Chiara Paci

Clotilde, un'applicazione per la gestione e l'analisi di corpus di interesse storico-filologico.

Lo strumento che si vuole realizzare servirà ad agevolare la creazione di corpus digitali di testi di interesse storico-filologico e a estrarre da questi stessi testi informazioni di carattere linguistico e storico.

L'obbiettivo principale è minimizzare il lavoro ripetitivo. Per questo è previsto che l'utente fornisca soltanto:

- un insieme di testi senza tag; al più l'utente dovrà inserire dei separatori (spazi, segni di interpunzione, ecc.) e marker di formattazione (allineato a destra o sinistra, sottolineato, ecc.);
- un insieme di regole con cui analizzare i testi;
- altri dati necessari all'analisi o alla contestualizzazione (lemmari, collocazione e datazione dei testi, ecc.).

Per l'inserimento sono previsti dei tool per agevolare i compiti ripetitivi e per l'import e la conversione di dati preesistenti.

Inoltre il programma potrà utilizzare fonti esterne (dizionari, database bibliografici, altri corpus, ecc.) e interagire con altri tool di analisi e gestione di corpus.

1. Interfaccia utente

L'interfaccia utente è composta di due parti, una di inserimento e una di visualizzazione.

1.1. Inserimento dati

L'inserimento dati riguarda i testi, le regole e i dizionari inseriti dall'utente, manualmente o con l'ausilio di altri tool.

Per l'inserimento di testi sono previste due modalità:

- *manuale*, in cui l'utente trascrive a mano il testo (o si avvale di un OCR e verifica l'output);
- *importazione*, in cui appositi tool convertono testi già trascritti (ad esempio in formato html o odt o provenienti da altri corpus) e li inseriscono nel database.

Anche per l'inserimento di metadati (collocazione, date, produttore, ecc.) sono previste le due modalità manuale e importazione.

Per l'importazione sarà creata un interfaccia che consentirà di supportare nuovi formati tramite plugin anche di terze parti.

Le regole grammaticali e i dizionari sono salvati nel database in un formato utile al programma per l'analisi e uguale per tutte le lingue.

All'utente sono forniti dei tool di inserimento (specifici per ogni lingua) per agevolarlo nell'operazione di data entry e di definizione delle regole e per consentirgli di importare i dati in modo massivo. Questi dati vengono poi convertiti nel formato voluto dal programma.

Anche qui verrà creata un'interfaccia che consenta di sviluppare plugin per supportare le diverse lingue o diversi criteri di definizione delle regole.

1.2. Visualizzazione ed estrazione dei dati

La visualizzazione riguarderà, di base, gli oggetti ricavati dall'analisi, quindi:

- testi analizzati e taggati;
- regole verificate e glossari;
- informazioni ricavate dai documenti, di tipo enciclopedico o quantitativo.

Sarà comunque possibile aggiungere dei moduli per personalizzare la visualizzazione e l'estrazione.

Sempre tramite plugin sarà possibile esportare i dati anche in formati adatti ad essere utilizzati con altri tool.

2. Interfaccia verso l'esterno

In un'ottica di riuso, il programma sarà in grado di ricercare autonomamente informazioni, per esempio in dizionari online, su Wikipedia, in database bibliografici, in progetti opendata, ecc.

Inoltre potrà interfacciarsi con tool esistenti di terze per parti sia per lo scambio di dati che per demandare l'esecuzione di alcune analisi.

3. Motore interno

Internamente, il programma è composto di due parti: un database e un motore di analisi.

Il database contiene i dati forniti dall'utente (dizionari, regole, testi e metadati), e le strutture di appoggio dei moduli di analisi del programma (cache, tabelle intermedie, ecc.).

Il motore di analisi è composto da moduli indipendenti di analisi, pensati come filtri in cascata. Possono essere di tre tipi:

- di analisi morfologico-sintattica basati su regole grammaticali: per compiere l'analisi applicano le regole impostate dall'utente in modo deterministico;

- ◈ di analisi morfologico-sintattica basati su dati statistici e algoritmi adattativi: eseguono l'analisi cercando di individuare delle regole in base a quanto appreso in precedenza;
- ◈ di analisi semantico-pragmatica: estraggono informazioni di tipo enciclopedico (luoghi citati, persone, identificazione di eventi, ecc.) o di tipo quantitativo, cercando di comprendere l'argomento e di inserirlo in un contesto.

L'utente potrà scegliere quali moduli inserire o escludere per il tipo di analisi che gli interessa. I moduli potranno essere aggiunti come plugin di terze parti sia come componenti creati ad hoc, sia come interfacce verso tool esterni.

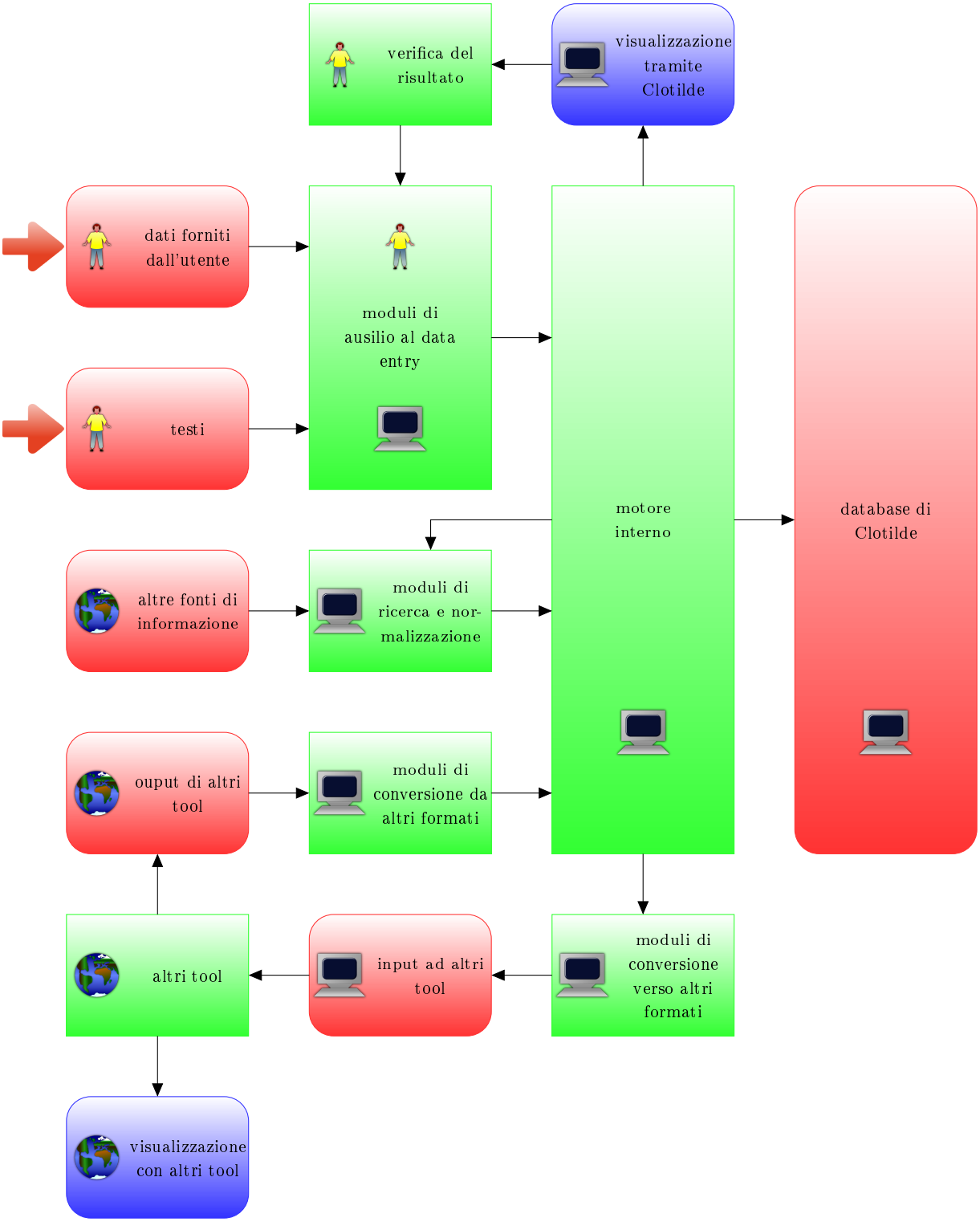

Figura 1: Schema delle attività dell'utente.

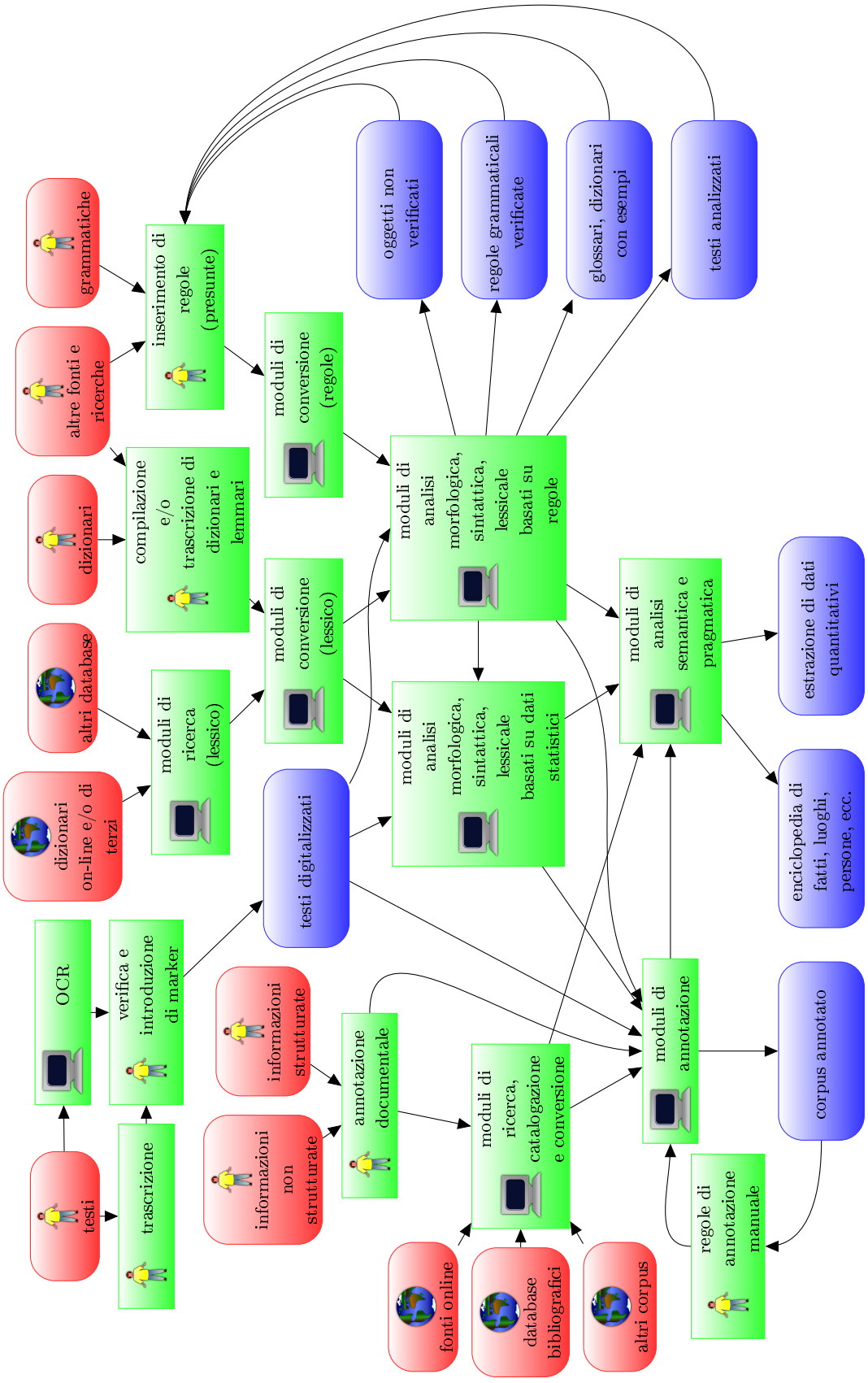

Figura 2: Schema del funzionamento.

Indice