***Disambiguazione di autori AGRIS: Identificare Entità all’interno di sorgenti dati incomplete***

La definizione di entità come elemento distinto ed univocamente identificato è una delle basi del web semantico e del progetto Okkam. La forte distinzione permette un nuovo orizzonte per le discipline di data retrieval e data integration, a patto di venire applicate con coerenza e consistenza di informazioni; senza tale presupposto questi sforzi in ambito di ricerca non avrebbero senso. E’ intuitivo ed immediato comprendere questi concetti, ma è altrettanto facile perderedi vista la rigorosità con cui essi devono essere seguiti nel realizzare applicazioni e sistemi che si basano su essi. In particolare, l’ambito che riguarda l’integrazione di dati, (records ricavati da sorgenti con origine, schemi e piattaforme differenti) richiede una forte sensibilità ed elasticità, a causa di informazioni che potrebbero risultare incomplete, ambigue, o addirittura errate.

Affrontando il progetto “OKKAM meet FAO” è emerso un obiettivo inerente all’identificazione di entità all’interno di sorgenti dati. Le sorgenti dati sono dei file XML contenenti metadati che riguardano articoli scientifici contenuti nel repositorio AGRIS. Tra questi metadati vi sono informazioni sugli autori degli articoli, individui per i quali vorremmo creare le relative entità che li identifichino nell’ENS. Tuttavia tali informazioni si presentano incomplete: i dati a nostra disposizione sono Cognome, Iniziale puntata del nome, spesso ma non sempre Nome dell’Affiliazione del soggetto autore. Tale incompletezza non ci consente di avere sufficienti informazioni per creare un’entità sull’ENS; ne risulterebbero delle entità ambigue e non-identificative. Una possibile soluzione consiste nell’affrontare questo problema manualmente, avvalendosi di risorse esterne per integrare le informazioni. Da tale processo manuale potrebbero emergere idee e ipotesi per la comprensione dei ragionamenti che stanno alla base di una ricerca di informazioni; la sequenza di passi seguiti per ovviare alla povertà di informazioni delinea un ragionamento logico compiuto da un esecutore umano.

Riferendosi al processo utilizzato per analizzare i dati in nostro possesso possiamo dire che l’approccio è stato naturale. Le fasi principali che lo caratterizzano sono due: la prima, una spontanea e intuitiva ricerca all’interno dei dati, la cui esecuzione non richiede nessuna nozione scientifica particolare, una semplice sequenza che comprende:

1. osservazione dei dati;
2. acquisizione di consapevolezza del problema;
3. deduzione di correlazione tra i vari autori;
4. intuizione riguardante le possibili soluzioni;
5. ragionamento delle nuove informazioni localizzate;
6. analisi del risultato.

tali passaggi possono apparire come un banale processo di lettura e studio, la loro ovvietà deriva dal fatto che l’esecutore è un essere umano senziente e con una certa cultura. Lo sforzo successivo a questa meta-analisi è la comprensione dell’algoritmo decisionale che sta alla sua radice, la formalizzazione del processo che il ragionamento umano compie affrontando un problema di questo tipo; quali deduzioni spingono verso una decisione piuttosto che un’altra.

La seconda fase, la formalizzazione del processo svolto in prima analisi, cerca di scindere ogni operazione in una serie di deduzioni logiche che possano essere meglio scomposte e valutate atomicamente. Per questo obiettivo vi è una sequenza di operazioni simile alla fase precedente, caratterizzati da più rigore nel metodo e nella valutazione:

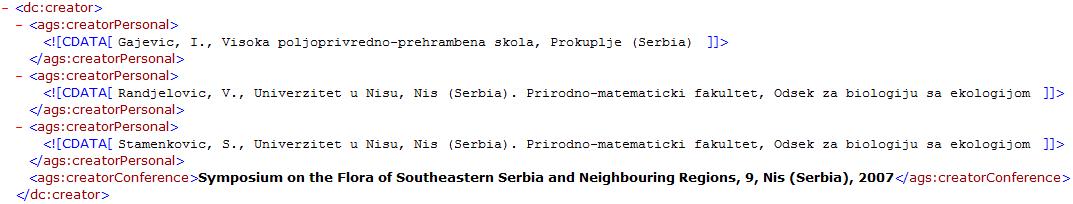
1. catalogazione dei dati;
2. individuazione degli attributi mancanti per la distinzione dell’entità;
3. messa in evidenza delle relazione di co-authoring e affiliazioni;
4. utilizzo di risorse web per ottenere informazioni aggiuntive sulle entità;
5. valutazione delle nuove informazioni localizzate, in termini di coerenza, consistenza, rilevanza;
6. formalizzazione del processo utilizzato.

La rilevante differenza che si nota ad una prima osservazione riguarda la profondità dell’analisi svolta: le due fasi hanno diversità sostanziali come per una poesia la sua mera lettura e la sua parafrasi, tuttavia sono l’una mezzo dell’altra (non è possibile parafrasare una poesia senza leggerla!!!). Si tratta di un approccio top-down alle informazioni contenute nei file di metadati XML di FAO: una prima osservazione generale, che genera nella mente umana (il reasoner) una serie di valutazioni più o meno corrette, una schematizzazione dei dati (capacità astrattiva della mente umana) ed infine una valutazione delle soluzioni in base all’obiettivo. In seguito alla “lettura” si procede alla fase di “parafrasi”, dove si cerca di scindere il reasoning della mente umana in operazioni più vicine ad un approccio algoritmico.

Esistono vari web-site che catalogano autori, delle più svariate discipline e campi di ricerca. Gli autori di AGRIS sono principalmente ricercatori che svolgono il loro lavoro attorno alle scienze agrarie, e AGRIS stesso figura come uno dei portali più completi per ricerca articoli e autori di questo campo. Il nostro obiettivo è quello di utilizzare tali risorse in Internet per estrarre informazioni. La ricerca nel campo agrario coinvolge molte altre discipline (ingegneria, informatica, biologia, chimica, etc), ed è per questo motivo che gli autori coinvolti potrebbero essere catalogati in siti di svariata origine.

**Catalogazione dei dati.**

I files XML del repositorio AGRIS contengono metadati riguardanti gli articoli di scienza agraria contenuti in esso. Questi files si prestano all’obiettivo di identificare entità in una sorgente di dati incompleta, essi contengono infatti informazioni incomplete (o considerati tali per il nostro scopo) degli autori di tali articoli scientifici. Una porzione di ipertesto contenente tale parte di dati compare in ***Fig.1***.



***Fig 1.*** *Parte di testo in sintassi XML contenente metadati su alcuni autori di journals del repositorio AGRIS*

Si evincono dall’immagine di esempio le particolarità delle informazioni a nostra disposizione:

* Il Linguaggio utilizzato è spesso differente dall’inglese;
* Cognome dell’autore specificato, Nome Proprio abbreviato alla prima lettera puntata;
* Affiliazione dell’autore espressa in maniera parziale, o addirittura del tutto assente.

Tali informazioni posso risultare sufficienti per un lettore che abbia conoscenza dell’affiliazione o del centro dove è stato pubblicato l’articolo, possono essere soddisfacenti per un utente casuale che intendesse prendere visione sommaria dell’articolo, ma sono assolutamente insufficienti per creare un’entità univocamente e distintivamente legata a quella persona. Successivamente, si può dire che senza la possibilità di individuare tali persone come entità distinte non è possibile nemmeno creare una rete di entità in relazione tra loro (documenti, articoli, autori, affiliazioni, centri di ricerca); un limite enorme alla rete di conoscenza navigabile che si vuole creare.

Ogni particolarità dei nostri dati pone una problematica: allo scopo di avere maggiori informazioni, navigazione di siti web in lingua straniera, con le relative difficoltà di comprensione; casi di omonimia ambigui e fuorvianti, riscontrabili in web-sites di differente origine; ambiguità del nome dell’affiliazione, e, di nuovo, difficoltà nel caso l’affiliazione avesse un web-site in lingua straniera.

Tuttavia discuteremo come affrontare questi casi successivamente, la catalogazione avviene prima. Nel nostro caso i file XML di AGRIS sono divisi per anni, poi mesi e ancora per nazioni/centri. Questa preventiva catalogazione aiuta il raggiungimento del nostro obiettivo, in quanto è possibile analizzare un file alla volta, con la garanzia che all’interno di essi gli autori citati lavorarenno per la maggior parte nella stessa nazione, con la conseguenza che le relazioni di co-authoring siano più probabili. Prendiamo ad esempio il file X , catalogato nell’anno 2009, mese di gennaio. I metadati riguardanti autori all’interno di tale file descrivono solo ed esclusivamente persone che hanno per lo meno lavorato in quella nazione.

Per procedere con la catalogazione possono essere utilizzati due metodi, a seconda degli utilizzi futuri che si prospettano per queste informazioni: foglio elettronico o tabella in un database. Si tratta semplicemente di sfogliare manualmente i metadati e completare i record di una tabella con i seguenti campi: Cognome, Nome, Affiliazione, Publisher, E-Mail, Web-Site, Co-Authoring, Note Aggiuntive. L’unico campo che avrà sempre un valore è Cognome, mentre gli altri saranno incompleti o spesso mancanti. Il campo “Co-Authoring” conterrà un’informazione riguardante il rapporto di collaborazione con altri autori catalogati nello stesso file; per esempio nel l’articolo F, di cui sono autori Persona A, Persona B, Persona C, potremo dire che A è co-autore di B e C, B è co-autore di A e C, e C è co-autore di A e B (proprietà transitiva). Nei processi di analisi successivi alla catalogazione esprimeremo come questa informazione si rivela fondamentale.

**Individuazione degli attributi mancanti per la distinzione dell’entità.**

La corretta creazione di un’entità di tipo Persona nell’Entity Name System richiede un certo numero e tipo di attributi, i quali devono costituire un insieme di attributi che renda l’entità 1. Coerente col mondo reale, 2. Univocamente distinguibile, 3. Sintatticamente e semanticamente corretta. Nel paragrafo dedicato al concetto di “entità distinta” si è già ampiamente parlato del significato di queste tre caratteristiche, per cui non ci dilungheremo oltre. Tuttavia è il caso di evidenziare cosa manca al nostro insieme di dati per giungere a tali presupposti. Le informazioni che troviamo nei metadati in XML sono le stesse che possono essere riscontrate nelle intestazioni dei paper del repositorio AGRIS: cognome, iniziale del nome, talvolta università o centro di ricerca per cui il soggetto lavora. E’ intuibile, ma anche facilmente verificabile sul web, che tali informazioni non sono sufficientemente distintive, nonostante si possa essere piuttosto sicuri della loro coerenza con il mondo reale. Ad esempio possiamo dire che il signor Zlatkovic B., ricercatore della “University of Nis (Serbia)” è PROBABILMENTE l’unico con tali cognome e iniziale del nome in quella università, ma è un’ipotesi che ha origine da un’intuizione umana, non ha nulla a che vedere con un ragionamento logico che una macchina potrebbe capire. Ne consegue che un’informazione così incompleta non è sufficiente a decidere che tale individuo è univocamente definito dal suo cognome, iniziale nome e affiliazione; in termini di basi di dati questi attributi non rappresentano una chiave primaria, almeno nelle condizioni in cui si presentano.

Possiamo decidere un set di attributi minimi per definire un’entità distinta, ma per concludere quali essi siano bisogna accettare un compromesso: alcuni dati, particolarmente personali, non è possibile rinvenirli tramite la rete o tramite risorse pubbliche. In altre parole dati sensibili come codice fiscale, numero di previdenza sociale, o particolarmente esplicativi come data di nascita e indirizzo di residenza, non è plausibile sperare di poterli rinvenire con facilità. La qualità di distintività delle entità sarebbe particolarmente elevata, tuttavia si ottiene un equivalente risultato con le risorse a disposizione sul web. L’insieme di attributi individuato è il seguente, attenendosi alla nomenclatura dell’ontologia di Okkam (ENS Core Vocabulary):

(first\_name , last\_name , affiliation , country )

L’insieme composto da questi quattro attributi (in cui il nome compare per esteso) rappresenta un insieme di informazioni che meglio identifica un individuo. Esso è una buona base di partenza per la creazione di un’entità di tipo Persona, le cui informazioni possono venir ampliate in futuro, rendendo l’entità sempre più distintiva. Tuttavia tale set di attributi non garantisce la completa univocità, a causa di una casistica molto semplice : l’omonimia. A favore di tale caso consideriamo l’ipotesi di due individui omonimi che frequentino lo stesso ambiente di lavoro.

Successivamente si vuole individuare quali informazioni è plausibile sperare di trovare tramite risorse esterne. Può risultare importante ampliare in tal maniera le informazioni per i motivi accennati poco sopra: più informazioni corrette e coerenti portano a maggior distintività. Non sappiamo in quale misura si esprima tale proporzione, ma è piuttosto intuitivo affermare che la proporzionalità è diretta e positiva. Coerentemente al tipo di risorse disponibile che abbiamo considerato nelle nostre ipotesi (repositori web di autori internazionali, motori di ricerca, web-sites di enti e affiliazioni di ricerca) riteniamo plausibile la possibilità di rinvenire indirizzi e-mail, pagine web, identificativi appartenenti ad altre community, inerenti alle persone che intendiamo registrare nell’ENS.

L’indirizzo e-mail è un dato che per sua natura non è completamente affidabile, tuttavia all’interno della comunità scientifica internazionale la pubblicazione del proprio indirizzo e-mail, funzionante e correlato ai propri dati, viene curata con particolare attenzione. E’ in assoluto il mezzo che più di altri permette di entrare in contatto con altri ricercatori, studenti, professori, enti, per cui nell’ambito che stiamo analizzando possiamo ritenere di utilizzarlo come attributo delle nostre entità. La pagina web, allo stesso modo, può risultare identificativa; verificata la coerenza dei contenuti della pagina, in base all’autore del repositorio AGRIS ad essa associata, possiamo specificarne l’indirizzo all’interno del profilo Okkam dell’entità. Identificativi alternativi, che sono associati alla medesima persona in community differente dalla nostra è l’ultima delle nostre scelte plausibili. Tale possibilità permette al profilo Okkam dell’entità che creiamo di avere delle relazioni con altri repositori; in tal modo la persona associata al suo profilo ENS sarà identificabile univocamente anche in altri sistemi.

Pertanto il set di attributi preferenziale che delineiamo per questo caso d’uso sarà il seguente, sempre coerentemente alla nomenclatura adottata nell’ENS Core Vocabulary:

(first\_name , last\_name , affiliation , country, mbox, homepage, Alternative IDs )

La ricerca manuale delle informazioni mancanti ha premura di verificare che le informazioni ritrovate siano corrette, tramite ragionamenti logici e mezzi che un operatore umano possiede (intuito, considerazioni, risorse web).

**Messa in evidenza delle relazione di co-authoring e affiliazioni**

Dopo aver deciso quali informazioni sono necessarie per completare i dati si vogliono evidenziare le relazione di collaborazione tra i vari autori del campione che analizziamo. Come detto in precedenza, le informazioni analizzate provengono tutte dallo stesso file XML, di determinati mese, anno e stato. Inoltre in ogni gruppo di metadati riguardanti un articolo compaiono più autori per lo stesso.

Nella tabella che realizziamo per contenere i dati estrapolati inseriamo due colonne dedicate allo scopo di questa fase : “co-authoring” e “affiliazioni”; come accennato qualche paragrafo precedente lo strumento software utilizzato per creare questa collezione può essere un foglio elettronico, o una tabella di una base di dati. Nello sviluppo della nostra analisi è stato utilizzato inizialmente un foglio elettronico, ma sono state poi sviluppate interessanti ipotesi riguardanti l’utilità di una base di dati per obiettivi successivi. La colonna “co-authoring” viene completata con informazioni riguardanti gli autori con cui l’individuo del record preso in considerazione ha collaborato. Una sigla o una serie di identificativi locali interni alla tabella sono sufficienti. La colonna “affiliazioni” riporta la stringa che meglio descrive l’ente per cui lavora l’autore, nel linguaggio più diffuso disponibile (i.e. lingua Inglese dove possibile).

L’importanza di questi due campi rispetto agli altri attributi emerge nell’obiettivo di coerenza e consistenza delle informazioni: mentre un nome ed un cognome servono per delineare un individuo, affiliazione e collaborazioni di co-authoring delineano il suo contesto di lavoro. Il ragionamento che sfrutta tali informazioni si esprime nella seguente procedura:

1. utilizzo del motore di ricerca con keywords nome autore, cognome autore, affiliazione (dove presente)
2. osservazione della lista risultati
   1. aprire un risultato
   2. valutare se il contesto che si presenta è plausibile con le informazioni
   3. cercare la citazione del nome dell’autore
   4. verificare la correttezza e coerenza dei dati
   5. cercare co-autori dell’autore in questione
   6. verificare la correttezza dei dati del co-autore
   7. registrare eventuali nuove informazioni
   8. ripetere con il risultato di ricerca successivo
3. valutazione complessiva di tutte le pagine osservate

Per entrare più nello specifico facciamo un esempio: immaginiamo di voler completare le informazioni della persona “Rossi M., Università di Bologna”. Esso corrisponde al record ( Cognome: Rossi, Nome: M., Affiliazione: Università di Bologna). Sappiamo anche dai dati ricavati dal relativo file XML che esso ha collaborato con Bianchi F. e Verdi L. . Nella fase di ricerca manuale proviamo a cercare sul web, tramite motore di ricerca, Rossi M. Università Bologna. Nel lunghissimo elenco di risultati che il motore di ricerca ci darà sfoglieremo i primi più plausibili, e tra questi ci orienteremo cercando l’individuo “Rossi M.” che abbia rapporti di conoscenza con “Bianchi F.” e “Verdi L.”. Fortunatamente tra i risultati ve ne sono due particolarmente interessanti (sempre per ipotesi): la pagina web del dipartimento di scienze agrarie, dove sono elencati tali “Rossi Mario” e “Bianchi Fabrizio”, ricercatori, e un link ad un articolo in formato PDF, i cui autori sono “Rossi M.” e “Verdi L.”, entrambi dell’Università di Bologna, rispettivamente dipartimento di Scienze Agrarie e dipartimento di Biotecnologie. Concludiamo quindi dall’esempio che con i dati a nostra a disposizione abbiamo saputo ampliare le informazioni in nostro possesso ragionando tramite rapporti di conoscenza e affiliazione degli individui.

Come spesso viene ripetuto, tale procedimento è logico ed abbastanza intuitivo per una mente umana, ma non lo è per la macchina. Tuttavia, concepirne le fasi atomiche può aiutare la formalizzazione di un processo più vicino all’elaboratore. Nel nostro caso potremo esprimere in pseudo-codice il ragionamento in questa maniera:

Dati A e B due individui, il primo estratto dai nostri file XML, il secondo rivenuto su una pagina web di un dipartimento universitario, allora possiamo definire che l’individuo è lo stesso con la seguente condizione:

*SE (Cognome\_A = Cognome\_B AND Cognome\_collaboratore\_A = Cognome\_collaboratore\_B AND affiliazione\_A = affiliazione\_B)*

*ALLORA A = B.*

Questa deduzione logica ci porta ad ottenere due nuove informazioni, coerenti e consistenti: Nome = “Mario” (esteso rispetto a prima), Affiliazione = “Università di Bologna, Dipartimento di scienze agrarie” (più ricco rispetto a prima).

La rara ipotesi di omonimia, accennata precedentemente, viene confutata dal controllo delle informazioni rinvenute. Qualora esistesse realmente un omonimo all’interno dello stesso dipartimento della stessa università, con gli stessi collaboratori, sarebbe comparso nella pagina in cui sono stati rivenuti i dati di esempio, dove può venir facilmente evidenziato il problema.

**Utilizzo di risorse web per ottenere informazioni aggiuntive sulle entità.**

Nel paragrafo precedente c’è stato un accenno a questa fase, per dimostrare come alcune informazioni sugli autori hanno particolare utilità nel ricostruire l’entità che verrà poi registrata nell’ENS. Tuttavia in questo paragrafo viene posto l’accento sulle risorse online utilizzate per ricercare identificativi alternativi (Alternative Ids) e community dove i nostri autori di AGRIS siano già stati registrati. Le risorse web individuate si dividono in due insiemi: il primo è quello dei repositori di autori scientifici, il secondo è quello più generale del motore di ricerca, tramite il quale si cerca di andare a scovare pagine web inerenti ai nostri dati.

Come accennato in introduzione, abbiamo deciso di ricercare ulteriori informazioni sugli autori catalogati in determinati siti-web, consapevolmente al fatto che tali portali contengono svariate informazioni riguardanti ricercatori di differenti discipline specifiche. I portali presi in considerazione sono:

* 1. DBLP Bibliography - <http://www.informatik.uni-trier.de/~ley/db/>
  2. RKB Explorer - <http://www.rkbexplorer.com/explorer/>
  3. ResearcherID - <http://www.researcherid.com/>
  4. CiteSeerX - <http://citeseerx.ist.psu.edu/about/site>
  5. arXiv.org - <http://arxiv.org/>

Il procedimento di ricerca a questo punto è oneroso, ma piuttosto banale: si tratta di utilizzare Cognome e iniziale del nome puntata per effettuare delle query nei motori di ricerca di tali siti, ottenere un elenco di risultati (vuoto o con una serie di autori) e sfogliare tra questi per individuare delle connessioni. In questi risultati i casi di omonimia si sprecano, in quanto la ricerca viene effettuata solo sulla base di cognome e iniziale del nome, tuttavia è possibile raffinare l’elenco ottenuto procedendo con la logica discussa nel paragrafo precedente. La maggior parte di questi siti riporta, per ogni autore, le sue collaborazioni; tale preziosa informazione viene quindi utilizzata al nostro scopo, effettuando un vero e proprio confronto tra la collaborazione riportata sul sito e i rapporti di co-authoring individuati dalla nostra analisi.

Purtroppo la fase di utilizzo dei repositori di autori non ha portato a risultati molto utili. Questa problematica non deriva dall’idea che c’è a monte del metodo, ma dalla particolarità dei dati che stiamo analizzando; avessimo trattato autori di un altro campo di ricerca, ad esempio Informatic Science, il sito DBLP Bibliography sarebbe stato molto più utile al nostro scopo. L’origine di questo problema sta nel fatto che i repositori presenti sul web contengono pochi, o nessun, records riguardanti la ricerca nel campo delle scienze agrarie, e, probabilmente, il più completo e ampio di questo tipo di repositori è proprio AGRIS, lo stesso da cui hanno origine i dati che vogliamo ampliare.

La successiva fase di analisi delle risorse web prevede l’utilizzo di uno strumento meno orientato al problema, ma forse più efficace: il motore di ricerca generico. Trattandosi di un caso di autori, i cui articoli sono pubblicati in vari siti, è ipotizzabile che essi possano avere pagine web personali (ipotesi per altro effettuata anche nel secondo processo di questa analisi), o possano comparire su pagine web di conferenze, centri di ricerca, community varie ed eventuali. Semplici query, utilizzando come parole chiave le informazioni a nostra disposizione (ricordiamo l’esempio “Rossi M. Università Bologna”) ci permettono di accedere ad un elenco molto lungo di risultati. La maggior parte delle pagine risultanti dalla ricerca saranno incoerenti con la nostra query, o inutilizzabili, tuttavia si può sperare di individuare una o più risorse in cui rinvenire informazioni mancanti, correlate all’autore su cui stiamo effettuando il processo. Tale procedimento è molto macchinoso e lungo, ma è in assoluto quello più fruttuoso, se applicato con la corretta logica e se verificata la coerenza dei dati.

**Valutazione delle nuove informazioni localizzate.**

Questa fase, posta in conclusione del processo, non è isolata come l’elenco di inizio capitolo può far intendere. Il processo di valutazione permea ogni fase di questa analisi, è l’elemento di controllo che guida le scelte dell’operatore, e non è formalmente corretto incapsularlo in una isolata, posta in una sequenza. Tuttavia discutiamo di questa parte per poter individuare con chiarezza di cosa si tratta.

Vengono considerati tre concetti fondamentali per l’analisi dei dati, sia per quelli che vengono localizzati con la ricerca, sia per quelli che abbiamo nei file XML del repositorio AGRIS: coerenza, consistenza, rilevanza.

1. *Coerenza*. Ogni dato che cataloghiamo deve presentare coerenza, rispetto alla realtà (non possiamo registrare nell’ENS persone inesistenti), al tempo (non possiamo registrare nell’ENS persone con dati obsoleti), e ai dati rinvenuti sul web (nel caso in cui il matching tra i nostri dati e quelli ottenuti non sia completamente esatto occorre invalidare le nuove informazioni e porsi delle domande). Tale concetto è considerato fondamentale per diversi motivi, primo tra tutti lo scopo del sistema del progetto Okkam. Si è già parlato di coerenza nella definizione di entità, ma è importante rimarcare come questa idea debba essere sempre considerata, senza lasciare spazio all’approsimazione.
2. *Consistenza*. Coerenza e consistenza sono concetti molto simili, sicuramente puntano allo stesso obiettivo (come tra l’altro in questo caso). Nella disciplina delle basi di dati si parla di consistenza in riferimento ai record, i quali, nel caso riportino la stessa informazione, devono essere sempre nello stesso stato (se l’informazione viene aggiornata da una parte, l’aggiornamento va effettuato anche dall’altra, devono quindi trovarsi nello stesso stato). Nel nostro caso proponiamo questo concetto distintamente dalla coerenza per ricordare che l’operatore, nell’analisi dei dati, deve verificare la coerenza delle nostre informazioni con l’esterno, ma deve anche effettuare il processo al contrario. Nessuno garantisce che i dati in nostro possesso siano corretti al 100%, tuttavia il sistema di analisi funziona stabilendo che la fonte di dati su cui fare affidamento sia il nostro repositorio di Journals scientifici.
3. *Rilevanza*. Alla fine del processo di ricerca, i dati che vengono collezionati avranno differenti livelli di rilevanza. Facciamo immediatamente un esempio per chiarire il concetto: se ho un record (first\_name = “Mario” , last\_name = “Rossi” , affiliation = “Università di Bologna” , country = “italy”, mbox “m.rossi@fakedomain.it”, homepage “http://www.unibo.it/mrossi” ), nel quale non è stato possibile individuare degli Alternative IDs, l’operatore può decidere di registrare lo stesso tale entità nell’ENS, in quanto la rilevanza e coerenza delle altre informazioni è sufficiente per l’obiettivo, ovvero per l’identificazione univoca e distinta dell’entità associata alla persona. Quindi il concetto di rilevanza di ogni singolo attributo va valutato per decidere quando ci si può ritenere soddisfatti delle informazioni raccolte per ogni singolo record. Tale concetto è lo stesso che ha guidato le scelte del set di attributi minimo e preferenziale, in precedenza.

Applicando tali concetti si raffina la catalogazione dei dati ad un insieme ordinato di informazioni pronte per essere utilizzate per la creazione di entità nell’ENS.

**Formalizzazione del processo utilizzato.**

La sequenza di fasi esposta a inizio capitolo e la loro esplicazione successiva non sono complete senza un’ulteriore chiarimento. Tali tipi di processi, in cui si alternano fasi di catalogazione, raccolta e controllo dati, avvengono solitamente con cicli a spirale. Nel nostro caso, mano a mano che si collezionano più informazioni riguardanti autori, diviene necessario estendere i controlli alle nuove informazioni. Ad esempio se al primo record analizzo l’autore “Rossi M.”, e questo nome si ripropone al record cinquantaquattresimo, sarà necessario ricontrollare la bontà delle informazioni di entrambe le registrazioni.

Pertanto il processo si esplicherà, in pseudo-codice, nella seguente maniera:

*tabella table = new Tabella*

*File filesorgente = new File(“sorgente.xml”)*

*table.aggiungi\_campo (“Nome”)*

*table.aggiungi\_campo (“Cognome”)*

*table.aggiungi\_campo (“Affiliazione”)*

*table.aggiungi\_campo (“MailBox”)*

*table.aggiungi\_campo (“Stato”)*

*table.aggiungi\_campo (“HomePage”)*

*table.aggiungi\_campo (“ID Alternativo 1”)*

*WHILE esistono record da analizzare in “sorgente.xml” {*

*record x = “sorgente.xml”.record(i)*

*table = CATALOGA( x, table)*

*x = ESTENDI\_INFORMAZIONI\_VIA\_WEB(record x)*

*table = CATALOGA( x, table) //aggiornamento*

*}*

La funzione “CATALOGA ()” è il mero processo di inserimento dei dati riguardanti il record all’interno della nostra tabella. Il risultato di questa ciclo è una “Grey Table”, in cui sono contenuti records riguardani entità candidati alla registrazione nell’ENS.

Nel dettaglio le funzioni vengono implementate in tal maniera:

*tabella CATALOGA( record x, Tabella t)*

*{*

*x = EVIDENZIA\_CO-AUTORI(x, t)*

*t.aggiungi\_record(x)*

*return t }*

All’interno del lavoro svolto, la tabella risultante da tale procedura è riportata come esempio in Appendice X.1. La procedura EVIDENZIA\_CO-AUTORI rappresenta un processo che per un esecutore umano è piuttosto banale, per una macchina invece risulta lievemente più complesso:

*record EVIDENZIA\_CO-AUTORI( record x, table t) {*

*IF (t == EMPTY) THEN return x*

*ELSE FOR (tutti i record di t)*

*IF (x.IDarticolo == t(i).IDarticolo)*

*THEN x.co-autori.AGGIUNGI(t(i).Nome+t(i).Cognome)*

*t(i).co-autori.AGGIUNGI(x.Nome+x.Cognome)*

*return x*

*}*

La successiva funzione per le ricerca di informazioni tramite risorse web:

*record ESTENDI\_INFORMAZIONI\_VIA\_WEB(record x){*

*FOR (tutti i repositori di autori)*

*{*

*List lista\_risultati1 = RICERCA\_AUTORE(x.Nome+x.Cognome)*

*FOR (tutti gli elementi della lista)*

*IF (lista\_risultati1(i).cognome == x.cognome)*

*THEN IF (lista\_risultati1(i).nome == x.nome)*

*THEN IF (lista\_risultati1(i).campo\_ricerca == x.affiliazione.campo\_ricerca)*

*THEN x.AGGIORNA(lista\_risultati1(i).dati)*

*ELSE IF (x.VER\_CO-AUTORI(lista\_risultati1(i).co-autori))*

*THEN x.AGGIORNA(lista\_risultati1(i).dati)*

*}*

*List Lista\_risultati2 = RICERCA\_GOOGLE(x.Nome+x.Cognome+x.Affiliazione)*

*FOR (tutti gli elementi della lista)*

*IF (lista\_risultati2(i).CONTIENE(x.cognome) AND lista\_risultati2(i).CONTIENE(x.nome))*

*THEN IF (lista\_risultati2(i).CONTIENE(x.affiliazione) OR lista\_risultati2(i).CONTIENE(x.co-autori) OR*

*lista\_risultati2(i).CONTIENE(x.co-autori.affiliazione))*

*THEN x.AGGIORNA(lista\_risultati2(i).dati)*

*}*

Le funzioni AGGIORNA e CONTIENE sono rispettivamente semplici funzioni per inserire nel record i nuovi dati ritenuti validi, e per verificare se nella pagina risultante compaiono i dati del soggetto analizzato. Nell’argomento della funzione RICERCA\_GOOGLE compare il campo Affiliazione del record x. Come annunciato prima talvolta questo valore non è specificato, quindi nell’argomento dato alla funzione comparirà semplicemente una stringa vuota al suo posto. Quando manca il valore dell’affiliazione la ricerca viene effettuata solo con Cognome e Nome (iniziale del Nome soltanto in molti casi). La condizione dell’ultimo ciclo IF prevede 3 casi, al presentarsi dei quali la condizione è verificata: la pagine di risultato è inerente all’affiliazione del soggetto, la pagina di risultato riporta i nomi dei co-autori del soggetto, la pagine di risultato è inerente all’affiliazione dei co-autori del soggeto. L’ultima condizione si fa necessaria qualora il campo Affiliazione fosse vuoto, e l’unica informazione contestuale riguardante il soggetto derivasse dai suoi co-autori.

Ritroviamo in queste semplici righe di pseudo-codice un approccio quasi ricorsivo, nel quale l’esecutore umano simulato amplia le sue valutazioni su un soggetto autore mano a mano che raccoglie nuove informazioni. Quando tali informazioni arrivano al livello di soddisfazione previsto (o viene concluso che non è possibile raggiungere tale livello) egli decide di passare al soggetto successivo, e così via.

**Conclusioni.**

La capacità di astrazione e ragionamento della mente umana è di gran lunga la più sviluppata. Il limite di tale potentissimo strumento è la mole di dati che riesce ad elaborare contemporaneamente, limite che viene agevolmente superato dagli elaboratori, al giorno d’oggi. Consapevoli di non poter espandere la “memoria” umana, la frontiera della ricerca si concentra (ormai da decenni) sull’espansione della capacità di ragionamento degli elaboratori. Per i casi di studio più vari vengono elaborate macchine dotate di intelligenza artificiale, mentre per casi meno ambiziosi e più specifici si istruiscono le macchine a elaborare automaticamente dati strutturati in modi specifici, per fornire risultati in maniera veloce, puntuale e coerente. Per il nostro caso di elaborazione dati abbiamo discusso un processo manuale di approccio a dati strutturati e non, dove l’esecutore umano si è trovato di fronte a delle informazioni riguardanti autori.

Uno dei risultati che si è voluto ottenere da tale processo riguarda la comprensione dei passaggi svolti. Nella sezione scorsa si è cercato di schematizzare con pseudo-codice le scelte che sono state effettuate per concludere il valore di ogni informazione di questo caso. Il valore di questa schematizzazione si esprime nel tentativo di formalizzare il ragionamento umano, chiaramente legato ad un caso d’uso molto specifico. Non si sta descrivendo come insegnare ad una macchina a riconoscere dati, ma si sta provando a evidenziare un processo automatico per riconoscere le informazioni riguardanti degli autori di un repositorio scientifico.

Oltre alla schematizzazione del metodo, tale processo manuale ha dato origine alle seguenti entità Persona:

|  |
| --- |
| * <http://www.okkam.org/ens/id452ff850-3825-4d6d-b46e-a9fef6b5d4ad> |
| * <http://www.okkam.org/ens/id290f0a1d-4113-4e6c-8422-ca347a173820> |
| * <http://www.okkam.org/ens/ida2891cd3-6617-4ac7-a8b9-d458ea27992a> |
| * <http://www.okkam.org/ens/idea69430b-87ac-44c4-b8f3-1b12c9198c83> |
| * <http://www.okkam.org/ens/id3ab4b169-2107-4b12-b31b-b10010732a98> |
| * <http://www.okkam.org/ens/idc627641d-c900-4306-9de0-d539c1514b4e> |
| * <http://www.okkam.org/ens/id075bb22d-e271-489e-ac89-f7f3b1be6818> |
| * <http://www.okkam.org/ens/id02b2fff8-02d5-4b9b-9c40-c6fd8b2138bb> |
| * <http://www.okkam.org/ens/ide45a85b0-e62d-40ea-b9c9-249c59ab309b> |
| * <http://www.okkam.org/ens/id364d43a7-d08c-4d3d-b3d8-828dab2442d0> |
| * <http://www.okkam.org/ens/idc6d788eb-e00f-476a-a5e7-a7d346fde35f> |
| * <http://www.okkam.org/ens/id6da10723-a403-4bad-b0ad-171e856e263e> |
| * <http://www.okkam.org/ens/id26830f68-e772-41dc-b8f4-8fa84dfaa54d> |
| * <http://www.okkam.org/ens/ida9050ca1-d362-498e-89bf-d66ed55973f7> |
| * <http://www.okkam.org/ens/id021235a3-0208-446a-95eb-f14e88b607e2> |
| * <http://www.okkam.org/ens/id934bb77a-a73c-4f67-b046-219641303f0b> |
| * <http://www.okkam.org/ens/ide1b3c9ac-fec8-4ce8-b033-5d17f1933a62> |
| * <http://www.okkam.org/ens/ide5d05354-a5c9-4252-8dd8-cd14aa09305c> |
| * <http://www.okkam.org/ens/ide585b701-c75c-45e3-a448-a85670475043> |
| * <http://www.okkam.org/ens/id357540d0-cf33-4511-a131-9cb5d91f0d1c> |

Tali entità, create dal processo svolto manualmente dall’esecutore umano, si inseriscono all’interno dei metadati degli articoli scientifici del repositorio AGRIS. Precedentemente è stato mostrato come tali metadati siano strutturati riguardo questa informazione ( Fig. 1 ), ed all’interno degli attributi del tag XML che si riferisce all’autore va inserito l’Okkam ID ad esso associato, allo scopo di arricchirne il valore semantico ed il potere di identificazione del soggetto. L’utilizzo di tali metadati arricchiti diviene più significativo: si hanno molte informazioni riguardanti i journals e gli articoli, come in principio, e si ha modo di identificare in maniera forte e disambiguata l’autore relativo. Questa possibilità aggiuntiva, offerta dall’entità distintiva e dal suo Okkam ID, permette di raggiungere tramite Web tutte le altre risorse, affiliazioni, community, riguardanti quell’autore, agevolando la ricerca di qualsiasi utente voglia interessarsi al suo ramo di ricerca ed alle sue pubblicazioni. Tale nuova frontiera viene ampliata in questo documento quando si parla di Sig.Ma (\*).

**Impieghi futuri e correlati.**

Il risultato di questo processo è una “Grey Table”, creata manualmente. Normalmente le “Grey Table” per le entità candidate vengono create tramite applicativi che agevolano la gestione e l’inserimento dei record, e non con inserimento manuale di dati. Tuttavia si può concludere che è possibile estendere questo tipo di lavoro in tal senso: un’applicazione che prenda in input le entità considerate valide dalla tabella e effettui la creazione dell’entità nell’ENS, tramite le Java APIs di Okkam. Per questo tipo di obiettivo ben si presta la tabella di una base di dati, piuttosto che un foglio elettronico. La valutazione formale del processo, schematizzando i ragionamenti umani in pseudo-codice, permette di avere un’idea delle intuizioni che la nostra mente elabora di fronte a tale situazione. La prima deduzione è che le nostre capacità mentali lavorano in maniera automatica ed inconscia con dati di tale semplicità; diversamente, è necessario uno sforzo quando viene chiesto di accorgersi di che tipo di ragionamento si sta effettuando, allo scopo di formalizzarlo. Ciò che per l’essere umano sembra scontato non lo è per la macchina; ne consegue che non è scontato realizzare strumenti di elaborazione automatica, quindi è necessario avvicinarsi a tali problemi molto gradualmente. La schematizzazione in pseudo-codice di un ragionamento può essere un primo passo in questa direzione.

La realizzazione di uno strumento in grado di prendere in input dei dati riguardanti un’entità, e sfruttare risorse dalla struttura eterogenea per ampliare tali dati in maniera coerente, avrebbe un grande seguito di approvazioni, risolvendo alcune delle maggiori problematiche della disciplina dell’elaborazione dati.

***Appendice X***

**X.1** *Tabella su Foglio Elettronico per registrare le entità Persona visionate – alcune colonne sono state omesse per motivi di visualizzazione.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| File di Origine | Cognome | Nome | Stato | Affiliazione | E-Mail | Web-Site |
| 02/ES901.xml | Villalba Fonfría | Angel | Spain | University of València (Spain) | [Angel.Villalba@uv.es](mailto:Angel.Villalba@uv.es) | <http://www.uv.es/~villalba/> |
| 02/ES901.xml | Bono Martinez | Emerito | Spain | University of València (Spain) | [Emerit.Bono@uv.es](mailto:Emerit.Bono@uv.es) |  |
| 02/ES901.xml | Tomas Carpi | Juan Antonio | Spain | University of València (Spain) | [Juan.A.Tomas@uv.es](mailto:Juan.A.Tomas@uv.es) |  |
| 02/ES901.xml | Gómez-Limón Rodríguez | José Antonio | Spain | University of Valladolid, Palencia (Spain) | limon@iaf.uva.es |  |
| 02/ES901.xml | Picazo Tadeo | Andrea J. | Spain | University of Valencia (Spain), Economy | [Andres.j.picazo@uv.es](mailto:Andres.j.picazo@uv.es) | http://www.uv.es/~ajpicazo/ |
| 02/ES901.xml | Reig Martínez | Ernest | Spain | University of Valencia (Spain), Economy | [Ernest.Reig@uv.es](mailto:Ernest.Reig@uv.es) | http://www.uv.es/ereig/ |
| 02/ES901.xml | Carús Ribalaygua | Luis | Spain | University of Zaragoza (Spain) | carus@unizar.es |  |
| 02/ES901.xml | Murgado Armenteros | Eva | Spain | University of Jaén (Spain) | emurgado@ujaen.es | http://www4.ujaen.es/~emurgado/ |
| 02/ES901.xml | Ciruela Lorenzo | Antonio Manuel | Spain | University of Málaga (Spain) | [ACL@UMA.ES](mailto:acl@uma.es) | http://webpersonal.uma.es/~ACL/acl.htm |
| 02/ES901.xml | Bruque Cámara | Sebastián | Spain | University of Jaén (Spain) | [sbruque@ujaen.es](mailto:sbruque@ujaen.es) |  |
| 02/ES901.xml | Martínez de Anguita | Pablo | Spain | University Rey Juan Carlos, Madrid (Spain) |  |  |