De Bases de Datos a Espacios de Datos: la Nueva Realidad de la e-Ciencia

From Databases to Dataspaces: the New e-Science Reality

C. Mendoza
Centro de Física, Instituto Venezolano de Investigaciones Científicas (IVIC), Caracas, Venezuela,
y Centro Nacional de Cálculo Científico Universidad de Los Andes (CeCalCULA), Mérida, Venezuela.
claudio@ivic.ve

Resumen

Describimos las actividades de data que se han llevado a cabo en los últimos 20 años en el Proyecto de la Opacidad y el Proyecto del Hierro, principalmente el desarrollo y mantenimiento de la base de datos atómicos TIPTOPbase y el servidor interactivo de opacidades OPserver. Dentro del contexto del nuevo paradigma computacional de la e-Ciencia, discutimos los cambios de enfoque que se tienen que introducir para poder manejar la nueva escala en volúmenes y diversidad de datos, sobre todo en relación a los estándares de intercambio de datos basados en XML que se están estableciendo.

Abstract

We describe the data activities that have been carried out in the past 20 years within the Opacity Project and Iron Project, namely the development and maintenance of the TIPTOPbase atomic database and the OPserver interactive opacity server. In the context of the new computational paradigm of e-Science, we discuss the approach changes that must be introduced to manage the new scale in data volumes and diversity, especially in relation to the data interchange standards based on XML that are being established.

1. Introducción

En el presente reporte hacemos un recuento de los aciertos y problemas que se derivaron de nuestra participación en colaboraciones internacionales para llevar a cabo proyectos de computación de alto rendimiento a largo plazo (más de 20 años), sobre todo en relación a las actividades de data. Estas últimas se emprendieron con el propósito de facilitar el acceso a los datos por parte de los usuarios, pero tuvieron que sobrevivir en un entorno evolutivo muy dinámico marcado por el establecimiento de las bases de datos relacionales, la Internet, el World Wide Web y, más recientemente, la e-Ciencia. En la Sección 2, describimos el Proyecto de la Opacidad y del Hierro, sobre todo en relación al portafolio computacional que se utilizó (Sección 3), y las actividades de data que se llevaron a cabo (Secciones 4-5). En la Sección 6 introducimos el nuevo paradigma computacional de la e-Ciencia, y discutimos en la Sección 7 los cambios que se pronostican en relación al manejo de datos científicos debido a su diversidad y localización en depositorios distribuidos, donde se aparenta establecer el XML como estándar para el intercambio de datos (Sección 8). En la Sección 9 presentamos algunas conclusiones.

2. Proyecto de la Opacidad y Proyecto del Hierro

Desde mediados de la década de los Ochenta, hemos participado en colaboraciones internacionales a largo plazo para calcular los conjuntos masivos de datos atómicos que se necesitan en aplicaciones astrofísicas, específicamente el Proyecto de la Opacidad (OP) [1] y el Proyecto del Hierro (IP) [2]. Estas iniciativas han involucrado a grupos de investigación de Canadá, Francia, Alemania, el Reino Unido, los Estados Unidos y Venezuela.

La opacidad es la propiedad física de un plasma que determina la transferencia de luz, y su estimación implica el cálculo de las probabilidades de absorción radiativa de todas las transiciones entre estados ligado-ligado, ligado-libre y libre-libre para cada uno de los iones contenidos en el medio transmisor. Esto permite obtener opacidades promedio y aceleraciones radiativas a partir de las características de la mezcla química del plasma y sus propiedades termodinámicas tales como la temperatura y densidad electrónicas. Para lograr la precisión requerida (mejor que el 10%), los modelos atómicos que se utilizan están caracterizados

datos para modelar plasmas fuera del equilibrio

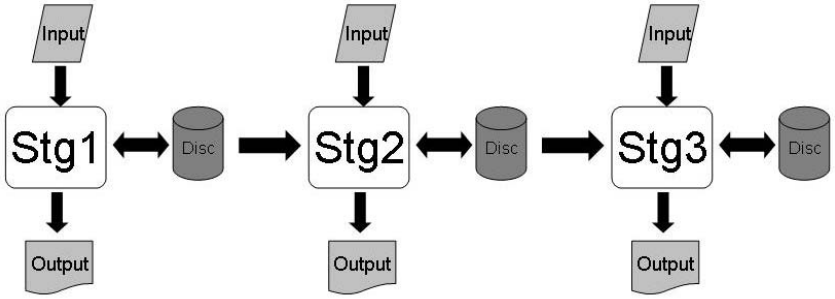


Figura 1. Estructura por etapas de un cálculo de la Matriz-R [6] indicando los archivos de entrada, salida e intermedios, Estos últimos suelen ser voluminosos.

por expansiones mecano-cuánticas complejas que implican cómputos a gran escala. La opacidad y las aceleraciones radiativas son componentes integrales en los modelos numéricos de estructura y evolución estelar, cuyas precisiones fueron seriamente cuestionada a comienzos de los Ochenta [3] dando lugar a varios proyectos de revisión, entre ellos el OP y el OPAL [4]. El éxito final de estas iniciativas queda registrado con el excelente grado de acuerdo que finalmente se logra entre ellas, lo que ha invalidado nuevas solicitudes de revisión basadas, por ejemplo, en re-estimaciones de las abundancias solares [5].

El IP, por otra parte, se ha concentrado en el cálculo de probabilidades de transición radiativa y secciones eficaces de excitación electrónica para los iones del elemento hierro. Estos parámetros permiten implementar diagnósticos para determinar las condiciones de los plasma astrofísicos (temperatura y densidad electrónicas, las abundancias y fracciones de ionización) aprovechando las líneas del hierro que se observan en la mayoría de los espectros astronómicos. Debido al alto grado de ionización de algunos de los iones del hierro de interés, el cálculo de las propiedades atómicas antes mencionadas implica – además de los complejos modelos atómicos ya mencionados – correcciones relativistas que complican los cómputos enormemente, acercándolos a los límites de las capacidades actuales de la computación de alto rendimiento.

Aunque estos proyectos han sido notoriamente productivos, tanto en número de publicaciones como en volúmenes de datos de confiable calidad, están lejos de satisfacer las necesidades actuales de datos atómicos en la astrofísica: por ejemplo, una base de (NLTE); niveles energéticos para estados de vacancia K y L con precisión espectroscópica y sus cascadas de decaimiento radiativo y Auger; átomos pesados (Z > 28) y datos moleculares en general. En la investigaciones de reactores de fusión tokamak, existe también la necesidad urgente de datos atómicos para el elemento pesado tungsteno (Z = 74). Por lo tanto, podemos concluir que existe una necesidad omnipresente para cálculos masivos de datos atómicos y moleculares.

3. Paquete computacional de la Matriz-R

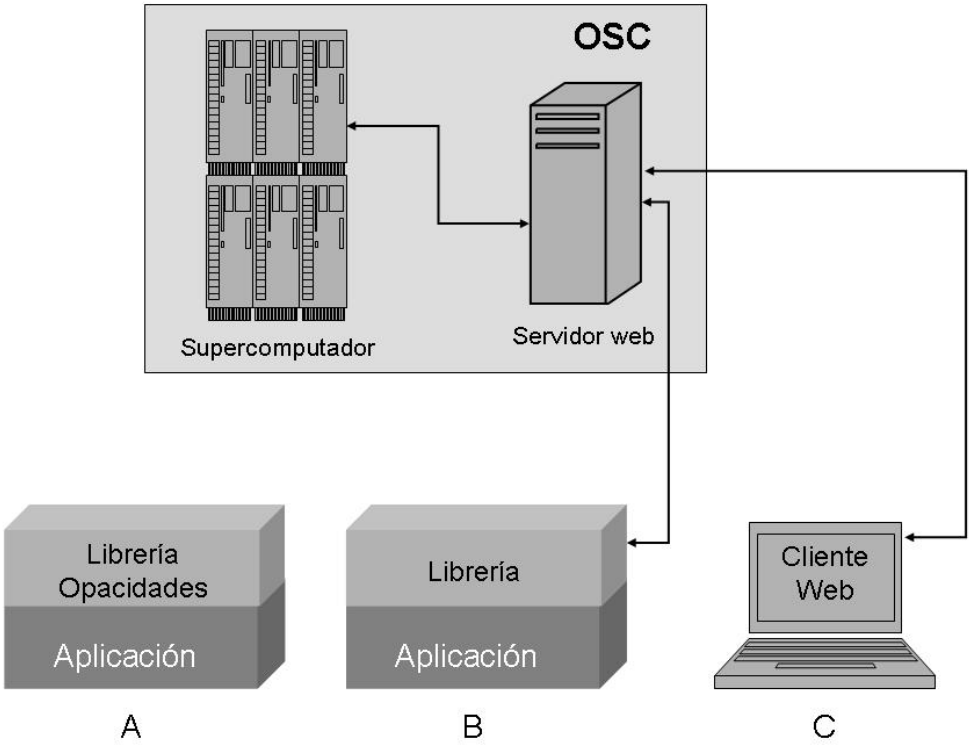
Los cómputos que se han llevado a cabo en el OP e IP se basan en el portafolio computacional de la Matriz-R [6] que esencialmente resuelve el problema mecano-cuántico de dispersión multi-canal de un electrón por un blanco atómico. Este último se modela con programas de estructura atómica tales como SUPERSTRUCTURE [7] y CIV3 [8], labor que requiere las destrezas de un experto. Como se indica en la Figura 1, una característica particular de estos cálculos es que se estructuran en una cadena de etapas interconectadas por archivos intermedios que pueden ser de gran volumen.

4. TIPTOPbase

En el OP e IP no sólo hemos atendido los cálculos de los datos atómicos básicos y derivados sino también

los aspectos relacionados a su acceso y utilización por

atributos. Tiene también una bitácora de búsquedas



la comunidad astronómica. Estas actividades comenzaron en el Centro Científico de IBM de Venezuela con bases de datos operadas con interfaces de usuario basadas en comandos mucho antes de la popularización de la Internet, y con la aparición el WWW a mediados de la década de los Noventa, progresivamente se conformaron en la base de datos en línea TIPTOPbase [9, 10]. El apoyo de los centros de datos, específicamente el Centre de Donées astronomiques de Strasbourg (CDS) [11] y el NASA High Energy Astrophysics Science Archive Research Center (HEASARC) [12], en el desarrollo, residencia, instalación y mantenimiento de TIPTOPbase ha sido decisivo.

TIPTOPbase contiene datos para iones con número atómico $1 \leq Z \leq 28$, y permite dos tipos de búsquedas: una basada en tablas de contenido activas y otra por medio de consultas delimitadas por rangos de

que guarda en cada sesión un registro de las consultas hechas por el usuario y permite bajar múltiples archivos de datos al final de la sesión. Cada vista de la base de datos genera tablas de números con sus respectivas unidades, hiperenlaces que apuntan hacia las fuentes de documentación (artículos en pdf o resúmenes) y permite procesamiento gráfico por medio de applets de Java interactivos.

5. OPserver

Las actividades de data dentro del IP/OP no sólo se han limitado a la implantación de bases de datos (TIPTOPbase, ver Sección 4) sino también a servidores que calculan datos derivados en línea. Este es el caso del OPserver [13], un servidor interactivo de opacidades y aceleraciones radiativas para aplicaciones

astrofísicas. Como se muestra en la Figura 2, este servidor, localizado en el Ohio Supercomputer Center (OSC) [14], permite tres modos para calcular opacidades promedio y aceleraciones radiativas desde la aplicación del usuario. En el modo A, el usuario baja desde el OSC e instala localmente tanto la base de datos de opacidades como una librería de funciones de comunicación que empalma a su aplicación. Esta opción requiere al menos 1 Gb de RAM ya que la base de opacidades se carga y maneja en memoria principal durante la ejecución de la aplicación. En el modo B, sólo se baja e instala la librería de funciones, accediendo a la base de opacidades remotamente en el OSC. Esta opción está destinada a ambientes de cómputo distribuido en grid. En el modo C, el usuario solicita un cómputo de opacidades en el OSC desde un cliente web, bajando los datos al disco local para ser entonces post-procesados por su aplicación.

6. La e-Ciencia

A pesar de todos los esfuerzos que se han hecho en el OP/IP para facilitar el acceso y uso de la data por parte de los usuarios, estamos conscientes que son insuficientes dentro del nuevo paradigma computacional que se establece rápidamente y se empieza a conocer como "e-Ciencia" [15]. Usando las propias palabras de uno de sus principales promotores, John Taylor (ex-Director General de los Research Councils, Office of Science and Technology, Reino Unido), "la e-Ciencia es ciencia computacionalmente intensiva". También ha declarado con mucha convicción que "la e-Ciencia está cambiando la dinámica de hacer ciencia". En realidad es ciencia computacionalmente intensiva pero a mayor escala, donde colaboraciones distribuidas globalmente se manejan sobre una Internet de segunda generación para minar grandes volúmenes de datos científicos y llevar a cabo simulaciones de alto rendimiento en la tera-escala y con visualizaciones sofisticadas. En el presente, la e-Ciencia es propulsada principalmente por campos tradicionalmente basados en datos como la física de partículas (Large Hadron Collider), la fusión nuclear (International Thermonuclear Experimental Reactor) y la astronomía (Virtual Observatories), pero también por nuevas áreas como las biociencias (la genómica, proteómica, farmacogenética, bioinformática), climatología y las ciencias sociales.

Como la física atómica computacional siempre ha respondido a las demandas y retos de disciplinas como la astrofísica y la física de plasmas de fusión, los cómputos masivos y actividades de data están destinados a prosperar en el nuevo ambiente de la e-Ciencia, y se deben ajustar a los requerimientos y formatos que exigen las nuevas organizaciones virtuales; por ejemplo, los observatorios virtuales. Sin embargo, nos vemos en la necesidad imperiosa de adaptar los portafolios de códigos computacionales (ver Sección 3) para trabajar en entornos distribuidos de grid y de redimensionar las bases de datos (ver Sección 4) hacia el nuevo concepto de "espacio de datos". Ya que los aspectos relacionados al grid van a ser tratados en esta reunión por Oldenhof & Mendoza, entre otros, discutimos en las próximas secciones el innovador concepto de "espacio de datos".

7. De bases de datos a espacios de datos

En el nuevo paradigma de la e-Ciencia, la mayoría de los conjuntos de datos científicos, debido a los volúmenes que se tienen que manejar, van a estar almacenados en repositorios distribuidos que no cumplen necesariamente con un modelo estándar de datos o la estructura específica de un manejador de bases de datos (DBMS). Como se indica en la Figura 3, la proximidad administrativa e integración semántica de dichos conjuntos de datos pueden ser muy diferentes; por ejemplo, los portales corporativos y organizaciones virtuales están lejos de alcanzar el grado de integración de un DBMS. Más aún, en relación a los diversos DBMSs con datos atómicos que se mantienen actualmente, se ha desarrollado un motor de búsqueda general y bastante útil bajo el nombre de GENIE [16], el cual actúa como un sistema de integración de datos. Sin embargo, no incluye búsquedas en el web.

Según Franklin et al. [17], los desarrolladores enfrentan ahora retos tales como implantar nuevos procedimientos de búsqueda y consultas en conjuntos de datos muy diversos y distribuidos; establecer nuevas convenciones para etiquetar, identificar e integrar la data; controlar el acceso, disponibilidad y recuperación; y gerenciar la evolución de la data y metadata. Más aún, proponen el nuevo concepto de "espacios de data" como abstracción para el manejo de data, y el desarrollo de "Plataformas de Soporte a Espacios de Datos" (DSSPs) en sustitución de los tradicionales DBMSs relacionales. Esto permite al desarrollador concentrarse en su aplicación más que en intentos infructuosos de integrarse a esquemas de manejo de datos de acentuada diversidad. El "espacio de datos" no es un enfoque de integración de datos sino de coexistencia de datos. El DSSP debe manejar datos y aplicaciones en una variedad de formatos sin

tener el control total sobre algunos grupos de ellos. Las consultas de datos entonces tienen que ofrecer diferentes niveles de servicio y en algunos casos retornar con respuestas aproximadas.

el cual hemos participado, que desarrolla un dialecto XML específico para el intercambio de datos atómicos y moleculares bajo el nombre de "Atomic and Molecular Data Markup Language" (AMDML). Éste

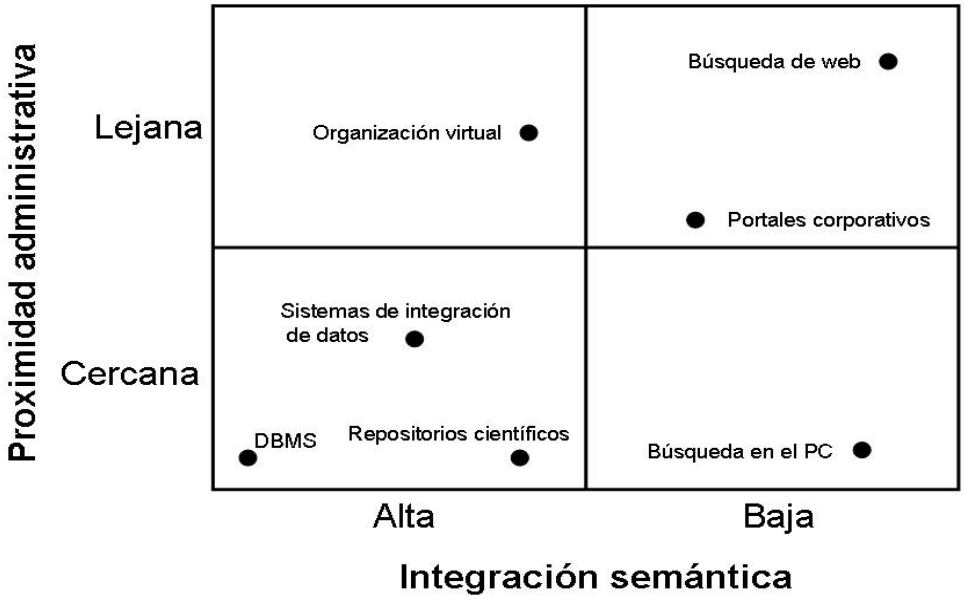


Figura 3. Un espacio para la solución del manejo de datos. Fuente: Ref. [17].

Una importante innovación en este contexto es el nuevo formato de documento abierto OASIS (ISO/IEC 26300) [18] basado en XML que permite la edición, almacenamiento e intercambio de datos de oficina (texto, hojas de cálculo, base de datos, gráficos y presentaciones) independientemente de la plataforma y aplicación. Otra es Google Docs & Spreadsheets [19] que ofrece el manejo de documentos centrado en la red, específicamente las posibilidades de crear, compartir, almacenar y publicar documentos utilizando el browser gratuitamente.

8. Data XML

Con el establecimiento de estándares como el OASIS (ver Sección 7), XML se está rápidamente convirtiendo en el vehículo por excelencia para intercambiar datos en el web. En este sentido, Yuri Ralchenko del National Institute of Standards and Technology (NIST) dirige un grupo internacional, en le dará a los archivos de datos atómicos y moleculares una semi-estructura flexible pero estandarizada que promoverá que se cataloguen en las búsquedas de Google.

Sin embargo, como discute Freire & Benedikt [20], la tecnología XML todavía no ha llegado a la madurez sobretodo en problemas básicos como almacenamiento, acceso, publicación y consultas. La data todavía se va a almacenar en DBMSs relacionales con consultas bajo lenguajes como SQL, lo que implica la implantación de interfaces de mapeo a XML que pueden ser bastante complicadas de lograr. Más aún, el parseo, validación y mantenimiento de archivos XML todavía son ineficientes.

9. Conclusiones

Estamos ya involucrados en un nuevo paradigma computacional que se empieza a conocer como e-Ciencia, caracterizado por la minería de grandes

volúmenes de datos científicos almacenados en repositorios distribuidos globalmente. En este contexto, los proyectos de computación de alto rendimiento que generen datos deben dedicar cierto esfuerzo al desarrollo de bases de datos de segunda generación ("espacios de datos" más que bases de datos) y a la definición de dialectos XML para el intercambio e indexación de dichos datos.

Referencias

[1] http://cdsweb.u-strasbg.fr/topbase/op.html
[2] http://www.usm.uni-muenchen.de/people/ip/iron-project.html
[3] Simon N.R., "A plea for reexamining heavy element opacities in stars", 1982, *Astrophys. J.*, 260, L87
[4] Iglesias C.A., Rogers F. J., "Updated Opal Opacities", 1996, *Astrophys. J*, 464, 943
[5] Asplund M., Grevesse N., Sauval A. J., "The Solar Chemical Composition", 2005, en Barnes T. G., Bash F. N., eds, ASP Conference Series Vol. 336, *Cosmic Abundances as Records of Stellar Evolution and Nucleosynthesis*, Astronomical Society of the Pacific, San Francisco, p. 25
[6] Berrington K.A., et al., "A new version of the general program to calculate atomic continuum processes using the R-matrix method", 1978, *Comput. Phys. Commun*., 14, 367
[7] Eissner W., Jones M., Nussbaumer H., "Techniques for the calculation of atomic structures and radiative data including relativistic corrections", 1974, *Comput. Phys. Commun.*, 8, 270
[8] Hibbert A., "CIV3 - A general program to calculate configuration interaction wave functions and electric-dipole oscillator strengths", 1975, *Comput. Phys. Commun*., 9, 141
[9] Cunto W., Mendoza C., Ochsenbein F., Zeippen, C. J., "Topbase at the CDS", 1993, *Astron. Astrophys.*, 275, L5
[10] http://cdsweb.u-strasbg.fr/topbase/home.html
[11] http://cdsweb.u-strasbg.fr/
[12] http://heasarc.gsfc.nasa.gov/
[13] Mendoza C., et al., "OPserver: interactive online computations of opacities and radiative accelerations", 2007, *Month. Not. R. Astron. Soc*., 378, 1031
[14] http://www.osc.edu/
[15] Hey T., Trefethen, A. "The data deluge: an e-Science perspective", en *Grid Computing: Making the Global Infrastructure a Reality*, 2003, Berman F., Fox G, Hey A.J.G. (Eds), Wiley, New York, p. 809
[16] http://www-amdis.iaea.org/GENIE/
[17] Franklin M., Halevy A., Maier D., "From databases to dataspaces: a new abstraction for information management", 2005, *SIGMOD Record*, 24, 27
[18] http://www.oasisopen.org/
[19] https://docs.google.com/
[20] Freire J., Benedikt M., "Managing XML data: an abridged overview", 2004, *Comp. Sc. Eng.*, 6, 12