briefing paper

digital preservation europe

Un centro per i dati mondiali del clima: conservazione dei dati degli Earth System Model

L'aumento delle capacità di calcolo per la produzione di dati dei modelli di simulazione del sistema Terra ha creato nuove sfide per la loro archiviazione e conservazione a lungo termine.
Non tutti i dati prodotti dal sistema possono essere conservati in un in un archivio di lungo periodo. Un nuovo concetto di archivio sviluppato dall'ICSU (International Council for Science) World Data Center for Climate (WDCC) separa l'archiviazione dei dati, includendo una data di scadenza a livello di progetto scientifico, e l'archivio documentato di lungo periodo. Questo nuovo sistema produce un archivio di lungo termine completamente documentato e qualitativamente controllato con un catalogo dati ricercabile.

Introduzione

Il volume dei dati in uscita da un earth system models dipende dalla risoluzione spaziale, dagli intervalli temporali in uscita, dal numero di variabili e dal formato di archiviazione dei dati. Lo sviluppo tecnologico ha determinato l'incremento del potere di calcolo consentendo una migliore risoluzione spaziale e temporale dei modelli e l'integrazione in essi di nuovi processi fisici e chimici. Nonostante il costo di archiviazione per ogni Terabyte diminuisca continuamente, tale diminuzione non tiene il passo con l'aumento nel potere di calcolo e con i conseguenti ritmi di produzione dei dati. In aggiunta alle implicazioni finanziarie della conservazione dei dati, il costo della qualità e della sicurezza aumenta con l'ingombro totale dell'archivio. Di conseguenza non è fattibile conservare tutti i dati prodotti in un archivio di lungo periodo e le strategie di archiviazione di lungo periodo devono essere modificate nell'ottica di limitare l'espansione dell'archivio di lungo periodo.

Strategia per l'archiviazione di lungo periodo e l'utilizzo interdisciplinare dei dati

Questa strategia separa i dati in due categorie o cicli di vita. La prima categoria o stadio include la gestione dei dati di progetto per progetti scientifici di breve periodo, la seconda categoria include i dati adatti alla conservazione di lungo periodo. Questa separazione assicura che la decisione di spostare dati dalla prima categoria alla seconda per l'archiviazione di lungo periodo, prenda in considerazione e valuti sia il rigore di buone pratiche scientifiche sia la limitata disponibilità di risorse.

Anche se la trasmissione dei dati al WDCC è tipicamente il risultato finale di progetti scientifici mirati alla pubblicazione scientifica, le regole della buona condotta scientifica richiedono che questi dati siano disponibili e accessibili per almeno dieci anni, al fine di permettere future verifiche dei risultati pubblicati. Questi dati inoltre fanno parte del processo generale di creazione della conoscenza e potrebbero essere usati in settori scientifici interdisciplinari. Questo duale aspetto applicativo dell'archiviazione di lungo periodo, specialmente il riutilizzo interdisciplinare, richiede standard di conservazione, qualità e usabilità più elevati rispetto al livello di progetto scientifico. In aggiunta a questo, se, dopo un certo numero di anni, i risultati finali e i dati prodotti non possono essere facilmente riprodotti e gli utenti provenienti da diversi settori interdisciplinari non hanno la capacità di analizzare i dati dei modelli numerici, nell'archivio di lungo periodo, a fianco dei dati scientifici, sarà necessaria la presenza di documentazione informativa.

Il ciclo di vita atteso di queste due categorie di standard si riflette nelle loro date di scadenza. I dati che hanno raggiunto la loro data di scadenza saranno rimossi dal sistema quando viene emesso un avviso. I dati per cui viene richiesta la conservazione saranno trasferiti sull'archivio di lungo periodo, conferendo loro un ciclo-vita di conservazione di 10 anni o più, a seconda delle regole stabilite. I dati che non richiedono ulteriore archiviazione saranno rimossi dall'archivio in accordo con il proprietario dei dati.

I dati trasferiti sull'archivio di lungo periodo, saranno memorizzati sui nastri del sistema di archiviazione di massa DKRZ (Deutsches Klimarechenzentrum) insieme con la documentazione completa e una seconda copia di sicurezza. Il WDCC, con il suo catalogo interrogabile, sta ancora costruendo il cuore di questo livello di archiviazione. Per i server di nuova generazione il concetto di documentazione del WDCC sarà esteso per coprire tutti i dati in questo livello gerarchico, sia in file che nel database, permettendo ai dati dell'archivio di lungo periodo di essere completamente documentati ed interrogabili.

Il principale servizio di pubblicazione è offerto dal WDCC come un servizio aggiuntivo per dati di interesse generale che dovrebbero essere riportati in pubblicazioni scientifiche e quindi ricercabili nei cataloghi standard delle biblioteche insieme alle altre pubblicazioni scientifiche. Il processo primario di pubblicazione dei dati è stato sviluppato insieme con la Technical Information Library di Hannover (TIB), ed è stato implementato nel profilo STD-DOI (Scientific and Technical Data - Digital Object Identifier, http://www.std-doi.de).

I dati primari che sono identificati come entità indipendenti nel contesto della letteratura scientifica sono adatti per il processo di pubblicazione STD-DOI. Questi dati, insieme ai loro metadati, attraversano un processo di revisione e uno di garanzia di qualità prima che i metadati per pubblicazioni elettroniche e il persistent identifier (DOI/URN) siano assegnati. I metadati per pubblicazioni elettroniche e i corrispondenti identifiers saranno registrati nel catalogo della biblioteca del TIB. In questo modo i dati primari pubblicati potranno essere ricercati e ottenuti insieme alle pubblicazioni scientifiche standard. I metadati del profilo STD-DOI sono aperti per la raccolta automatica e per l'integrazione in sistemi informatici alternativi.

Conservazione dei dati, mantenimento della Qualità e Usabilità

Il nuovo ed espanso concetto di archiviazione di lungo termine del WDCC e del DKRZ è basato sulla necessità di provvedere una più comprensiva gestione dei dati a livello di archiviazione a lungo termine. Questo ovviamente limita la quantità di dati che possono essere trattati.

La conservazione del Bit stream sarà assicurata da copie aggiuntive su nastro archiviate in un altro luogo e in un altro formato. In aggiunta a questo sarà definito il numero massimo di accessi al nastro e il numero di accessi al nastro sarà registrato. Quando il numero massimo di accessi sarà raggiunto i dati sul nastro originale saranno migrati su un nuovo nastro e il vecchio nastro verrà rimosso.

Provvedere garanzia di qualità per i risultati dei modelli numerici è reso complesso dalla vasta mole di dati immessi e archiviati. Non è possibile controllare ciascun singolo numero in uscita. Per riuscire in questa sfida saranno effettuati dei test a campione a vari livelli di complessità. La garanzia di qualità degli output dei modelli numerici richiedono controlli semantici e sintattici. I controlli semantici implicano controlli del comportamento dei modelli numerici stessi confrontati con osservazioni e altri modelli.

Informazioni aggiuntive

Il WDCC fornisce dati e servizi per la ricerca degli Earth system con specifica enfasi sulla modellazione numerica e le relative osservazioni. Il WDCC, fondato nel 2003, è coordinato dal Model and Data Group del Max-Planck-Institute for Meteorology in collaborazione con il German Climate Computing Centre di Amburgo. Ad oggi il WDCC archivia e diffonde più di 340 Terabytes (TB) di dati degli earth system model e le relative osservazioni. Tutti i dati WDCC sono accessibili attraverso un'interfaccia web standard (http://cera.wdc-climate.de). Il WDCC ha approssimativamente 1000 utenti nominate e nel 2007 ha superato i 650,000 download per un totale di 200 TB. Il WDCC supporta anche la cooperazione internazionale per lo sviluppo di federazioni di archivi e reti..

Questo risulta come la maggior parte del processo di valutazione scientifica. Il controllo sintattico considera gli aspetti formali dell'archiviazione dei dati e assicura che l'archiviazione sia il più possibile priva di errori. Questo include l'esame della corrispondenza fra i metadati e i dati climatici, la completezza dei dati climatici e dei relativi metadati, i range standard dei valori, e il loro ordinamento temporale e spaziale.

Sebbene nella maggioranza dei casi possono essere condotte automaticamente, queste prove richiedono tempo. Tuttavia esse sono necessarie per garantire la validità dell'archivio.
La garanzia di qualità negli archivi di lungo termine WDCC e DKRZ è assicurata da un processo in tre stadi:
• il controllo semantico è effettuato ad un livello scientifico durante lo sviluppo del corrispondente progetto scientifico al fine di decidere sulla validità e utilità dei risultati dell'attuale modello. L'esito positivo è il criterio base per migrare i dati dalla gestione dei dati di progetto all'archivio di lungo termine;
• la documentazione dei dati e le routine di prove sintattiche sono condotte durante il processo di integrazione dei dati nell'archivio di lungo termine;
• la più completa disamina dei dati è condotta in connessione con il processori pubblicazione dei dati STD-DOI. La pubblicazione include revisioni più dettagliate e garanzie di qualità dei metadata e dei dati climatici.

L'usabilità dell'archivio di lungo termine sarà migliorata da una documentazione dei dati climatici completamente ricercabile nel sistema catalografico del WDCC. In aggiunta il WDCC offre un accesso web per quei dati climatici che sono salvati come campi individuali bidimensionali nelle tabelle del database. Ad oggi il WDCC offre accesso web a più di 340 TB di dati climatici che sono archiviati in 120,000 tabelle di database e in sei miliardi di voci di tabelle individuali. Le dimensioni medie di una singola voce della tabella è calcolabile in 60 KB e questo corrisponde al livello di granularità dell'accesso offerto.

L'usabilià dei dati deve inoltre essere supportata a un livello tecnico. Il trasferimento tecnologico dell'archivio deve essere compatibile verso il basso al fine di rendere leggibili i vecchi dati nel futuro. I relativi software dovranno essere trasferiti su nuove piattaforme. Gli strumenti per l'elaborazione dei dati e i formati di accesso ai dati devono poter continuamente funzionare anche con dati più vecchi provenienti dall'archivio di lungo termine.

Conclusioni

Il concetto del WDCC per la conservazione dei dati degli Earth system model qui presentato migliora l'affidabilità degli archivi di lungo termine nei limiti dei fondi disponibili.
A causa della crescita esponenziale del tasso di produzione dei dati, è stato scelto di focalizzare le risorse sulla tutela dei dati, come la documentazione degli archivi, conservazione del bit-stream, garanzia di qualità e usabilità, invece che sulla gestione tecnica di una vasta mole di dati.

Bibliografia

[1] ICSU World Data Center Climate (WDCC): http://www.wdc-climate.de

[2] German Climate Computing Centre (DKRZ): http://www.dkrz.de

[3] Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., Wächter, J.
Data Publication in the open access initiative Data Science Journal, Vol. 5, p79-83, 2006.

[4] Lautenschlager, M., Stahl, W.
Long-Term Archiving of Climate Model Data at WDC Climate and DKRZ
In: E Mikusch (Ed.): PV2007 - Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data, Conference Proceedings. DLR, German Remote Sensing Data Center, Oberpfaffenhofen, 2007

[5] Toussaint, F., Lautenschlager, M., Luthardt , H., World Data Center for Climate Data - Support for the CEOP Project in Terms of Model Output
Journal of the Meteorological Society of Japan, Vol. 85A, pp. 475-485, 2007

Autore: **Michael Lautenschlager**, *Model and Data / Max-Planck-Institute for Meteorology, Hamburg, Germany* , michael.lautenschlager@zmaw.de
Traduzione: Giorgio Trumpi, gtrumpy@gmail.com