briefing paper

digital preservation europe

Preservación de un sistema de modelo de datos de la Tierra

El aumento de la capacidad de proceso de los ordenadores para la producción de un sistema de modelo de datos de la Tierra ha creado nuevos retos para su preservación y archivación a largo plazo. No todos los datos producidos en el sistema pueden almacenarse en un archivo a largo plazo. Un nuevo concepto de archivo desarrollado por el ICSU (International Council for Science) y el World Data Center for Climate (WDCC) separa el almacenamiento de datos, incluyendo una fecha de caducidad en determinados niveles científicos del proyecto, y un archivo documentado a largo plazo. Este nuevo sistema produce una completa documentación, un archivo a largo plazo de calidad comprobada con un catálogo de búsqueda de datos.

Introducción

El tamaño de datos de salida desde los sistemas de modelos de la Tierra dependen de su resolución espacial, intervalos de salida temporales, el número de variables y formato de almacenamiento de los datos.
Los desarrollos tecnológicos han conllevado un incremento de la capacidad de proceso hacia un modelo de resolución temporal y espacial más preciso y a la integración en modelos de procesos físicos y químicos adicionales.
A pesar de que el coste de almacenamiento por TB decrece continuamente, no es suficiente debido a los incrementos en capacidad de proceso y las tasas de creación de datos conectados. Además de las implicaciones financieras de la preservación de datos, el coste de la verificación de la calidad y la formalización de la usabilidad, incrementa junto con el tamaño total de los archivos de almacenamiento masivo. Así pues, no es viable almacenar todos los datos producidos en un archivo de datos a largo plazo y tiene que modificarse la estrategia de archivación a largo plazo para limitar el crecimiento de un archivo de datos a largo plazo.

Estrategia para archivos a largo plazo y uso de datos interdisciplinarios

Esta estrategia separa los datos en dos categorías o etapas del ciclo de vida. La primera categoría o etapa incluye gestión de datos del proyecto en el nivel de proyectos científicos con una ciclo de vida limitado, la segunda categoría o etapa incluye datos accesibles disponible para archivar a largo plazo. Esta división asegura un conocimiento, la decisión científica de mover datos desde la primera categoría a la segunda para la archivación a largo plazo, teniendo en cuenta el rigor de la praxis científica y la disponibilidad limitada de recursos de forma equilibrada.

Como los datos transmitidos hacia el WDCC son típicamente los resultados finales de proyectos científicos basadas en publicaciones científicas, las reglas de la buena práctica científica requiere que estos datos deban estar disponibles y accesibles durante al menos diez años para permitir las futuras verificaciones de los resultados. Estos datos son también parte del proceso general de creación del conocimiento y podrían utilizarse en campos científicos interdisciplanarios. Este aspecto dual de aplicación de archivación a largo plazo, especialmente la reutilización interdisciplinaria, todavía requiere altos estándares de preservación de datos, comprobación de la calidad y usabilidad que datos en el nivel de proyectos científico. Además de esto, como los resultados finales y productos de datos no se pueden reproducir fácilmente después de un número de años y los datos de los usuarios de una amplia audiencia interdisciplinar podrían no tener las habilidades necesarias para manejarse con modelos numéricos de datos, hay un requisito en el archivo a largo plazo para almacenar metadatos de documentación a través de datos científicos.

Los ciclos de vida esperados de estas dos categorías de datos estándar se reflejan en sus fechas de caducidad. Los datos que lleguen a su fecha de caducidad se borrarán del sistema después de que el evento sea comunicado. La identificación de datos que se requerirá más adelante para su archivación se trasladará hacia el nivel de archivo a largo plazo facilitando un ciclo de vida de diez años o más de acuerdo con las normas de la biblioteca. Los datos que no requieran almacenamiento posterior se borrarán del archivo en coordinación con el propietario de los datos.

Información adicional

El WDCC facilita servicios y datos para la investigación del sistema de la Tierra con enfásis específico en modelos numéricos y observaciones relacionadas.

El WDCC, fundado en 2003, está siendo operado por el Grupo de Modelos y Datos en el Instituto de Meteorologia Max-Planck en cooperación on el Centro de Computación Alemán del Clima en Hamburgo.

Actualmente el WDCC archiva y difunde más the 340 Terabyte (TB) del sistema de modelos de datos y observaciones relacionadas.

Todos los datos del WDCC están accesibles a través de una interfície web estándar (http://cera.wdc-climate.de)

El WDCC tiene aproximadamente 1000 usuarios y en 2007 más de 650.000 datos fueron desacargados haciendo un total de 200 TB.

El WDCC también trabaja en cooperación internacional en desarrollos de redes y federación de archivos.

Los datos que han sido trasladados al nivel de archivo a largo plazo se almacenaran en cintas del sistema de almacenamiento masivo de la DKRZ (Deutches Klimarechenzentrum) junto con una completa documentación y una segunda copia de seguridad. La WDCC con su catálogo de búsqueda de datos está actualmente construyendo el núcleo de este nivel de archivo. Para la siguiente generación de ordenadores servidores, el concepto de documentación de WDCC se expandirá para cubrir todos los datos en este nivel de jerarquía bien en ficheros o en un sistema de base de datos, permitiendo que el archivo de datos a largo plazo esté completamente documentado y puedan buscarse datos en él.

El servicio primario de publicación de datos se ofrece por el WDCC como un servicio adicional para datos de interés general que deberían referenciarse directamente en publicaciones científicas y que pueden buscarse en catálogos de biblioteca estándar junto con las publicaciones científicas. El proceso primario de publicación de datos ha sido desarrollado junto con la Technical Information Library, Hannover (TIB), y ha sido implementado en el perfil STD-DOI (Scientifical and Technical Data – Digital Object Indentifier, http://www.std-doi.de).

Los datos primarios que se han identificado como entidades independientes de datos en el contexto de literatura científica están disponibles para el proceso de publicación del STD-DOI. Estos datos, junto con sus metadatos, pasan un proceso de revisión y un proceso de comprobación de la calidad antes de que se asignen los metadatos para publicaciones electrónicas y un identificador persistente (DOI/URN). Los metadatos para publicaciones electrónicas y su identificador correspondiente se registran en el correspondiente catálogo de la biblioteca del TIB. Ahora los datos primarios publicados se pueden localizar y acceder juntos con publicaciones científicas estándar. Los metadatos del perfil del STD-DOI están disponibles para la recolección de datos para archivación web y para la integración en sistemas alternativos de información.

Preservación de Datos, Comprobación de la Calidad y Disposición de la Usabilidad

El nuevo y expandido concepto de archivo a largo plazo del WDCC y del DKRZ está basado en la necesidad de facilitar de un servicio de administración de datos más entendible en un nivel de archivo a largo plazo. Esto naturalmente limita la cantidad de datos que se pueden manipular.

La preservación del bit stream será segura debido a copias adicionales de cintas en un lugar separado y en un formato diferente. Además se establece el acceso a un número máximo de cintas y se registra el número de accesos a cintas. Si el número máximo de accesos se alcanza, los datos de la cinta original se migran a una nueva cinta y la cinta vieja se elimina del silo de cintas.

Facilitar la verificación de la calidad para resultados de un modelo numérico se vuelve compleja por una cantidad ingente de datos introducidos y almacenados. No es posible inspeccionar cada único número en la salida. Para resolver este reto realizan test puntuales en varios niveles de complejidad. La comprobación de la calidad para las salidas de modelos numéricos requiere evaluaciones sintácticas y semánticas. Las evaluaciones semánticas comprende la evaluación del comportamiento de un modelo numérico en si mismo comparado con observaciones y otros modelos. Esto forma la mayor parte de los procesos de evaluación científicos.

La evaluación sintáctica considera los aspectos formales de la archivación de datos y verifica que la archivación está libre de cualquier error posible. Esto incluye examinar la consistencia entre los metadatos y los datos del clima, todos los datos del clima y sus relativos metadatos, el margen de valores estándar y la ordenación temporal y espacial.
Aunque en la mayoría de los casos pueden realizarse automáticamente, estas pruebas llevan tiempo realizarlas. Aún así son necesarias para asegurar la confidencialidad en los archivos de datos.
La comprobación de la calidad en un archivo a largo plazo del WDCC y del DKRZ se realiza como un proceso de tres etapas:
*La comprobación semántica se lleva a cabo a nivel científico durante el tiempo de ejecución del correspondiente proyecto científico para decidir la validez y la utilidad del actual modelo de resultados. Una decisión positiva es el criterio básico para migrar los datos desde la gestión de un proyecto de datos a un archivo da datos a largo plazo.
*la documentación de datos y las rutinas de prueba semántica se realizan durante el proceso de integración de datos en un archivo a largo plazo.
*el examen más completo de datos se realizará en conexión con el proceso de publicación de datos del STD-DOI. La proceso de publicación incluye una revisión más detallada y una verificación de los metadatos y los datos del clima.

La usabilidad de un archivo a largo plazo se mejorará con un sistema de búsqueda de documentación completa de las entidades de datos de clima en el sistema de catálogo del WDCC. Adicionalmente el WDCC ofrece acceso a bases de datos basadas en formato web para aquellos datos de clima que estén almacenados en campos individuales de dos dimensiones las tablas de la base de datos. Actualmente el WDCC ofrece acceso web a más de 340 TB de datos de clima que están almacenados en 120.000 tablas de bases de datos y en seis billones de entradas de tablas individuales. El tamaño medio de una entrada simple en la tabla se calcula para 60 KB y esto corresponde con el nivel de granuralidad de acceso a los datos ofrecidos.

La usabilidad de los datos debe sostenerse en el nivel técnico. La transferencia tecnológica del archivo debe ser custodiada de forma compatible para disponer de los datos viejos técnicamente legibles en el futuro. El software vinculado, debería migrarse a nuevas plataformas. Las herramientas de proceso de datos y el acceso a los formatos de los datos de las bibliotecas están continuamente necesitando trabajar aún con datos viejos desde un archivo a largo plazo.

Conclusiones

El concepto WDCC para preservación de un sistema de modelos de datos de la Tierra presentado aquí mejora la fiabilidad de una archivo a largo plazo dentro de unas limitaciones existentes presupuestarias.
Debido a la margen de crecimiento exponencial en la producción de datos, se determinó enfocar recursos hacia la administración de datos, tales como documentación de archivos, preservación de bit-stream, la verificación de la calidad y facilitando la usabilidad, en lugar manipulaciones técnicas de una vasta cantidad de datos.

Referencias

[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.
Publicación de datos bajo la iniciativa open acces
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, Actas de la conferencia. DLR, German Remote Sensing Data Center, Oberphaffenhofen, 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

Author:**Michael Lautenschlager**, *Model and Data / Max-Planck-Institute for Meteorology, Hamburg, Germany*
michael.lautenschlager@zmaw.de
Traduccion: **Juan-José Boté**, juanjobote@inalbis.net