostra riflessione ad un caso determinato e specifico: la possibile integrazione tra gestione dei documenti e gestione dei contenuti all'interno di *workflow* di *Business Process Management* (BPM). Il contesto tecnologico di implementazione e sperimentazione è "josh Protocol!" [6], piattaforma di gestione della co-

noscenza, i cui componenti di protocollo e gestione documentale sono frutto di una collaborazione tra l'azienda produttrice e il Laboratorio di Documentazione dell'Università della Calabria [7].

Per loro stessa natura, i *workflow* di BPM si configurano come strumenti di ricerca attivi non solo sull'intero patrimonio documentale ma anche in ragione delle relazioni che gli stessi intrattengono con i dati e le informazioni provenienti dalle fasi dei *workflow* di processo. I BPM, infatti, nascono e vengono normalmente elaborati nella fiduciosa aspettativa di poter fornire risposte concrete all'esigenza di interoperabilità tra sistemi informativi esistenti e specifici applicativi progettati per l'accesso e la distribuzione dei dati. Tuttavia, malgrado gli sforzi effettuati, le tecnologie messe a punto per la classificazione delle informazioni e la gestione dei contenuti agiscono, di fatto, prevalentemente su dati strutturati, mentre l'integrazione della conoscenza destrutturata presente all'interno dei processi richiede ancora il prevalente ricorso ad attività manuali, vincolate all'utilizzo di personale specificatamente qualificato con conseguente dispendio di tempo e incremento dei costi. In tale prospettiva, la definizione concettuale di regole e la realizzazione di applicativi adattivi capaci di utilizzare le risorse terminologiche per generare descrittori informativi da testi non strutturati mediante procedure di analisi concettuale, assume una sicura rilevanza teorica ed operativa [8]. Particolarmente evidente – dopo un primo periodo d'uso – è l'incremento dei benefici che l'adozione di simili strumenti comporta in termini di contenimento dei costi, riduzione sostanziale dei tempi di ricerca e di recupero delle informazioni. Al più tradizionale approccio della strutturazione dei dati destrutturati si sostituisce la modellizzazione concettuale di *features* capaci di descrivere comunque modelli documentali diversificati contestualizzando la rilevanza dei termini estratti come possibili descrittori [9]. In questa prospettiva, nella quale la virtualizzazione dei supporti affievolisce la significanza dei legami organici tra gli atti di uno stesso complesso, classificazione e indicizzazione si configurano come delle vere e proprie chiavi di accesso alle varie fasi che accompagnano il ciclo vitale del documento, dalla sua produzione, alla selezione, alla conservazione temporanea e permanente, alla consultabilità e fruibilità.

Il modello è un sistema di accesso all'informazione, costituito da un'integrazione tra tecniche di *information retrieval* e *ontology engineering*, combinate mediante un approccio *knowledge-based*, concettuale e automatico-statistico. In particolare, la base di conoscenza del sistema è strutturata su uno schema concettuale elaborato mediante la combinazione di dati estratti dall'organigramma e dal titolario di classificazione associati a tipologie documentali a loro volta relazionate alle norme e ai regolamenti in modo da permettere la strutturazione e la definizione delle *query*.

Il modello progettato consente, nel dettaglio, di effettuare:

- la ricerca di informazione estratta mediante analisi semantica, automatica e semi-automatica, dei testi, condotta su basi di dati testuali non strutturate;

- la ricerca di documenti attraverso tecniche avanzate di interazione con l'utente, (*profiling*, espansione semantica delle *query*, *relevance feedback*, *clustering* dei documenti);
- la valutazione della *performance* del motore di ricerca e sua ottimizzazione attraverso l'analisi dei flussi di documenti e dei processi di *business*;
- la costruzione di un'ontologia dell'organizzazione in grado di guidare il processo di ricerca delle informazioni tramite la definizione di classi, relazioni, e strutture informative rilevanti;
- l'estrazione di informazione e l'identificazione di contenuti informativi specifici all'interno dei documenti e del flusso;
- il trattamento automatico del testo per l'identificazione di occorrenze di informazione strutturata nei documenti;
- la costruzione di strumenti di annotazione automatica dei documenti per la strutturazione di relazioni tassonomiche e l'implementazione di *corpora* terminologici.

Il modello concettuale qui sommariamente delineato, applicato ad amministrazioni con strutture decisionali definite ma con produzione documentale non formalizzata, potrebbe produrre vantaggi operativi anche in tempi estremamente brevi aumentando considerevolmente il supporto informativo e la conseguente capacità di scelta dei decisori.

In ogni caso l'integrazione di metodologie e di strumenti tecnologici per l'automazione delle operazioni di classificazione e di indicizzazione all'interno di piattaforme di gestione delle conoscenze è uno stimolante terreno di confronto per l'elaborazione di esperienze e studio di casi la cui riproducibilità è in grado di apportare benefici, ad ampio raggio, in ogni organizzazione sia sul piano organizzativo sia su quello delle *performance*.

Non sono da sottovalutare, tuttavia, le criticità che il modello ancora presenta relativamente agli aspetti linguistico terminologici determinati dall'analisi dei contenuti applicata a contesti fortemente eterogenei e dalla necessità di usare una pluralità di vocabolari di dominio nell'interazione dei quali emergono significative problematiche di gestione delle sinonimie e di definizione delle relazioni, specie in contesti vocati ad una naturale multilinguismo [10]

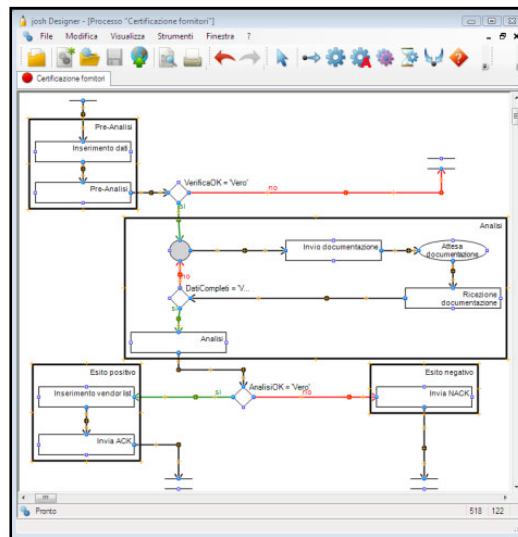
Il caso applicativo.

Quello della gestione dei dati strutturati è stato tradizionalmente considerato un mondo separato da quello dei dati non strutturati. Il primo, fatto essenzialmente di *database*, dispone di tecnologie e metodologie consolidate e stabili, sia pure in evoluzione; il secondo, fatto soprattutto di documenti, non solo testuali, e anche da processi e altri oggetti è, viceversa, sostanzialmente privo di *standard*, presenta problemi meto-

dologici e di interoperabilità. L'interpretazione, l'elaborazione, la semplice estrazione di dati da documenti – intesi in senso lato – è una operazione tuttora particolarmente complessa e difficilmente generalizzabile.

Le tecnologie di *Business Process Management* (BPM) si possono intendere ed utilizzare come un ponte fra i due mondi. Ad esempio, la piattaforma *josh* dispone di un *Workflow Management System* (WFMS) che:

- consente la descrizione semi-formale dei processi con un linguaggio grafico di modellazione [11],
- manda in esecuzione tali diagrammi, chiamando direttamente in causa le persone coinvolte nel processo, attraverso il loro *browser* Internet.



Si tratta di una tecnologia che costituisce una efficace modalità di rappresentazione della conoscenza incorporata nei processi ma anche di un meccanismo di astrazione e disaccoppiamento dei dati dai processi, che però consente di associare gli uni agli altri, ad esempio associando metadati (strutturati) ai documenti (non strutturati).

Il modo tradizionale di sviluppare quelle applicazioni che necessitano, durante l'esecuzione, dell'intervento successivo di diversi operatori, implica l'incorporazione del flusso di azioni all'interno del codice sorgente con, nella migliore delle ipotesi, una parziale configurabilità di opzioni attraverso impostazioni utente. In pratica il progettista/sviluppatore codifica sia la logica dell'applicazione sia le singole *form* che determinano l'interazione dell'applicazione stessa con l'utente.

Ciò che, invece, un sistema *software* di BPM permette, è di disaccoppiare la logica, che viene disegnata graficamente, dalle *form* che possono essere generate automaticamente o continuare ad essere sviluppate con linguaggi tradizionali, ma che comunque beneficiano della parcellizzazione e modularizzazione dello sviluppo di singole componenti di gran lunga più piccole e più semplici di una grande applicazione monolitica che, peraltro, coinvolgendo una elevata quantità utenti, deve spesso possedere elevati requisiti di solidità e scalabilità.

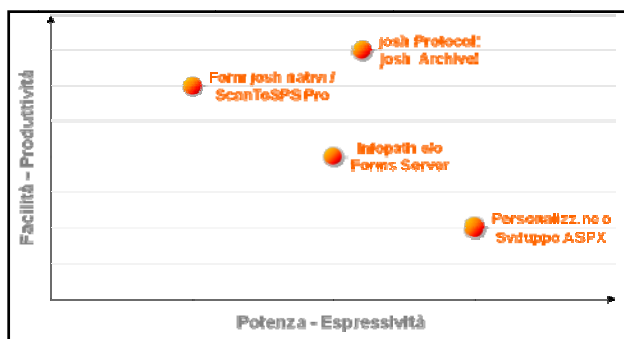
Questa modalità operativa configura un nuovo modo di costruire applicazioni *software*, in cui i processi da analizzare ed automatizzare trovano collocazione nell'ambito di processi generali di un livello di astrazione più elevato, per i quali è l'esperto di dominio applicativo che analizza e definisce quel flusso delle azioni (il *workflow*) che poi, contestualmente o successivamente viene formalizzato in forma grafica, utilizzando lo specifico linguaggio dello strumento software. Tipicamente, una volta descritto un processo in termini di *workflow*, è attraverso successivi passi di raffinamento (*task activity*) che si giunge a definire il dettaglio dei singoli *task*. Il *task* è un'unità elementare di lavoro all'interno di un processo ed è eseguita da un singolo *attore* (tipicamente umano, ma talvolta automatico). Alcune *task activity* consistono nella visualizzazione e/o modifica di dati strutturati collocati su *database* e manipolati attraverso delle *form*.

The screenshot shows a web browser window with the address bar containing a URL. The page content includes a header with the 'UBAE' logo and a title 'UBAE - Istruttoria'. Below the header, there is a form titled 'Dati Richiesta' with several input fields and dropdown menus. The fields are: 'Codice NDG' (text input with value '1234566'), 'Società Richiedente' (text input with value 'TEST'), 'Paese' (text input with value 'TEST'), 'Tipologia Cliente' (dropdown menu with value 'Cliente Corporate'), 'Cliente già affidato' (checkbox with value 'No'), 'Cliente con garanzie' (checkbox with value 'No'), 'Pratica in via d'urgenza (per le vie brevi)' (checkbox with value 'No'), and 'Pratica per Proroghe Garanzia' (checkbox with value 'No'). At the bottom of the form, there are 'Save' and 'Back' buttons. The browser's status bar at the bottom shows 'Done' and 'Trusted sites'.

In josh la costruzione delle *form* di accesso ai *database* si può realizzare in diverse modalità, che hanno un livello decrescente di facilità d'uso (produttività) e parallelamente crescente in termini di espressività (potenza), che sono:

- *form* di josh nativi generati attraverso un *wizard* [12] da cui si scelgono le variabili di processo da editare o visualizzare sul *browser* Internet;
- Microsoft Infopath e/o Forms Server; lo strumento Microsoft per generare e pubblicare *form* che debbono interagire con josh attraverso una modesta attività di sviluppo;
- personalizzazione o sviluppo di *task activity* e di *form*, in ambo i casi con una vera e propria programmazione in ASP.NET;

È opportuno precisare che in josh esistono alcune applicazioni verticali che, oltre a specializzare josh e SharePoint nella risoluzione di specifiche tematiche, gestiscono alcuni dati strutturati *out-of-the-box*. È il caso, ad esempio, in josh Protocol!, dei dati di protocollo dei singoli documenti protocollati registrati e archiviati.



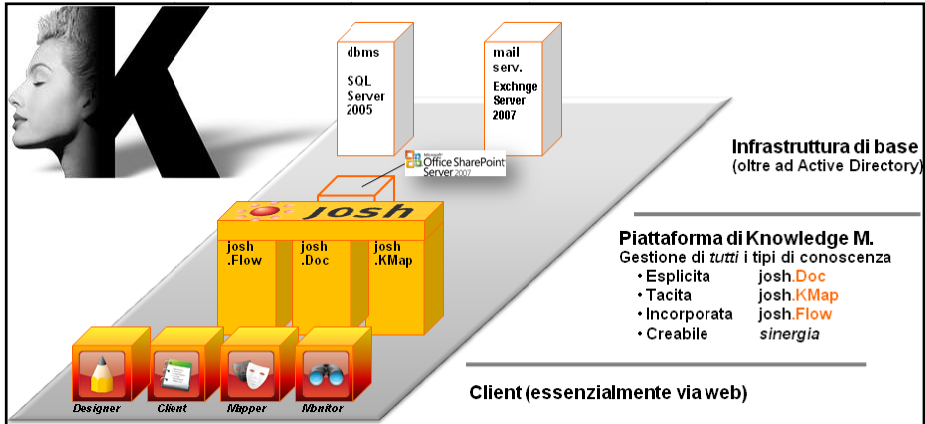
Josh è una piattaforma *software* di *Knowledge Management* (Gestione della Conoscenza Organizzativa - KM) di tipo *enterprise* che consta di tre moduli principali, ciascuno corrispondente ad una tipologia di conoscenza gestita, e, in particolare:

- a) Josh.Doc: un sistema di gestione dei documenti formalizzati;
- b) Josh.KMap: un sistema di mappatura delle competenze individuali e per la formalizzazione della gestione della conoscenza tacita;
- c) Josh.Flow: un *workflow management* per gestire la conoscenza Incorporata nei processi.

Su questa piattaforma si innestano le componenti verticali come joshArchive! (Archiviazione e Conservazione Sostitutiva) e joshProtocol! (Protocollo Informatico) come anche altre applicazioni *ad hoc* che di volta in volta possono essere sviluppate.

Questo approccio comporta una serie di benefici. In primo luogo il fatto che la logica con cui si sviluppano queste “nuove” applicazioni è direttamente presidiata dall’esperto del dominio applicativo e non più dal solo sviluppatore; conseguentemente le

modifiche che consentono la reale riconfigurabilità dei processi aziendali in maniera reattiva divengono rapidissime ed economiche oltre che sicuramente pertinenti.



Infine, non meno importante, il fatto che il sistema BPM consente di migliorare l'efficacia delle ricerche e la produttività individuale, rilevando quali documenti vengono maggiormente consultati durante i *task* di un processo in modo tale di poter successivamente proporre agli utenti i documenti più utili in quel contesto specifico (perché già usati dai colleghi) e/o classificarli automaticamente in base al loro utilizzo, aggiungendo le parole chiave legate ai *task* in cui vengono impiegati, i quali si configurano – di fatto – come dei *training set*. Su quest'area sono possibili numerose migliorie, coerentemente al modello presentato nella prima parte dell'articolo.

In conclusione, l'esperienza di sinergia realizzata è stata sicuramente positiva ed ha portato alla realizzazione di un prodotto commerciale che rappresenta un primo importante tassello concettuale verso l'integrazione dei sistemi e l'estrazione di conoscenza da testi non strutturati.

Note

- [1] «Prima di entrare nel vivo della nostra teoria, è bene chiarire con tre osservazioni, somiglianze e differenze fra i concetti di conoscenza e di informazione. La prima osservazione è che la conoscenza, diversamente dall'informazione, concerne le credenze e il coinvolgimento. È cioè funzione del punto di vista, della prospettiva o dell'intenzione del singolo. La seconda osservazione è che la conoscenza, diversamente dall'informazione, riguarda l'azione. È sempre diretta a un fine. La terza osservazione è che la conoscenza, come

- l'informazione, concerne significati; è specifica del contesto e relazionale». Ikujiro Nonaka, Hirotaka Takeuchi, *The knowledge creating company*, Oxford University Press, 1995, trad. it. 1997, p. 94.
- [2] Laurent Gille, *La protezione della proprietà intellettuale fattore della divisione internazionale della conoscenza*, in: Antonio Pilati, Antonio Perucci (a cura di), *Economia della Conoscenza. Profili teorici ed evidenze empiriche*, Il Mulino, Bologna, 2005, p. 211.
- [3] «Sous l'acronyme de CMIS sont regroupées les spécifications d'une interface, des techniques, un langage commun et des protocoles qui doivent permettre de développer des moyens de consultation et d'échanges d'objets (documents, fichiers) entre les référentiels de plusieurs logiciels de GEIDE ou d'ECM». *CMIS: une interface de services pour l'interopérabilité entre des solutions de gestion de contenu ou de GEIDE*, in: "MOS", n. 251, settembre 2008, p. 5. Cfr. Anche Alain Garnier, *L'information non structurée dans l'entreprise*, Lavoisier, Paris, 2007.
- [4] Sigmund Koch, *Information Theory*, in: *Psychology: A Study of a Science*, 1959, pp. 614-615.
- [5] Nel 1948 Claude E. Shannon pubblicava *A mathematical theory of communication* ("Bell System Technical Journal", vol. 27, pp. 379-423 and 623-656, July and October, 1948) gettando le basi della teoria dell'informazione. Egli partiva dall'idea che un messaggio inviato attraverso un qualsiasi canale subisce nel corso della trasmissione deformazioni diverse per cui al suo arrivo esso ha perso una parte delle informazioni che conteneva originariamente. Egli stabilì quindi una correlazione tra tale perdita di informazioni e l'entropia, ovvero la funzione matematica che esprime la degradazione dell'energia che si verifica in ogni trasformazione del lavoro meccanico in calore, in quanto la trasformazione inversa dal calore al lavoro meccanico non risulta mai completa. In base a questa analogia la quantità di informazione trasmessa può essere calcolata come entropia negativa giacché nella trasmissione dei messaggi come nella trasformazione dell'energia, l'entropia negativa decresce continuamente in quanto quella positiva (perdita di informazione o degradazione di energia) cresce continuamente.
- [6] L'azienda produttrice è ItConsult di Fermignano (Urbino).
- [7] Referente di progetto per l'Università della Calabria è stata la prof. Anna Rovella, coautrice del presente testo.
- [8] «The link between terminology and cognitive science is created by concepts. Concepts are units constituting the basis of knowledge and concept systems describe the way each field organizes knowledge. In this sense, the theory of terminology and the theory of knowledge are closely related». Maria Teresa Cabré, *Terminology. Theory, methods and applications*, Amsterdam, 1998, p. 52.
- [9] Cfr. C. Beghtol, *Semantic Validity: concepts of warrant in bibliographic classification systems*, in: "Library resources and technical services", 30 (1986), n. 2, pp. 109-125, nonché Alberto Cheti, *Le categorie nell'indicizzazione. Indagine su alcuni modelli di analisi e di organizzazione concettuale*, in: "Biblioteche oggi", n. 8 (1990) n. 1, pp. 29-49.
- [10] «Lo sforzo di determinazione degli elementi costitutivi di una terminologia, dei valori semantici dei termini delle regole di combinazione accettate è lo sforzo di conferire a parti

del linguaggio verbale le caratteristiche dei codici più semplici e dei calcoli. Costruire un linguaggio formale significa costruire un'area d'uso della lingua in cui ogni discorso sia un testo, ogni comprensione un processo di interpretazione certa e a termine. Un linguaggio speciale, tanto più se le sue regole costitutive sono esplicitate ed esso sia dunque formalizzato, è per così dire un tentativo di servirsi dei materiali della lingua per uscire fuori dalla storia, fuori dalla durata e dalla fluttuante massa parlante, per arrivare a costruire testi valevoli oltre il tempo e la contingenza in cui dapprima si produssero». Tullio De Mauro, *Minisemantica*, Laterza, Bari, 1982, pp. 148-149.

- [11] In particolare, josh utilizza l'evoluzione di un linguaggio di modellazione denominato WIDE (*Workflow on Intelligent Distributed database Environment*), frutto di un progetto di ricerca ESPRIT, cofinanziato dalla Commissione Europea, che ha avuto come *partner* accademici il Politecnico di Milano e la University of Twente (NL). Il progetto è descritto in Paul Grefen, Barbara Pernici, Gabriel Sanchez, *Database support for Workflow Management - The WIDE Project*, Kluwer Academic Publishers, 1999.
- [12] Ossia un'autocomposizione, in cui l'utente è assistito a passo a passo.