Data Storage and Documentation

Francesco Vespignani
francesco.vespignani@unitn.it

Indice

Perchè questo corso

Perchè questo corso

- difficoltà ad *insegnare* la programmazione
- consapevolezza di carenze (mie) relativamente a questo aspetto della ricerca
- rapida evoluzione delle tecnologie relative ai dati
- necessità di sviluppare sistemi di gestione di dati scientifici

Difficoltà ad insegnare la programmazione

- tipico processo di apprendimento implicito: learning by doing
- spesso ostacolo all'utilizzo di software di analisi dati (R, matlab) è legato all'importazione e gestione dei dati

Workaround: rendere interessante la tecnologia.

Carenze

La maggior parte dei dati *importanti* che possiedo sono su google mail o google drive.

Non è sempre chiaro chi sia, all'interno di un gruppo di ricerca, il responsabile della gestione dei dati.

Abbiamo responsabilità nei confronti di chi ci finanzia, chi ci aiuta (partecipanti), la comunità scientifica relativamente alla conservazione e condivisione di dati.

Evoluzione tecnologica

Elementi di natura differente che convergono nel rendere interessante e complesso questo argomento:

- Big data, data science e ampi interessi economici relativamente alla gestione e condivisioni di dati
- Replicability issues e sviluppo di *best practice* nella ricerca scientifica
- Intrinseca evoluzione del web (web 3.0)

Sviluppo e definizione di nuovi standard.

Gestire la complessità

Aumentare la consapevolezza delle problematiche tecnologice ed economiche sottese.

Interagire con tecnici e *decisori* relativamente agli investimenti in strumenti, conoscenze e professionalità relative a questi aspetti (migrazione email unitn a google, disponibilità illimitata di spazio in *google drive*, centralizzazione dei servizi di calcolo e storage).

Chi e come deve fornire strumenti adeguati ai ricercatori per conservare, condividere e comunicare i dati? Ricercatori stessi, tecnici, servizio bibliotecario?

Tecnologie dell'informazione

Standard e lock-in tecnologico

- Sviluppare uno standard tecnologico fornisce un grosso vantaggio
- Adeguarsi ad uno standard comporta lock-in tecnologico

Dal punto di vista dell'individuo è importante trovare il giusto compromesso fra lo scommettere ed investire (switching costs) su *nuovi potenziali standard* e (ri)utilizzare ciò che già conosce (subire il lock-in).

Su questi argomenti si veda ad esempio Shapiro & Varian (1999) Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press.

Standard

Standards are published documents that establish specifications and procedures designed to ensure the reliability of the materials, products, methods, and/or services people use every day (...)
Standards form the fundamental building blocks for product development by establishing consistent protocols that can be universally understood and adopted. This helps compatibility and interoperability (...)
In summary standards fuel the development and implementation of technologies that influence and transform the way we live, work and communicate.

ieee standard development.

Tipi di Standard

- Standard de jure, istituzionali gestiti da organizzazioni (ISO, IEEE, IETF/RFC)
- Standard de facto, open, industriali e commerciali (formato .doc di microsoft, DXF di autocad)
- Standard aperti (linux, TCP/IP, ASCII, bitcoin) e chiusi (dxf)

Tipi di Standard

IEEE corporate office

Satoshi Nakamoto, inventor of bitcoin

Data Storage and Documentation

Conservazione e comunicazione di dati

Cosa c'entrano gli standard tecnologici con la ricerca scientifica?

Conservare per se o condividere con altri:

- conservare privatamente i dati per poter sfruttare appieno le potenzialità del lavoro intellettuale che ha permesso la loro creazione (pubblicare)
- condividere i dati per permettere alla comunità scientifica il massimo sfruttamento delle informazioni contenute nei dati stessi

Dati e documentazione

I dati hanno senso e sono correttamente interpretabili solo se accompagnati da metadati, informazioni che descrivono i dati.

Tipi di metadati:

- informazioni per poter accedere ai dati ed interpretarli in modo corretto
- informazioni relative alla procedura con la quale i dati sono stati raccolti e manipolati (metodo)
- informazioni amministrative e legate alla preservazione dei dati (creatore, licenza)

La distinzione fra dato e sua documentazione non è sempre netta (anche la trascrizione di una procedura è un dato, alcuni dati sono "auto-documentati").

Archiviazione non è un semplice deposito

Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a project and incorporate a schedule for depositing products over the course of a project's life cycle and the creation and preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method.

Jacobs & Humphrey (2004) Preserving research data. ACM Comm., 47(9), 27-29.

Dati e documentazione

Il tipo e dettaglio nel quale i dati vanno descritti dipende dalla finalità della archiviazione e comunicazione, adeguandosi alle massime di cooperazione della comunicazione.

Esempi:

- descrivere una procedura nel metodo di un articolo scientifico o in guidelines di laboratorio
- commentare una funzione per se stessi, per collaboratori o per il pubblico generale

Alcune possibili regole

- stabilire (in anticipo o in corso d'opera) dei *deliverables* che richiedono archiviazione
- evitare multipli backup parziali, cercare ove possibile di avere un'unica *fonte* di dati e di documentazione
- scegliere dei *formati* che consentano di accedere a dati e metadati in modi differenti da differenti utenti
- valutare la sostenibilità nel tempo dell'archivio (preservazione e aggiornamento)

Gli strumenti

Depositi o archivi?

- google drive e altri cloud (dropbox, sourceforge, github)
- reserach gate, accademia.edu
- open science framework osf.io

Quanto scommettere su uno specifico standard (switching costs)?
Come utilizzare strumenti come OSF?

Conoscenze

- formati di dati e metadati (documentazioni)
- poter accedere ai formati in lettura e scrittura (in modo automatizzato)
- conoscere piattaforme di archiviazione e sistemi di gestione delle versioni
- strumenti amministrativi (md5, doi, licenze)

Questo corso

- lettura e scrittura di files
- markdown e pandoc
- git e github
- jupyter notebook
- proprietà intellettuali

Basi sui files

- files di testo e binari
- immagini raster e vettoriali (bmp, png, svg)
- pdf (latex), html e xml