Organizzazione dei dati in EGRID

Come i dati vengono immagazzinati nella griglia computazionale EGRID + convenzioni + modi di lavorare.

- Ripartizione dati tra vari SE.
- Organizzazione all'interno di SE + Convenzioni su nomi dei file + procedure operative in genere.
- Impatto dell'utilizzo dei nomi logici.
- Impatto dei requisiti di sicurezza.
- Strategie per attenuare limitazioni di banda.

Ripartizione dati tra SE

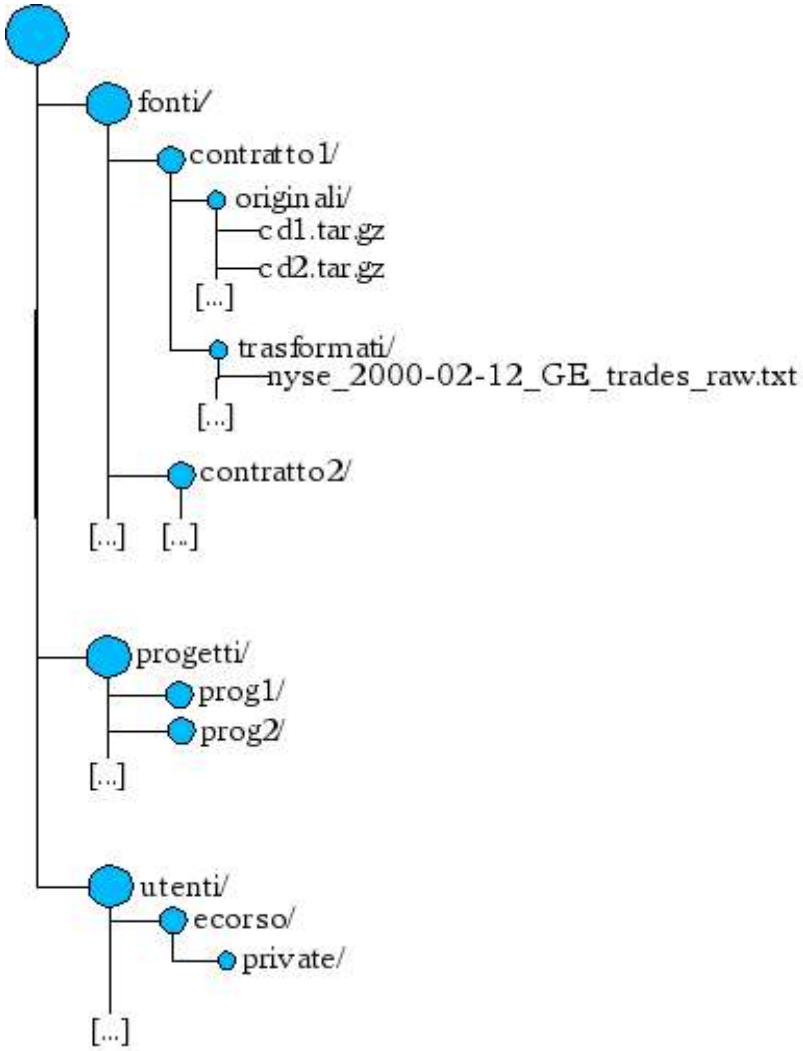
- EGRID è costituita da più SE sparsi geograficamente.
- Architettura a Stella: centro=INFN Padova, nodi periferici=gruppi di ricerca.
- INFN Padova: 2,6TB + archiviazione affidabile.
- Nodi Periferici: zona tampone per minimizzare spostamento ingenti quantità dati.
- L'architettura è un primo passo per mitigare le limitazioni di banda.
- Sfrutta tecnologia RLS di EDG: ma ha limiti rispetto requisito sicurezza EGRID = soluzione temporanea!

Organizzazione in directory

- Consentire amministrazione nodi periferici libera + garantire SE principale a Padova.
- SE principale: organizzazione directory e permessi **prefissati.**
- Nodi periferici: senza imposizioni **ma** vivamente consigliata sincronia con Padova per facilitare lavoro utenti! (strumenti forniti ai sistemisti).

Organizzazione in directory

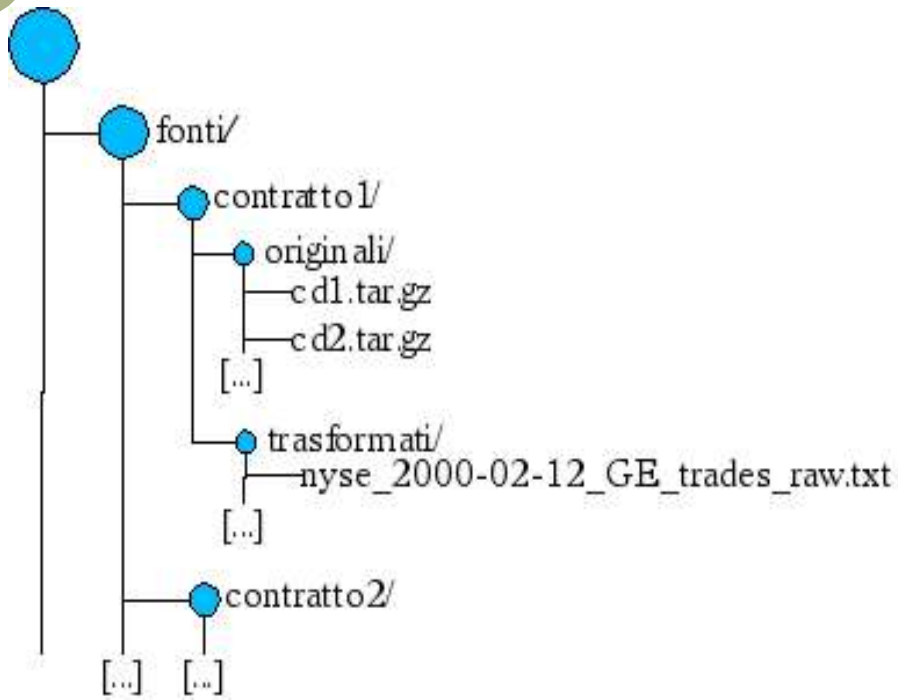
- Albero directory a Padova:



Tre directory principali:

- fonti
- progetti
- utenti

Organizzazione in directory - fonti



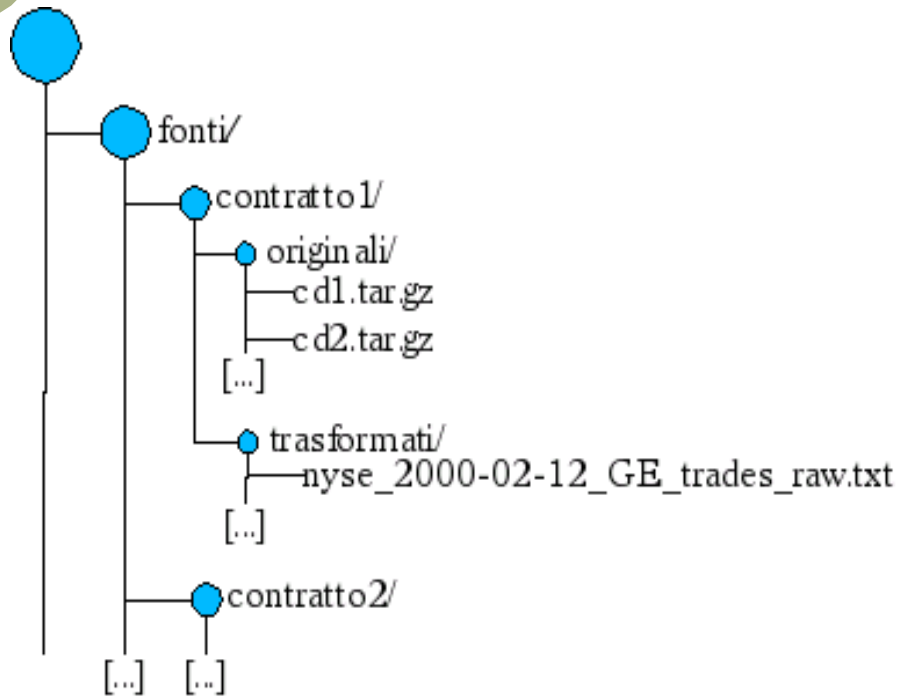
Una sottodirectory per ciascun contratto.

Per ciascun contratto: originali + trasformati.

originali: contenuto CD/DVD dei contratti.

trasformati: formato dei dati più utile ai ricercatori (riorganizzazione contentuto + rinominazione per individuare contenuto).

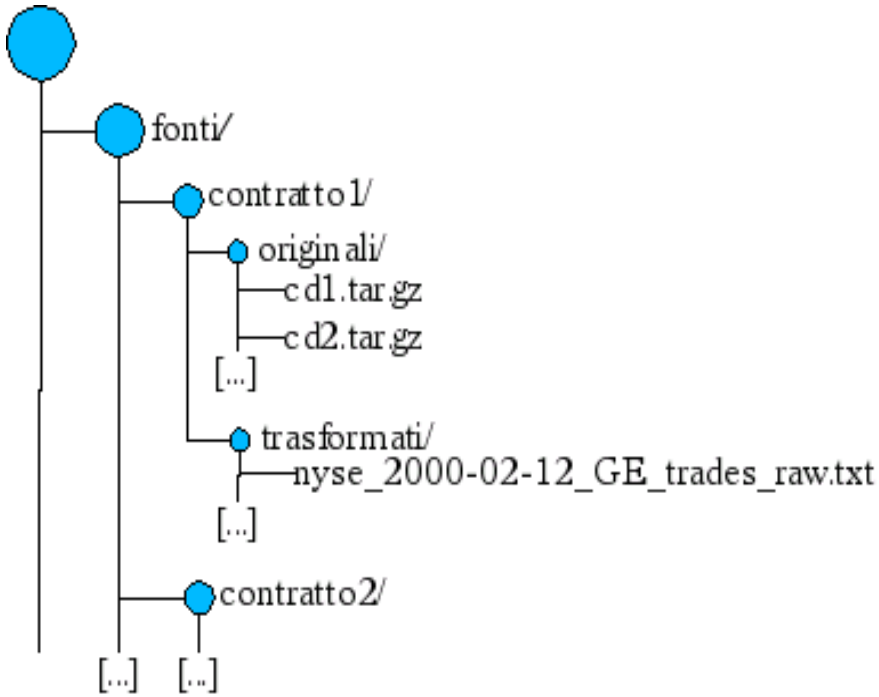
Organizzazione in directory - fonti



- Solo gli utenti facenti parte del contratto possono leggere – tutti gli altri utenti EGRID non riescono. Garantita integrità dei dati!
- Solo Utente Ammnistratore del Contratto ha diritti di scrittura: carica i dati + esegue trasformazioni. Garantita integrità dati!

Organizzazione in directory - fonti

Convenzioni e procedure standard

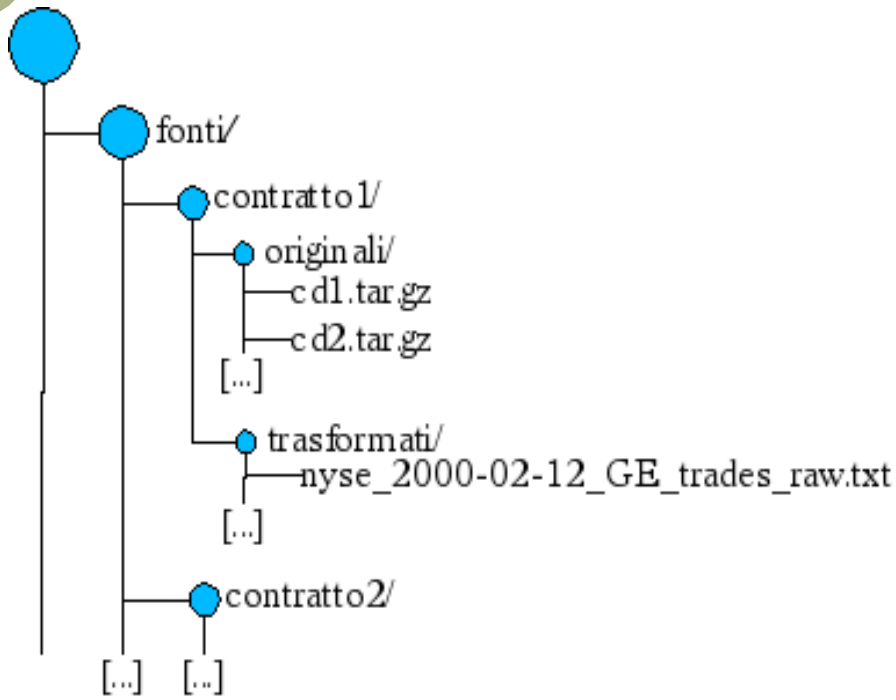


Per la creazione della directory di contratto:

- E-mail firmata digitalmente con richiesta di creazione che specifica il nome da assegnare, gli utenti con diritto di lettura + utente amministratore contratto

Organizzazione in directory - fonti

Convenzioni e procedure standard

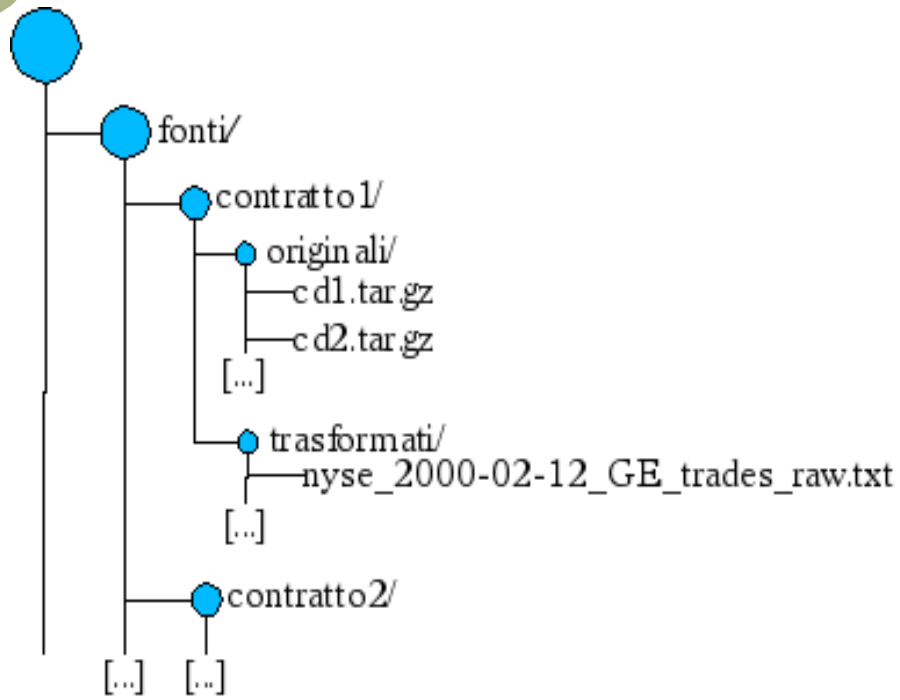


Caricamento CD/DVD - il contenuto va preparato:

- In una macchina non in griglia si crea directory con nome arbitrario.
- Vi si copia il contenuto del CD/DVD.
- Si crea tar.gz della directory.
- Si carica il file in griglia.

Organizzazione in directory - fonti

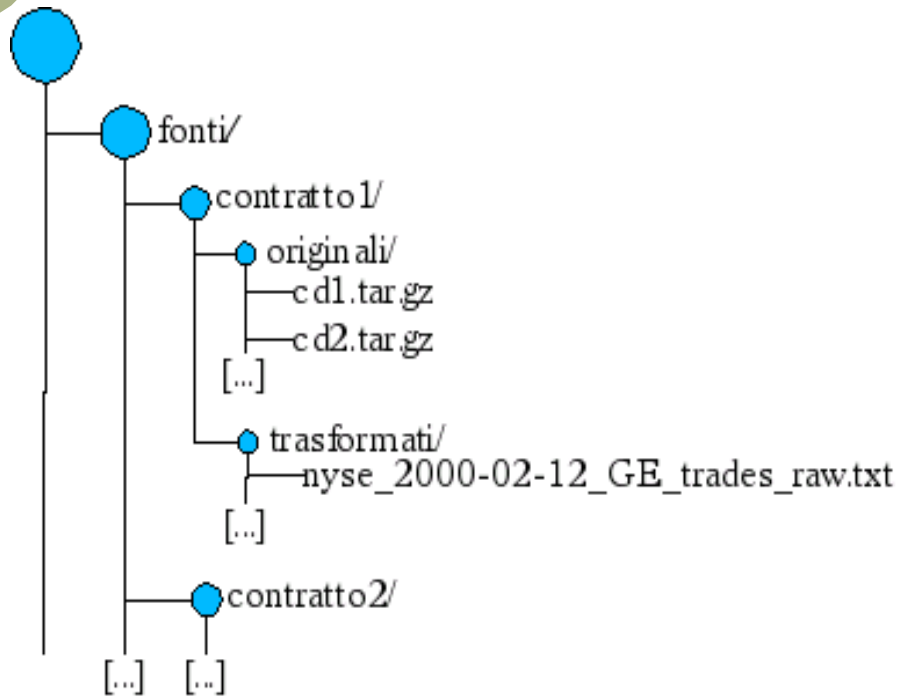
Convenzioni e procedure standard



- Il tar.gz comprime i dati per risparmiare spazio.
- La directory è necessaria per convenzione sui programmi di trasformazione.

Organizzazione in directory - fonti

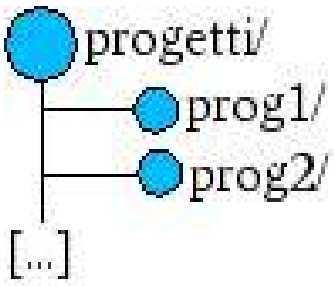
Convenzioni e procedure standard



I dati trasformati:

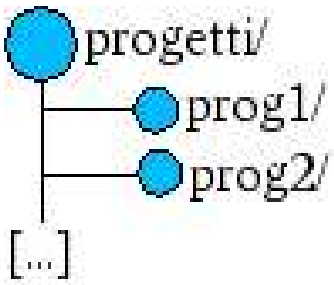
- Naming Convention per descrivere contenuto file: centrale per lavoro successivo dei ricercatori.
- Esempio nyse_2000-02-12_GE_trades_raw.txt

Organizzazione in directory - progetti



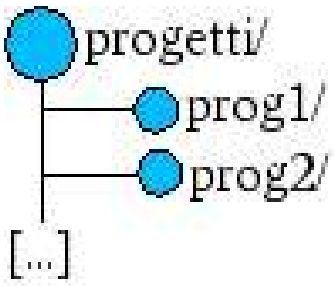
- Fornisce grado di flessibilità ai ricercatori per organizzarsi come meglio credono.
- Utenti EGRID liberi di creare directory + impostare gruppi utenti + diritti.
- Esempi: creazione gruppi all’interno del proprio istituto; creazione co-operazione tra ricercatori di gruppi diversi.

Organizzazione in directory - progetti



- Utente ha gli strumenti per gestione autonoma: non dipende dai servizi centrali EGRID.

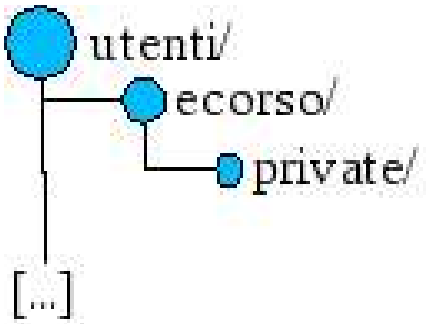
Organizzazione in directory - progetti



Permessi:

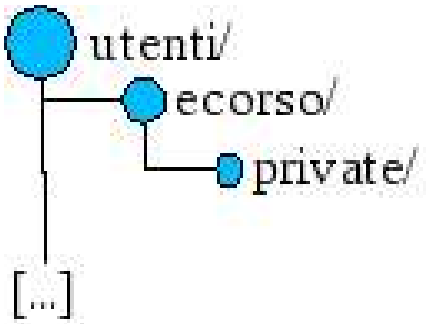
- Utenti possono creare directory + cancellare quelle proprie.
- Possono scrivere nelle directory proprietaria + quelle in cui è stato dato permesso.
- Possono leggere solo dove è stato dato permesso dagli altri utenti.

Organizzazione in directory: utenti



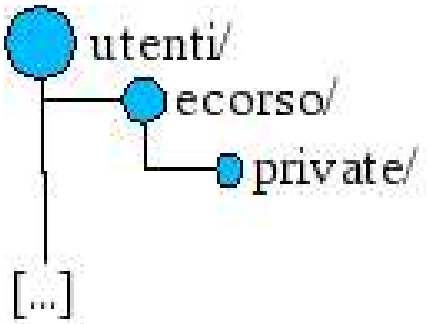
- Spazio personale per il lavoro in griglia di ciascun ricercatore: cioè la propria home!
- Ricercatore salva i propri file, risultati parziali, organizza il lavoro come meglio crede.

Organizzazione in directory: utenti



- Per convenzione il nome assegnato alla home di griglia coincide con username.
- Creazione home + impostazione permessi: alla richiesta di un utente di accedere in griglia.

Organizzazione in directory: utenti



Permessi:

- Proprietario legge e scrive.
- Tutti utenti EGRID leggono.
- *private* è blindata: solo proprietario vi accede.

Organizzazione in directory

I file caricati nello SE principale di Padova vengono organizzati all'interno delle strutture mostrate e i diritti di accesso seguono da dove sono stati collocati.

Nomi logici: convenzione sull'assegnazione

- File fisico individuato da un nome logico: stesso nome individua medesimo file su diversi SE.
- RLS di EDG consente nomi logici diversi dal nome fisico.
- Con migliaia di file questa flessibilità origina problemi (individuazione file).
- Si rende necessaria convenzione su come assegnare un nome logico ad un file fisico.

Nomi logici: convenzione sull'assegnazione

- Nei nomi logici è consentito lo '/'.
- Ci si basa sulla struttura di directory di Padova.
- La convenzione: Nome logico = intero percorso fisico del file a Padova.
- Per esempio: file dati.txt in /utenti/ecorso a Padova, avrà nome logico lfn:/utenti/ecorso/dati.txt.
- Nome logico che inizi per lfn:/utenti/ indica file collocato nella cartella /utenti/ dello SE principale.
- Allora nomi logici potranno solo iniziare per lfn:/fonti/, lfn:/progetti/ e lfn:/utenti.

Nomi logici: strategie di ottimizzazione

- Meccanismo nome logico è alla base dell'evitare scaricamenti ripetuti dello stesso file da e verso centro stella (problema limitazioni di banda).
- In EGRID operazioni spostamento file suddivisa in due fasi: prima interazione con SE locale, successiva interazione con SE principale.
- Caso caricamento iniziale CD
- Caso scaricamento dati su propria macchina UI dell'utente

Nomi logici: strategie di ottimizzazione

- Implicazione diretta è che si rende necessario controllo grado riempimento SE periferici.
- Svuotamento SE periferici: la copia principale è a Padova, è sempre disponibile, è a Padova che avviene il grosso dell'elaborazione dati.
- Comandi EDG Replica Manager + Comnadi EGRID – ma comandi EGRID più facili.

Impatto dei Nomi Logici sul modo di lavorare

- Con un nome logico s’individuano più file fisici: ciò ha impatto diretto sul modo di lavorare dell’utente in particolare per dati che vengono spostati da Padova al nodo periferico.
- A Padova l’integrità dei dati è garantita dalla struttura di directory, dai permessi e dalle procedure: nello spostamento si esce da questo contesto controllato e l’integrità potrebbe venire meno.
- Per ovviare questa evenienza si propongono due modi d’agire per elaborazioni su nodi periferici con dati inizialmente presenti a Padova.

Impatto dei Nomi Logici sul modo di lavorare

- Vengono distinti due casi.
- Uso personale + modifica contenuto = copia periferica con nuovo nome logico (così indica esplicitamente la diversità rispetto il file di partenza).
- Convenzione nel caso di nodo periferico gestito liberamente: nuovo nome logico inizia col nome della città in minuscolo + percorso locale completo
- Se invece più ricercatori utilizzano lo stesso file originale: eseguire replica locale che ne conserva nome logico ed evita ad altri utenti locali il tempo di scaricamento

I Nomi Logici

- Convenzione sul modo di assegnare i nomi logici.
- Modo operativo per minimizzare limitazioni di banda
- Distinzione tra operazioni che modificano dati e mantengono integrità.

Impatto dei requisiti di sicurezza

- Conseguenze nomi logici = operazioni che modificano file scaricato + operazioni che mantengono integrità.
- Effetto operativo di richiedere all'utente di assegnare o meno nuovo nome logico – rende esplicita la differenza.
- Rimane da garantire gli altri ricercatori dell’integrità del file!

Impatto requisiti di sicurezza

- In linea di massima questo si ottiene impostando diritti di sola lettura alla directory *locale* che accogli i dati.
- Questo è un problema non banale: operazioni di sola lettura nello SE di Padova (che garantiscono l'integrità) diventano anche di scrittura nel nodo locale (altrimenti non si trasferisce il file)!

Impatto dei requisiti di sicurezza

Per affrontare il problema si propone una strategia transitoria in attesa del completamento delle soluzioni più robuste:

- Garantiamo piena integrità di Padova.
- Garantiamo che nel nodo periferico non accedano dalla griglia utenti non-locali.
- All'interno del nodo periferico i gruppi creati e permessi dei gruppi sono impostati secondo politiche locali, rimanendo in ultima analisi responsabile dei file scaricati il ricercatore che li ha copiati.

Impatto dei requisiti di sicurezza

- Conseguenza pratica è che il ricercatore deve conoscere la struttura di directory periferica e i permessi di accesso + organizzarsi con l'amministratore locale per le proprie esigenze.
- In altre parole arginiamo possibili interazioni esterne e responsabiliziamo in locale l'amministratore + ricercatore.