Protocollo
per la Documentazione dei Progetti di Ricerca

v0.2 (Versione Preliminare)

Nicola Tommasi
nicola.tommasi@univr.it
C.I.D.E. Centro Interdipartimentale di Documentazione Economica

21 agosto 2013

Key words: documentazione, replicabilità, ricerca empirica

Abstarct

Questo protocollo descrive come organizzare una ricerca empirica per consentire ad una persona terza di replicare facilmente ed esattamente ogni passaggio della gestione e dell'analisi dei dati e quindi anche di replicare i risultati ottenuti.

1 MOTIVAZIONE

Il principio guida che ispira questo protocollo è di fornire uno schema per l'organizzazione dei lavori di ricerca empirica. La speranza è di riuscire ad implementare una serie di regole che consentano la replicazione con il minimo sforzo di tutti i passaggi di preparazione dei dati, di analisi e di ottenimento dei risultati riportati nella stesura finale dei lavori. A tal fine è necessario che all'interno dei do-file tutte le operazioni di caricamento e salvataggio di file avvengano tramite percorsi relativi. Con queste regole si spera anche di ottenere maggiore organizzazione ed efficienza nel lavoro che abitualmente ricercatori e studenti conducono presso le strutture del CIDE. L'osservanza del protocollo dovrebbe consentire agli studenti di poter interagire meglio con i propri docenti ed inoltre dovrebbe incentivare il loro senso di responsabilità e di rendicontazione del proprio lavoro. Una copia finale di tutti i documenti elettronici contenuti nella cartella del progetto sarà conservata negli archivi informatici del CIDE.

2 SCHEMA E PRINCIPI

Ogni progetto di ricerca sarà organizzato come di seguito illustrato. Innanzitutto si crea una cartella *cognome_nome*. Questa al suo interno conterrà almeno le seguenti cartelle e file:

1. una cartella *data* e la suo interno altre due cartelle
   (a) la cartella *raw_data*
   (b) la cartella *out_dta*
2. una cartella *metadata*
3. una cartella *docs*
4. una cartella *graphs*
5. un file di testo *readme.txt*
6. un file *master.do*
7. un file *import.do*
8. un file *cleaning.do*
9. un file *results.do*

Vediamo ora qual è la funzione di ciascuna cartella e di ciascun file.

cartella *data*

Contiene al suo interno la cartella *raw_data* in cui verranno messi tutti i file di partenza dei dati nel loro formato originale. Se questi file sono in un formato direttamente caricabile da Stata[1] si lasciano come sono altrimenti, oltre al file in formato originale, bisogna inserire anche la sua conversione in un formato importabile da Stata. La cartella *out_dta* invece ha la funzione di raccogliere tutti i file in formato .dta prodotti durante l'elaborazione dei dati. Conterrà anche tutti i file di dati che eventualmente dovessero essere prodotti in un formato diverso dal .dta.

cartella *metadata*

Contiene tutti i file che servono da documentazione per i dati contenuti nella cartella *row_data*. Solitamente sono file di documentazione forniti assieme ai dati come la definizione delle variabili, gli schemi di codifica dei valori o i metodi di campionamento e di raccolta dei dati. Se il file è unico deve avere lo stesso nome del relativo raw file, altrimenti si crea una cartella sempre con il nome del relativo raw file e al suo interno si inseriscono tutti i file di documentazione.

cartella *docs*

Contiene tutta la documentazione relativa al progetto come paper di riferimento o istruzioni su come organizzare l'elaborazione dei dati.

cartella *graphs*

È una cartella opzionale, nel senso che se non vengono prodotti grafici che debbano essere salvati si può evitare di crearla. Se invece si producono grafici e questi devono essere salvati come file (.gph, .eps, .eps ...) questa è la cartella che li conterrà. Se i grafici sono molti, si possono organizzare in sottocartelle della principale.

file *readme.txt*

Questo file contiene una panoramica di tutto il materiale che è stato assemblato nella cartella *cognome_nome*. In particolare dovrà contenere:

- ▷ la lista di tutti i file contenuti nella cartella *data* con descrizione della loro posizione all'interno della cartella , del loro contenuto e del loro formato
- ▷ la fonte dei dati ed eventualmente le istruzioni per riottenere gli stessi dati
- ▷ la lista in ordine di esecuzione di tutti i .do files contenuti nella cartella *cognome_nome* e una breve descrizione della loro funzione
- ▷ una referenza (e-mail, numero di telefono...) per contattare l'autore del lavoro in caso di necessità di ulteriori informazioni

[1] .dta, .xls, .xlsx, .csv, tutti i dati in formato testo delimitato e non delimitato. Per questi formati e altri non direttamente caricabili si veda l'appendice finale.

file *master.do*

Contiene la sequenza di lancio dei do-file. I tre do-file indicati sono caldamente consigliati ma non obbligatori, ovvero se le operazioni da compiere non sono molte il do-file potrebbe essere unico e quindi diventerebbe inutile anche il master.do. D'altra parte, se il progetto di ricerca fosse particolarmente complesso, è consigliabile aumentare il numero di do-file in aggiunta a quelli consigliati.

file *import.do*

Lo scopo di *import.do* è di importare i dati da ciascun file della cartella *row_data* in formato .dta e salvarli nella cartella *out_dta*. Se tutti i dati fossero già in formato Stata questo do-file non deve essere creato. Ciascun .dta file creato avrà lo stesso nome del corrispondente file in *row_data*. Per esempio se in *row_data* c'è il file *unaid.txt*, il corrispondente file in *out_dta* si chiamerà *unaid.dta*.

file *cleaning.do*

Lo scopo di *cleaning.do* è di processare i dati al fine di arrivare al dataset finale da usare per le analisi. Quindi carica i dataset presenti in *out_dta* e applica procedure di pulizia dei dati, di merge e di append, quindi salva il file così prodotto sempre in *out_dta* con il prefisso *clean_* o *final_*. É consigliabile in questa fase anche condurre delle analisi di esplorazione dei dati e di sperimentazione per verificare preventivamente le analisi che si intendono effettuare in *results.do*.

file *results.do*

Contiene i comandi per generare per generare le variabili necessarie, per generare tabelle e figure, regressioni e tutti i risultati pubblicati nel report finale della ricerca.

Appendice

Segue una panoramica dei principali comandi per caricare o importare dati in Stata. Per maggiori informazioni `help <command>`

- ▷ `use` carica dataset in formato Stata (.dta)
- ▷ `insheet` legge file in formato testo (ASCII) con una osservazione per riga e dove i valori sono separati da una tabulazione, da una virgola o da altro delimitatore
- ▷ `infile` esistono due versioni di questo comando. La prima versione legge file in formato testo libero in cui i dati sono separati da uno o più spazi bianchi o da tabulazioni, la seconda versione legge dati in formato testo fisso con l'ausilio di un dictionary file
- ▷ `infix` legge dati in formato testo fisso con l'ausilio di un dictionary file
- ▷ `import excel` importa dati da Excel, sia nel formato .xls che nel formato .xlsx. Consente di selezionare il foglio o un range di celle da cui importare i dati

- ▷ `import sasxport` importa dataset in formato SAS XPORT
- ▷ `odbc` consente di caricare (o di vedere) dati da una fonte Open DataBase Connectivity (ODBC)
- ▷ `xmluse` legge dati XML in formato Excel o dati XML in formato Stata
- ▷ `haver` Carica dati dal database Haver Analytics
- ▷ `freduse` consente di caricare dati direttamente dal database Federal Reserve Economic Data (FRED) della Federal Reserve Bank of St. Louis directly (user written command)
- ▷ `insheetjson` importa tabelle da fonti JSON su internet (user written command)
- ▷ `usespss` importa dati da SPSS nel formato .sav (user written command)

Riferimenti bibliografici

[1] Ball Richard and Medeiros Norm. 2011. Teaching Students to Document Their Empirical Research.