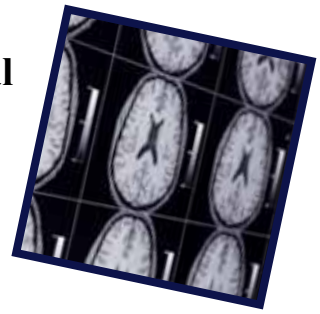
Progetto DECIDE: un esempio di infrastruttura al servizio della comunità biomedica



Valeria Ardizzone

Consortium GARR

Abstract. Il morbo di Alzheimer è la causa più comune di demenza, stimata nel 40-70% dei casi tra la popolazione oltre i 65 anni di età, e la sua diffusione è destinata ad aumentare notevolmente nei prossimi decenni. La demenza ha un enorme impatto sociale sulle famiglie, sui governi e sui loro settori sanitari e sociali. Per questo motivo, la malattia di Alzheimer non è solo una priorità europea, ma planetaria. Il modo in cui i ricercatori guardano al morbo di Alzheimer e alle demenze in generale è cambiato notevolmente: la demenza ora è vista come una fase avanzata dell'evoluzione della malattia, mentre lo scopo prioritario è quello di identificare la malattia di Alzheimer nella sua fase iniziale di sviluppo, utilizzando una combinazione di risultati prodotti dall'imaging strutturale (MRI), funzionale (FDG-PET), molecolare (PIB-PET) e dai test biochimici (analisi del CSF). Anche le analisi di test elettroencefalografici (EEG), in termini di densità di potenza e coerenza spettrale, stanno acquisendo sostegno da parte della comunità scientifica, poiché possono essere utilizzati per lo screening preliminare di grandi campioni di popolazioni.

1. Introduzione

Il progetto DECIDE [1] è incentrato sul fornire un valido supporto ai neurologi e alle varie figure mediche coinvolte nella diagnosi e prognosi delle malattie neurodegenerative. Il progetto, impiegando la e-Infrastructure basata su Grid, prevede la realizzazione di un servizio la cui comunità di riferimento sia in primo luogo quella clinica piuttosto che quella della ricerca. Il principale obiettivo è di fornire ai medici degli ospedali, e non solo dei grandi centri specializzati di ricerca e ricovero, strumenti efficaci per determinare i marcatori clinici per la diagnosi precoce dei disturbi neurologici (malattie neurodegenerative come l'Alzheimer) e psichiatrici (schizofrenia), insieme con la loro rilevanza prognostica.

Il raggiungimento dello scopo richiede lo sviluppo di una nuova infrastruttura nella cui progettazione, essendo il progetto focalizzato più sull'utenza della comunità clinica, è stata tenuta in grande considerazione la presenza dei vincoli specifici in termini di facilità d'uso, di standardizzazione, di sicurezza, di riservatezza dei dati. Particolare attenzione è stata inoltre prestata alle questioni etiche e giuridiche connesse alla gestione, all'elaborazione ed alla distribuzione dei dati. Il servizio offerto a questa comunità ha bisogno di essere compatibile con la routine clinica, e quindi essere semplice da usare, robusto, ragionevolmente veloce e non richiedere troppa interazione con il sistema.

2. Metodologia

L'idea ispiratrice alla base del progetto non è stata quella di fare qualcosa "per" una comunità, ma piuttosto "insieme" ad una comunità. Gli utenti del servizio sono stati coinvolti da subito nella fase di sviluppo del servizio, al fine di garantire che le loro esigenze fossero prese in considerazione per raggiungere la piena fruibilità del servizio finale nell'ambiente clinico.

La piattaforma DECIDE è costituita da tre strati differenti, e cioè le reti della ricerca, le risorse Grid e le applicazioni specifiche del dominio di riferimento.

- La connettività di rete fornita dalla dorsale europea GÉANT [2] e dalle NRENs (le reti nazionali della ricerca e dell'istruzione) dei paesi che partecipano al progetto, interconnette con collegamenti ad alta capacità diversi tipi di strutture, quali centri clinici e di ricerca ed istituti universitari di ricerca;
- L'infrastruttura Grid è utilizzata come piatta-

Workshop GARR Calcolo e Storage Distribuito - Selected Papers. Roma 29-30 novembre 2012
ISBN: 978-88-905077-4-8 © 2013 Associazione Consortium GARR

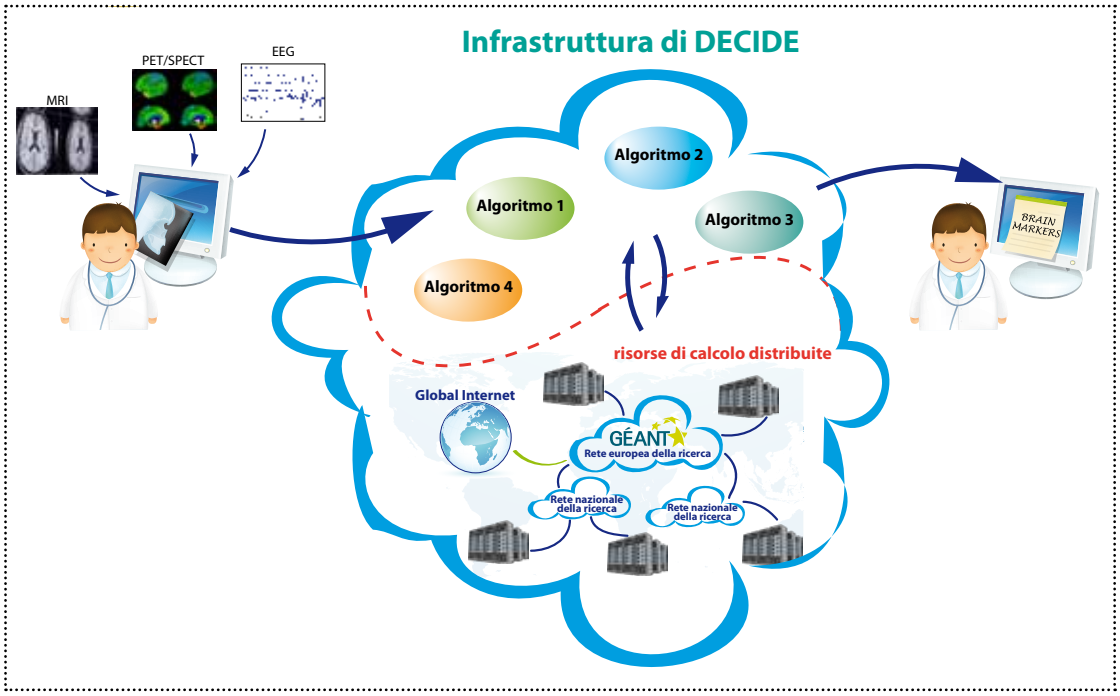


Fig 1 - Rappresentazione della eInfrastructure DECIDE

forma per consentire la collaborazione tra tutti i partner, una sorta di "collante" tecnologico atto ad armonizzare e unificare gli sviluppi, e come bacino di risorse di calcolo e storage in cui grandi volumi di dati possono essere ospitati in modo sicuro e le analisi possono essere eseguite;

- L'uso dei dati medici diversamente acquisiti *(Magnetic Resonance Imaging - MRI, Positron Emission Tomography - PET e Elettroencefalografia - EEG)* consente la combinazione di approcci diagnostici complementari sulla diagnosi delle malattie neurodegenerative, consentendo sinergie tra i diversi ambiti clinici e sostenendo possibili studi di correlazione tra i diversi approcci neurologici.

Una decisione chiave, presa sin dalle prime fasi del progetto, è stata quella di adottare, ove possibile, standard esistenti, al fine di contribuire a ridurre i costi di sviluppo e manutenzione, semplificare l'adozione e l'integrazione con altri servizi, e di ampliare la comunità di utenti. Il servizio DECIDE si basa su standard a tutti i livelli:

- dal punto di vista ICT, questo è vero dal livello di middleware e di rete (EMI / gLite [3], adottato su siti di produzione ufficiali EGI [4], la Grid Infrastructure europea), fino al livello del portale del progetto, il cosiddetto "Science Gateway" [5] basato su *Liferay portal framework* [6] (JSR 168/286 [7] per le portlets, SAML [8] per l'autenticazione, LDAP per il database di utenti, PKCS #11 per la crittografia e SAGA [9] per l'interfaccia con il *middleware*);
- dal punto di vista clinico, il progetto ha documentato e reso disponibile al pubblico le procedure di preparazione del paziente, preparazione esami, acquisizione e controllo di qualità dei dati, con il duplice obiettivo di migliorare la qualità e il contenuto informativo dei dati acquisiti e garantire che essi siano coerenti con i dataset di riferimento del progetto.

A livello tecnologico, si è deciso di sviluppare e implementare i servizi sulla base dell'infrastruttura Grid e del middleware sfruttando due caratteristiche:

- disponibilità di strumenti per integrare le risorse geograficamente distribuite, dove per "risorse" intendiamo principalmente le banche dati di immagini necessarie per addestrare gli algoritmi per il calcolo del volume di regioni cerebrali da immagini MRI, o quelle necessarie per fare confronti statistici di immagini FDG-

PET con i casi di pazienti "normali";
- capacità di stabilire criteri di autorizzazione fino al livello del singolo utente, requisito importante sia perché consente ai proprietari dei dati di mantenerne il controllo, pur facendo utilizzare da altri utenti l'informazione in essi contenuta tramite le applicazioni integrate, sia perché permette di controllare con precisione quale utente può avere accesso a una data applicazione.

In passato, l'adozione e l'impiego della tecnologia Grid, soprattutto da parte di utenti non esperti di IT, sono stati severamente limitati dalla scarsa usabilità e dai vincoli imposti dall'utilizzo di certificati personali, rigide procedure di sicurezza, interfacce a riga di comando, script di esecuzione scritti in linguaggi particolari. In DECIDE, queste problematiche sono state risolte con l'introduzione di uno *Science Gateway* (SG), ossia un portale web che integra un insieme di strumenti e di applicazioni, realizzati su misura del progetto, per soddisfare le esigenze della comunità. Il problema dell'identificazione degli utenti senza far uso di certificati personali Grid, è stato risolto con l'integrazione nel portale SG di certificati robot e mediando l'accesso ad essi tramite le Federazioni di Identità (quale l'italiana IDEM [10]). Questa configurazione combinata permette l'identificazione sicura degli utenti e risolve automaticamente i problemi della loro gestione in relazione al ciclo di vita delle identità digitali. È importante notare che nel corso degli ultimi due anni, l'efficacia dell'approccio SG è stata ampiamente riconosciuta, come testimonia il fatto che più di dieci Autorità europee di Certificazione hanno aggiornato le loro procedure per il rilascio anche di certificati di robot, e dalla presenza di numerosi (~ 30 nel solo EGI) SG che servono le varie comunità.

Gli utenti del servizio DECIDE sono stati classificati in tre gruppi, secondo le funzionalità messe a disposizione dal sistema:

1. "Neurologi": questi professionisti si prendono cura dei pazienti durante l'intero processo diagnostico, dalla diagnosi alla terapia. Hanno bisogno solo di richiedere degli esami da far eseguire ad altri utenti del servizio e recuperare i referti, che poi saranno combinate e usate per fare la diagnosi.
2. "Medici": questi professionisti (radiologi, neurofisiologi, medici nucleari) forniscono informazioni diagnostiche ai neurologi, per il test specifico di competenza.
3. "Scienziati": questi utenti possono avere diversi profili scientifici (fisici, matematici, statistici, ecc.). Hanno a che fare con gli algoritmi diagnostici e collaborano con i medici, fornendo la conoscenza e la comprensione della metodologia sottostante.

A ciascun gruppo di utenti corrisponde una vista personalizzata del portale e un crescente grado d'interazione con il servizio al quale sono autorizzati. Una specifica attività formativa è propedeutica all'uso del servizio per ciascuna categoria di utenti.

3. Descrizione della tecnologia

Come accennato nel paragrafo precedente, l'attuale implementazione del servizio è basata su middleware Grid: come primo passo, la Virtual Organization (VO) vo.eu-decide.eu è stata registrata nell'EGI Operations Portal ed i siti del progetto sono stati configurati per supportare tale VO ed offrire un insieme di servizi Grid (topBDII, WMS, LFC, CE, SE, UI). Per garantire il livello di servizio richiesto per la categoria degli utenti clinici, si è deciso di fare affidamento esclusivamente sui siti di produzione certificati EGI, poiché obbligati a rispettare certi livelli di disponibilità e affidabilità. Per quanto riguarda invece l'uso del servizio in termini di ricerca, le applicazioni possono funzionare su qualsiasi sito che supporti la VO del progetto. La scelta dei siti per l'esecuzione del lavoro è attivata richiedendo uno specifico software tag, che è pubblicato dal VO *Software Manager* tramite apposita procedura.

Il portale SG è uno strumento estremamente potente che rende la Grid utilizzabile dalle diverse comunità di utenti. Esso si basa sul f*ramework Liferay* ed è un contenitore di *portlet* 2.0 che supportano lo standard JSR-286. All'in-

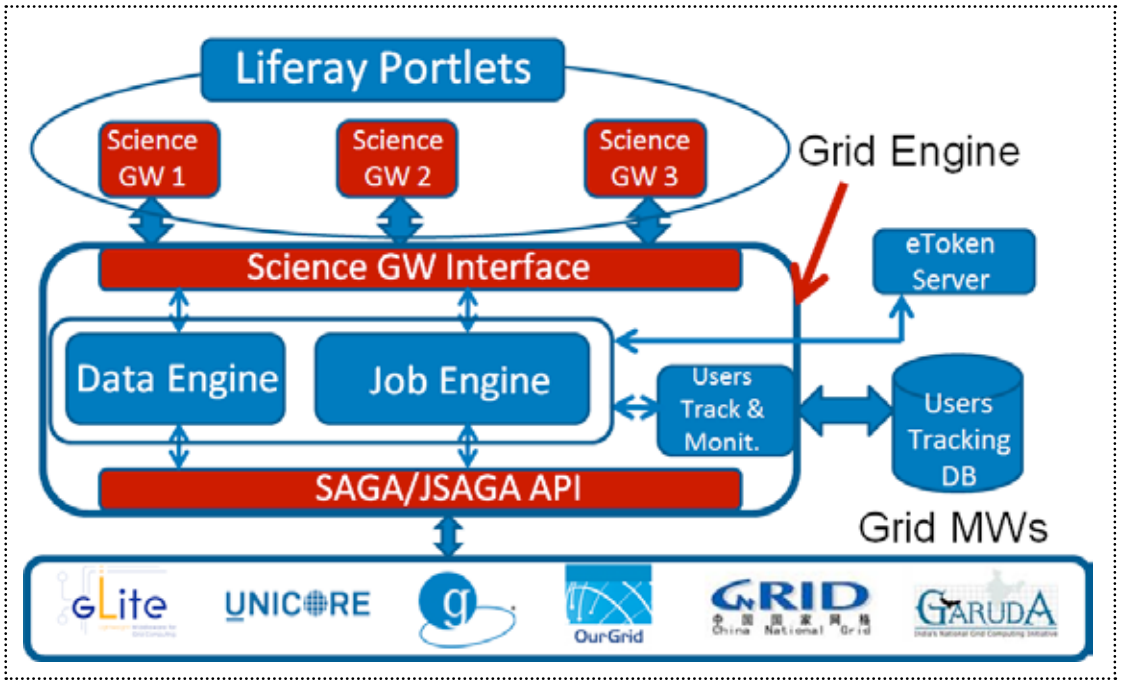


Fig 2 - Rappresentazione schematica delle componenti del portale SG

terno di questo framework, le applicazioni sono costruite e personalizzate in base alle esigenze degli utenti, da esperti sviluppatori di software, combinando, ove possibile, le portlet esistenti o scrivendone delle nuove. L'interazione con i servizi di Grid è mediata dal *Grid Engine*, uno strato di software compatibile con lo standard SAGA, in particolare con la sua applicazione JAVA (JSAGA) [11], progettato per interagire con una serie di middleware. Il Grid Engine isola efficacemente le applicazioni dai livelli sottostanti: invece, il portale SG può sottomettere con successo jobs a diverse infrastrutture basate su differenti middleware quali gLite, Unicore, Globus (in uso presso EGI), OurGrid (Brasile), CNGrid (Cina) e Garuda (India). Questo isolamento renderebbe l'eventuale passaggio ad altri paradigmi di calcolo (ad esempio, il cloud) piuttosto semplice.

Per evitare l'uso dei certificati personali degli utenti, il portale SG interagisce con il cosiddetto eToken server, grazie ad una leggera criptolibrary [12]: tale server contiene i certificati robot (uno per ogni applicazione/ruolo, memorizzati su una smartcard USB fisicamente collegata al server) e gestisce la creazione di *proxy* per conto dell'utente. Poiché con quest'approccio ogni utente è incanalato attraverso lo stesso tipo di proxy, l'associazione di ogni attività Grid all'identità digitale dell'utente è realizzata attraverso la registrazione ed il suo tracciamento da uno speciale database, lo *UserTracking DB*: questo rende il portale SG compatibile con le procedure di sicurezza EGI. Per quanto riguarda gli aspetti di sicurezza, l'eToken server appartiene ad una rete privata ed è accessibile solo dal portale SG; inoltre, a differenza di altri Science Gateway di uso comune, i certificati digitali non viaggiano mai sulla rete, solo i proxy di breve durata lo fanno (12 ore al massimo).

Nel portale SG i meccanismi di autenticazione e autorizzazione di un utente sono stati disaccoppiati. Per l'autenticazione, SG supporta diverse Federazioni d'Identità, come eduGAIN [13] e IDEM. Una Federazione d'Identità "*catch-all*" chiamata GrIDP [14] è stata creata e mantenuta dal progetto, per consentire la registrazione di utenti non appartenenti alle Federazioni. Lo SG può essere facilmente esteso ad altre Federazioni d'Identità che supportano lo standard SAML2 nelle sue implementazioni di Shibboleth e SimpleSAMLphp [15]. Il supporto alle Federa-

zioni d'Identità è un potente motore per la diffusione e l'adozione del servizio, grazie al fatto che non appena dei nuovi Provider d'Identità si uniscono ad una Federazione, questi si traducono in nuovi bacini di potenziali utenti del portale DECIDE.

Una volta che un utente è autenticato, deve essere autorizzato ad accedere ai servizi DECIDE. Per prima cosa si deve registrare al portale, operazione questa che innesca la creazione di una voce in un database LDAP, poi avrà bisogno di frequentare un corso di formazione anche online sull'uso del servizio scelto ed il ruolo richiesto, al fine di ottenere la relativa qualifica. Solo a questo punto alla voce relativa all'utente nel database LDAP viene associato il ruolo corrispondente appena conseguito. Dal punto di vista del portale SG, i benefici di questo meccanismo di autenticazione e di autorizzazione sono:

- offloading della gestione delle identità: non c'è bisogno di mantenere, proteggere, aggiornare, e verificare la validità delle credenziali utente sul SG;
- esecuzione delle procedure per garantire che tutti gli utenti del servizio siano stati adeguatamente formati;
- pieno controllo su chi è autorizzato ad utilizzare il servizio.

Dal momento che l'infrastruttura del progetto richiede la gestione di particolari tipi di dato dei pazienti, molta attenzione è stata posta sui problemi di sicurezza. Una tipologia di dato è particolarmente critico: le immagini FDG-PET che compongono il dataset di riferimento per i casi normali. A differenza delle immagini MRI, infatti, non esistono banche dati pubbliche per la loro conservazione,e forti vincoli legali, etici e anche finanziari limitano notevolmente la possibilità per gli ospedali e per i centri di ricerca di acquisirli. Una delle applicazioni offerte dal servizio DECIDE esegue un confronto tra l'immagine del paziente con un "cervello medio" costruito combinando le immagini di un certo numero di soggetti normali, al fine di evidenziare le regioni statisticamente significative di "ipometabolismo del glucosio". I migliori risultati sono stati ottenuti con l'uso di 50-100 immagini di pazienti normali, ma il dataset è stato allargato in modo da permettere di filtrare su parametri quali il sesso, l'età del paziente, il produttore e il modello dello scanner di acquisizione delle immagini. Le immagini dei soggetti normali sono una ricchezza immensa per un ospedale: basti pensare che un ospedale abbastanza grande può riuscire ad ottenerne mediamente una decina all'anno, sempre che abbia i fondi necessari per acquisirle.

Per permettere ai centri di ricerca medica coinvolti nel progetto di condividere efficacemente tali preziose risorse senza rinunciare alla proprietà, è stato adoperato il servizio middleware "*SecureStorage*" [16], una soluzione Grid-aware per memorizzare i dati crittografati su degli Storage Element: in questo modo, neppure gli amministratori di sistema degli Storage Element sono in grado di accedere ai dati riservati. I file dei dati riservati sono criptati da un *server KeyStore* (KS), ospitato nel dominio amministrativo del proprietario dei dati (ospedale o clinica) e poi copiati su uno o più *Storage Element*: i job delle applicazioni in esecuzione sui *Worker Nodes* possono contattare il KS per recuperare la chiave di decriptazione e, se le autorizzazioni possedute sono soddisfatte, l'applicazione può avere accesso al dato ed utilizzarlo per l'analisi statistica.

Le informazioni relative a ciascun file di dati sono memorizzate in un catalogo di metadati attraverso il servizio gLibrary [17], garantendo l'accesso e la gestione da parte di utenti ed applicazioni.

4. Conclusioni

I servizi che sono stati realizzati nel progetto DECIDE sono stati offerti agli utenti finali attraverso un nuovo tipo di portale SG, basato su Liferay e sugli standard più comuni, arricchito da un sofisticato meccanismo di autenticazione ed autorizzazione in grado di facilitare l'accesso e l'uso della tecnologia Grid, garantendo nel contempo il rispetto ed il controllo di ruoli e privilegi personalizzabili. Il DECIDE SG, inoltre, con-

sente la creazione e la gestione di grandi librerie digitali distribuite di immagini mediche e offre il servizio per la loro criptazione prima di memorizzarli sull'infrastruttura.

La sostenibilità dell'infrastruttura è garantita dal fatto che tutti i siti che fanno parte del servizio di produzione appartengono a organizzazioni che fanno parte delle iniziative Grid nazionali stabilite nei vari Paesi. Diverse iniziative programmate sono state previste per raggiungere la sostenibilità a lungo termine e descritte nel piano di sostenibilità realizzato prima del termine del progetto (corsi di formazione sull'uso dei servizi offerti, nuovi finanziamenti ottenuti partecipando a varie call nazionali ed europee, ecc.) e alcune delle azioni previste hanno già portato all'ottenimento di nuovi finanziamenti per proseguire il lavoro, a testimonianza del forte interesse suscitato da questo progetto.

La lezione principale che scaturisce da questa esperienza è che il coinvolgimento degli utenti già dalla fase di progettazione è fondamentale per identificare i bisogni e le esigenze specifiche delle comunità da raggiungere. In una fase successiva, la presenza di una comunità di motivati *early adopter* e il sostegno di alcuni opinion leader nel campo della ricerca e cura della Malattia di Alzheimer assicurano che il servizio fornito sarà accettabile per la più ampia comunità. Inoltre, problemi di sostenibilità devono essere seriamente considerati, poiché l'orizzonte temporale di un ospedale o di un centro di ricerca è tipicamente di diversi anni. A questo proposito, la scelta di aderire a standard e di sviluppare il servizio utilizzando un approccio stratificato garantisce che il prodotto risultante possa essere facilmente adattato alle future tecnologie.
Riteniamo che l'approccio DECIDE giocherà un ruolo significativo nel mettere a disposizione dei cittadini europei procedure cliniche di qualità scientificamente avanzata.

Ringraziamenti

Il lavoro descritto in questo articolo è stato realizzato all'interno del progetto DECIDE (*Diagnostic Enhancement of Confidence by an International Distributed Environment*), finanziato dalla Commissione Europea nell'ambito del 7° Programma Quadro per la Ricerca e lo Sviluppo Scientifico e Tecnologico. Si ringrazia il consorzio del progetto per aver reso possibile questi risultati. Maggiori informazioni su DECIDE sono disponibili su: www.eu-decide.eu

Riferimenti Bibliografici

[1] DECIDEProject:https://www.eu-decide.eu/

[2] GEANT Project: http://www.geant.net.

[3] gLite middleware: http://glite.cern.ch , European Middleware Initiative, http://www.eu-emi.eu/

[4] European Grid Infrastructure (EGI): http://www.egi.eu.

[5] Science Gateways are discussed in: Wilkins-Diehr N., Gannon D., Klimeck G., Oster S., Pamidighantam S. (2008), TeraGrid Science Gateways and Their Impact on Science, IEEE Computer 41(11), 32-41.

[6] The Liferay portal framework: http://www.liferay.com.

[7] The JSR 286 standard: http://www.jcp.org/en/jsr/detail?id=286.

[8] The SAML standard: http://saml.xml.org.

[9] The SAGA OGF Standard Specification: http://www.gridforum.org/documents/GFD.90.pdf.

[10] The IDEM identity Federation: http://www.idem.garr.it.

[11] The JSAGA website: http://grid.in2p3.fr/jsaga/

[12] La Rocca G., Barbera R,, Ciaschini V, Falzone A., Monforte S. (2011). A new "lightweight" Crypto Library for supporting a new Advanced Grid Authentication Process with Smart Cards. Proceedings of Science (ISGC 2011 & OGF 31), 29.

[13] The eduGAIN inter-federation: http://www.edugain.org.

[14] The GrIDP federation: http://gridp.ct.infn.it.

[15] The Shibboleth System: http://shibboleth.internet2.edu

[16] Scardaci D., Scuderi G. (2007), A Secure Storage Service for the gLite Middleware, Proceedings of the Third International Symposium on Information Assurance and Security, p. 261-266.

[17] Calanducci A. et al. (2007), A Digital Library Management System for the Grid, Fourth International Workshop on Emerging Technologies for Next-generation GRID (ETNGRID 2007) at 16th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE-2007), GET/INT Paris, France, June 18-20, 2007.

Valeria Ardizzone
valeria.ardizzone@garr.it

Laureata presso l'Università degli Studi di Catania nel 2003, ha lavorato presso l'Istituto Nazionale di Fisica Nucleare nell'ambito della tecnologia Grid Computing, partecipando alle tre edizioni del progetto europeo "Enable Grid for E-sciencE". Dal 2008 al 2010 è stata responsabile dell'Attività Formazione in Commissione Calcolo e Reti dell'INFN.
Dal 2011 è dipendente presso il Consortium GARR come Technical Coordinator del Progetto DECIDE.
È stata membro del Program Committee in più di una ventina di Scuole Internazionali sul Grid Computing, ha partecipato come tutor esperto della tecnologia Grid Computing ad oltre 100 eventi di training in tutto il mondo ed ha collaborato attivamente alla realizzazione di diverse National Grid Initiative (NGI).