briefing paper

digital preservation europe

Le Centre mondial des données climatologiques: la préservation des données du système de modélisation de la Terre

Les capacités croissantes de calcul pour la production des données du système de modélisation de la Terre ont créé de nouveaux défis pour son archivage à long terme et sa préservation. Toutes les données produites par le système ne pourront pas être conservées à long terme. Un nouveau concept d'archivage, développé par le ICSU (le Conseil international de la science) et le World Data Center for Climate (WDCC) différencie le stockage des données, qui comporte une date d'expiration au niveau du projet scientifique, et l'archivage à long terme, documenté. Ce nouveau système inclut donc un archivage à long terme entièrement documenté, un contrôle de qualité et un instrument de recherches des données.

Introduction

La taille des données en sortie du système des modèles dépend des résolutions spatiales, des intervalles temporels de sorties, du nombre de variables et du format de stockage de ces données. Les développements technologiques ont entraîné une augmentation des puissances de calcul conduisant à un modèle de résolution spatiale et temporelle plus fin, ainsi que l'adjonction de processus chimiques et physiques supplémentaires dans les modèles. Malgré la baisse continuelle des coûts de stockage du terra octet, il n'est pas possible de suivre ces puissances de calcul croissantes et les vitesses de production des données qui en résultent. Outre les implications financières de la préservation des données, le coût du contrôle qualité et de la fiabilisation des utilisations augmente avec le volume de la masse d'archives conservé. Par conséquent, il n'est pas possible, dans un archivage à long terme, de stocker toutes les données produites. La stratégie d'archivage à long terme doit être modifiée afin de limiter l'expansion des données à archiver.

Stratégie pour l'archivage à long terme et une utilisation interdisciplinaire des données

Cette stratégie sépare les données en deux catégories ou deux phases de cycle de vie. La première catégorie ou phase inclue le projet de gestion des données au niveau des projets scientifiques avec une durée de vie limitée ; la seconde catégorie ou phase comprend les données éligibles à l'archivage à long terme. Cette séparation assure une décision scientifique consciente pour le déplacement des données de la première catégorie à la seconde, pour l'archivage à long terme, prenant en compte à la fois la rigueur des bonnes pratiques scientifiques et la disponibilité limitée des ressources.

Les données transférées au WDCC sont principalement les résultats finaux des projets scientifiques sur lesquels les publications scientifiques sont basées ; les règles de bonnes pratiques scientifiques requièrent que ces données soient disponibles et accessibles pendant au moins dix ans afin de permettre des vérifications ultérieures sur les résultats publiés. Ces données participent à l'enrichissement de la connaissance générale et peuvent être utilisées dans des domaines scientifiques interdisciplinaires. Ce double aspect applicatif de l'archivage à long terme, et plus spécialement la ré utilisation interdisciplinaire, requière même plus de standards de préservation de données, plus d'assurance qualité et d'utilisabilité que pour les données au niveau du projet scientifique. De plus, comme les résultats finaux et les élaborations des données ne peuvent pas aisément être reproduits après un certain nombre d'années et que les utilisateurs de ces données, d'une audience interdisciplinaire plus large, pourront ne pas avoir les compétences nécessaires pour traiter ces modèles de données numériques, il est indispensable pour l'archivage à long terme de stocker les métadonnées de documentation avec les données scientifiques.

Le cycle de vie attendu de ces deux catégories de données est fourni par leur date d'expiration. Les données atteignant leurs dates d'expiration seront retirées du système après l'émission d'une alerte. Les données identifiées comme requerrant une conservation plus longue seront placées au niveau archivage à long terme, garantissant une durée de vie de dix ans et plus, selon les règles du centre de conservation. Les données qui ne nécessitent plus d'être conservées sont retirées de l'archivage, en accord avec le propriétaire.

Informations complémentaires

Le WDCC fournit des services et des données pour le système de recherches sur la Terre avec un accent spécifique sur la modélisation numérique et les observations qui lui sont liées. Le WDCC fondé en 2003 est géré par le groupe Modèles et Données de l'institut météorologique Max-Planck en collaboration avec le Centre allemand de calcul de climatologie à Hambourg.
A ce jour, le WDCC conserve et dissémine plus de 340 terra octets de données du système de modélisation de la Terre et des observations qui y sont rattachées. Toutes les données du WDCC sont accessibles via l'interface web standard (http://www.cera.wdc-climate.de). Environ 1000 utilisateurs sont enregistrés au WDCC ; en 2007 le WDCC a enregistré plus de 650 000 déchargements de données totalisant environ 200 TB.
Le WDCC favorise une coopération internationale pour le développement du travail en réseau et un archivage fédéré.

Les données placées au niveau archivage à long terme seront stockées sur des cassettes du système de stockage de masse du DKRZ (Deutsches Klimarechenzentrum) avec la documentation complète. Une copie de sécurité sera constituée. Le WDCC, grâce à ses instruments de recherches des données élabore actuellement le cœur de ce niveau d'archivage. Avec la nouvelle génération de serveurs le concept de la documentation du WDCC sera étendu afin de couvrir tout ce niveau hiérarchique, soit par des fichiers soit par une base de données, permettant ainsi aux données archivées à long terme d'être dûment documentées et interrogeables.

Le service de publication des données primaires est offert par le WDCC en tant que service additionnel pour les données d'intérêt général qui devraient être référencées directement dans les publications scientifiques et qui sont dès lors interrogeables, ainsi que ces publications, à travers les catalogues standards des bibliothèques. Le processus de publication de ces données primaires a été développé conjointement avec la bibliothèque d'information technique de Hanovre (TIB) et a été implémenté dans le profil du STD-DOI (Scientific and Technical Data – Digital Object Identifier, http://www.std-doi.de).
Les données primaires, qui sont identifiées comme des entités de données indépendantes dans le contexte de la littérature scientifique, rentrent dans le processus de publication de STD-DOI. Ces données ainsi que leurs métadonnées, subissent une procédure d'évaluation et d'assurance qualité avant que les métadonnées pour la publication électronique et pour l'identifiant persistant (DOI/URN) ne leurs soient assignées. Ces métadonnées pour la publication électronique et l'identifiant correspondant sont enregistrées dans le catalogue de la TIB. Dès lors, ces données primaires publiées ainsi que les publications scientifiques standards sont accessibles et interrogeables. Ces métadonnées du profil STD-DOI sont ouvertes aux robots du web et aux intégrations dans des systèmes d'information alternatifs.

La préservation des données, l'assurance qualité et l'utilisabilité

Le nouveau concept élargi de l'archivage à long terme du WDCC et du DKRZ repose sur le besoin de fournir une gestion plus complète des données au niveau de l'archivage à long terme. Ceci limite naturellement les masses de données qui pourront être prises en compte.
La préservation des chaînes de bits sera sécurisée par des copies supplémentaires de bandes dans des endroits distincts et dans un format différent. De plus, un nombre d'accès maximum est fixé et le nombre d'accès, pour une bande, est enregistré. Lorsque le nombre d'accès maximum est atteint, les données de la bande concernée sont migrées sur une nouvelle bande ; l'ancienne est retirée du silo.

Offrir une assurance qualité pour les résultats de modèles numériques est rendu complexe par la vaste quantité de données enregistrée et stockée. Il est impossible d'examiner un à un les chiffres produits. Pour répondre à ce défi, des sondages sont effectués, à des niveaux de complexité variés. L'assurance qualité des produits des modèles numériques requière des contrôles sémantiques et syntaxiques. Le contrôle sémantique entraîne l'examen du comportement du modèle numérique et des comparaisons avec d'autres observations et d'autres modèles. Tout ceci constitue la part principale de la procédure d'évaluation scientifique. Le contrôle syntaxique prend en compte les aspects formels de l'archivage des données et s'assure que l'archivage comporte le moins d'erreurs possibles. Cela comprend l'examen de la consistance entre les métadonnées et les données climatologiques, la complétude de ces données et de leurs métadonnées associées, les fourchettes des valeurs standard, et leurs classements temporels et spatiaux.

Même si dans la majorité des cas ces contrôles sont faits automatiquement, ils prennent du temps. Ils sont cependant nécessaires pour rendre fiables les données archivées.
Au WDCC et AU DKRZ l'archivage à long terme relève d'une procédure qui comporte trois phases :

• Le contrôle sémantique, qui est fait au niveau scientifique pendant le déroulement du projet afin de statuer sur la validité et l'utilité des résultats produits par le modèle. Le critère de base pour migrer les données du projet scientifique vers l'archivage à long terme, est l'établissement d'une décision positive.
• La documentation des données, qui est faite, ainsi que l'exécution des routines de contrôles syntaxiques, pendant la procédure d'intégration des données pour l'archivage à long terme.
• Les contrôles de données les plus complets, qui sont faits selon la procédure STD-DOI de publication des données. La procédure de publication comporte plus d'examens détaillés et plus d'assurance qualité des métadonnées et des données climatologiques.

L'utilisation de l'archivage à long terme sera améliorée par une documentation complète et interrogeable des entités de données climatologiques dans le système de catalogage du WDCC. En outre, le WDCC offre des accès web aux données pour les données climatologiques qui sont stockées en tant que champs bi dimensionnels dans les tables de la base de données. Actuellement, le WDCC offre un accès Internet à plus de 340 terra octets de données climatologiques stockées dans 120 000 tables de bases de données et six milliards d'éléments individualisés de tables. La taille moyenne d'une occurrence dans une table est calculée comme étant de 60 kilo octets ; ceci correspond au niveau de granularité d'accès offerts aux données.

L'utilisation des données doit aussi être possible au niveau technique. La technologie de transfert d'archives doit être de compatibilité descendante pour que les données anciennes puissent être techniquement lisibles dans le futur. Les logiciels associés devraient aussi être migrés vers de nouvelles plateformes. Les outils de traitements des données et les formats de données pour les accès aux bibliothèques nécessitent un travail continu, même avec les données plus anciennes de l'archivage à long terme.

Conclusions

Le concept du WDCC pour la préservation des données du système de modélisation de la Terre décrit ici améliore la fiabilité de l'archivage à long terme dans les limites budgétaires actuelles.
A cause de la vitesse d'accroissement exponentiel de la production des données, il a été décidé de concentrer les ressources sur la bonne gestion des données, telles la documentation archivistique, la préservation des chaînes de bits, l'assurance qualité et de larges possibilités d'utilisation, au lieu de se concentrer sur des manipulations techniques de vastes quantités de données.

References

[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

Michael Lautenschlager, *Model and Data / Max-Planck-Institute for Meteorology, Hamburg, Germany*
michael.lautenschlager@zmaw.de