

El Observatorio Virtual

EL OBSERVATORIO VIRTUAL SE ESTÁ REVELANDO COMO UN EFICAZ INSTRUMENTO CIENTÍFICO
 Por Juan de Dios Santander (IAA-CSIC)

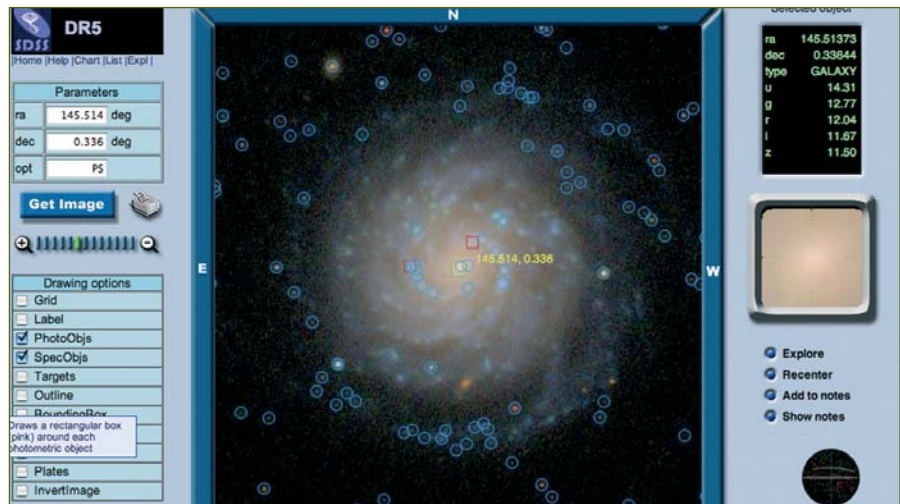
YA EN 1968, ARTHUR C. CLARKE ESCRIBIÓ EN SU OBRA 2001: UNA ODISEA ESPACIAL:

Desde que orbitaron los primeros satélites, hacía unos cincuenta años, billones y cuatrillones de impulsos de información habían estado llegando del espacio, para ser almacenados para el día en que pudieran contribuir al avance del conocimiento. Solo una minúscula fracción de esa materia prima sería tratada; pero no había manera de decir qué observación podría desear consultar algún científico, dentro de diez, o de cincuenta, o de cien años. [...] Formaban parte del auténtico tesoro de la Humanidad, más valioso que todo el oro encerrado inútilmente en los sótanos de los bancos.

Clarke siempre fue un visionario, y dio en el clavo cuando intuyó que los bancos de datos de información astronómica irían creciendo enormemente con el tiempo, y que la preservación de esa información, para obtener nuevos resultados de ella, sería clave.

En la actualidad estamos asistiendo a este proceso: se está generando una ingente cantidad de datos provenientes de un número creciente de instrumentos astronómicos con cada vez mejor resolución, lo que está multiplicando las necesidades de almacenamiento. Por ejemplo, se han puesto en marcha grandes proyectos de estudio exhaustivo del cielo, como el *Sloan Digital Sky Survey* (SDSS), que generan por sí mismos billones de datos. Se prevé que solo los datos elaborados del SDSS ocuparán el equivalente a 50 unidades de disco duro de 80GB, o cerca de 1000 DVDs.

Como Clarke vivía en un tiempo en que los ordenadores no se conectaban entre sí, y sólo se comunicaban mediante el intercambio de discos, no pudo anticipar -al menos en



su obra 2001- el advenimiento de Internet: un sistema de conectividad global que permite que cualquier ordenador se comunique con otro, sin importar la localización geográfica ni las diferencias entre sistemas operativos. Además, las capacidades de almacenamiento y el ancho de esas conexiones no han hecho más que crecer.

Sin embargo, eso no resuelve el problema del todo: con una ADSL de 20 megas se tardaría más de 18 días en bajar todos los datos del SDSS. Ahora pensemos que miles de astrofísicos necesitan acceder a esos mismos datos, ¿qué ordenadores necesitarían para tratar la información? y ¿cómo encontrarían lo que buscan? Parafraseando a Jack Swigert del Apolo 13: "Tenemos un problema aquí" Para resolverlo, se necesita:

1. Que los diferentes instrumentos unifiquen la forma en que publican su información: así cualquier programa podrá acceder a cualquier dato astronómico, y no habrá que crear, ni saber usar, un programa por instrumento.
2. Que los datos elaborados por un grupo de investigación puedan ser utilizados por cualquier otro. Aparte del uso de un formato común, es necesario adjuntar la lista de operaciones que se realizaron con los datos originales, para determinar su calidad, y poder recuperarlos y revisar su tratamiento en caso necesario.
3. Que el movimiento de la información sea mínimo: los análisis se realizarán en los ordenadores que contengan los datos, y sólo viajarán los resultados de los análisis.
4. Que existan programas que encuentren

automáticamente relaciones entre conjuntos de datos diferentes, sin tener que descargar ninguno de esos conjuntos de datos; esto se conoce como "minería de datos".

5. Que los datos disponibles estén permanentemente actualizados, y que se notifique automáticamente la aparición de nuevos servicios de datos astrofísicos, o incluso de modelos teóricos para poder hacer comparaciones.

El Observatorio Virtual (OV) es la infraestructura que proporcionará esas capacidades, y las siguientes secciones nos servirán para ver cómo.

1. Unificación de formatos

La comunidad astronómica, desde finales de los años 70, ha contado con un formato común: el formato FITS, *Flexible Image Transport System*. Sin embargo, este formato común es exclusivo de la astrofísica; además, está lleno de modismos propios de cada instrumento, puesto que los avances tecnológicos hicieron necesario añadir nuevas capacidades, muchas veces de forma independiente para instrumentos similares. Como el número de instrumentos y extensiones era al principio manejable, la necesidad de que el formato de los archivos fuese realmente único era menor: bastaba con el uso de las herramientas de manipulación de datos propias de cada instrumento.

Por otra parte, con el avance de las comunicaciones, dentro del mundo de la empresa se ha ido imponiendo un estándar de descripción de datos, que comparte algunas similitudes con las necesidades de la astrofísica. Y

es que en el mundo de los negocios, casi todo el mundo necesita hablar de lo mismo: inventario, unidades vendidas, ingresos... aunque nadie haga negocios exactamente igual. Así que si las empresas quieren compartir información, es necesario poder describir esas diferentes formas de hacer negocios. Del mismo modo, en astrofísica, los científicos estudian fenómenos similares, con instrumentos parecidos, pero nadie lo hace exactamente de la misma forma. Así que el problema de comunicación de las empresas y de la astrofísica es parecido. Este estándar de comunicación común a la astrofísica y los negocios es el XML. XML es un lenguaje de marca, esto es, un lenguaje que se introduce junto con los datos para marcarlos, especificando la función o el papel de cada uno de esos datos, para eliminar ambigüedades. Tiene la ventaja de que existen multitud de herramientas que pueden manipular archivos XML -por su origen técnico/empresarial-, frente al formato FITS, que sólo puede ser manipulado por herramientas específicas de la astrofísica.

Sin embargo, XML es incluso más flexible que FITS, lo puede llevarnos a la falta de uniformidad en la forma de aplicar XML sobre los datos astrofísicos. Ahí entra en juego la Alianza Internacional del Observatorio Virtual (IVOA, *International Virtual Observatory Alliance*), que tiene el papel de normalizar los modelos de datos astrofísicos y los protocolos de acceso (o cómo conversan los diferentes ordenadores entre sí). Algunos ejemplos de modelos de datos por normalizar serían: cómo se especifica y almacena el rango de frecuencias en el que opera el filtro del telescopio, cómo se proporcionan las coordenadas de un objeto en el cielo o la cantidad de luz que recibimos del mismo.

Así, los modelos de datos y protocolos establecidos por IVOA, utilizando XML como lenguaje de marca, unifican la forma de describir y comunicar datos astrofísicos, y per-

miten la creación de programas que sirvan para múltiples instrumentos de forma más sencilla.

2. Historial de operaciones

El viaje de la luz hasta nosotros es bastante laborioso, y hay que tener en cuenta muchos efectos. Por ejemplo, la luz que llega desde una supernova que se encuentra en una galaxia muy lejana llegará enrojecida por el efecto Doppler, con una determinada polarización, atravesará la atmósfera terrestre, y se atenuará por el camino. Cuando llegue al sensor, se registrará su energía con un cierto grado de precisión, y existirán unos deter-



Observatorios Virtuales nacionales o supranacionales que son miembros de IVOA. El SVO es el Observatorio Virtual Español.

minados niveles de ruido. La información registrada será procesada con un determinado conjunto de programas usando ciertos parámetros, y se publicará, finalmente, una imagen, un espectro, o alguna otra unidad científica de información, cuya interpretación depende de la forma en que se observó (por cuánto tiempo se hizo, hacia dónde apuntaba en concreto el instrumento, las condiciones del cielo, los ajustes del instrumento, etcétera). Asimismo, también dependerá de qué programas se utilizaron para tratar los datos, de los procesos que se siguie-

ron, en qué orden, y con qué ajustes.

El registro de forma sistemática y estandarizada de las operaciones realizadas sobre los datos nos permitirá establecer la calidad de los mismos, y determinar si son aptos para el estudio que se quiere realizar, o si es necesario volverlos a analizar.

3. Análisis remoto

Recordemos el SDSS, ese estudio del cielo que ocuparía miles de DVDs. La forma más sencilla de recibir toda esa información sería, aparentemente, pedir esos mil DVDs, o bien ocho enormes discos duros con los datos. Aún así, necesitaríamos ordenadores capaces de manejar semejante volumen de información. Los PC típicos solo pueden manejar una milésima parte de toda esa información de una sola vez... y las estaciones de trabajo (ordenadores de uso científico-técnico) más potentes solo podrían manejar entre un uno y un cinco por ciento. No todo el mundo puede permitirse ordenadores tan grandes, y al trabajar por separado se pierden muchas posibilidades de encontrar interrelaciones entre los datos y los estudios de diferentes grupos.

La solución es hacer que los datos viajen lo mínimo posible, y para ello es necesario proporcionar programas que permitan a los usuarios acceder a datos ya elaborados -con lo que se optimiza el tiempo de cómputo-, o realizar sus propias manipulaciones sobre los datos remotamente. Esas manipulaciones quedarán registradas de forma automática, de modo que quien utilice esos datos sabrá qué es lo que se ha hecho exactamente con los originales.

También existe una tendencia a lo que se llama "red de computación", que consiste en permitir el acceso a los recursos de computación de la misma forma en que accedemos a la red eléctrica: estamos conectados a la red y pagamos por su uso. Eso permite repartir las tareas de proceso, cuando sea necesario, a la red, y la propia tarea determinará el tipo de máquina o máquinas que

OV en Internet

IVOA <www.ivoa.net> Página de la Alianza Internacional del Observatorio Virtual. Nivel avanzado.

SVO <laeff.esa.es/svo> Página del Observatorio Virtual Español. Dispone de materiales de formación en castellano sobre el Observatorio Virtual. Nivel medio-alto.

Aladin Sky Atlas <aladin.u-strasbg.fr> *Aladin* fue el prototipo europeo de OV, y es igualmente atractivo para aficionados y profesionales. El programa se puede descargar de forma gratuita, y accede a múltiples fuentes de información gracias a los protocolos del OV. Las últimas versiones integran *VOspec* en el mismo paquete. Los materiales ofrecidos -en inglés- permiten

aprender mucho tanto de *Aladin* como de astrofísica. Nivel medio-alto.

Datascopie <heasarc.gsfc.nasa.gov/vo> Portal del *National Virtual Observatory* de los EEUU que permite acceder a toda la información públicamente disponible de múltiples clases de instrumentos: radio, infrarrojo, luz visible, ultravioleta, rayos-X, rayos Gamma... Nivel medio-alto.

SDSS <www.sdss.org> El *Sloan Digital Sky Survey* es una recopilación sistemática de imágenes que cubrirán la mitad norte del cielo, con cinco filtros diferentes, con el objetivo de proporcionar un mapa en tres dimensiones para cerca de un millón de galaxias y cúasares. Para los objetos más brillantes, proporcionará espectros detallados. Existen materiales educativos -también en inglés- que enseñan, por ejemplo, cómo encontrar asteroides.

necesite. De esa forma se pueden aprovechar recursos de super-computación disponibles en otro lugar sin tener que invertir en capacidad de cálculo adicional. Esto se complementa con la posibilidad de crear grandes sistemas de cómputo asociando muchos ordenadores pequeños, con lo que la capacidad de computación de "la red" se podría ampliar tanto como se quisiera.

4. Minería de datos

Con volúmenes de datos tan grandes, no sólo se vuelve difícil su tratamiento, sino también la extracción de resultados. El mejor detector de patrones que conocemos es el cerebro humano... siempre que el conjunto de patrones por descubrir, o el ámbito en el que se busquen, sea reducido, porque si no llegan el cansancio y el aburrimiento, y se deja de ver coincidencias. Si queremos buscar, literalmente, una aguja entre mil millones de pajares, necesitamos utilizar ordenadores.

La minería de datos pretende encontrar conexiones entre diferentes variables, que muchas veces no son evidentes. Se utilizan técnicas clásicas de inteligencia artificial, como redes neuronales o análisis de probabilidad, basadas en el entrenamiento de un programa para reconocer los patrones que queremos encontrar y clasificarlos con una posibilidad de error que queremos minimizar. Algunas de esas técnicas permiten localizar objetos exóticos o inusuales que no se ajustan a ninguna clasificación particular, por lo que se espera que se puedan realizar nuevos descubrimientos aplicando esas técnicas a los datos ya existentes.

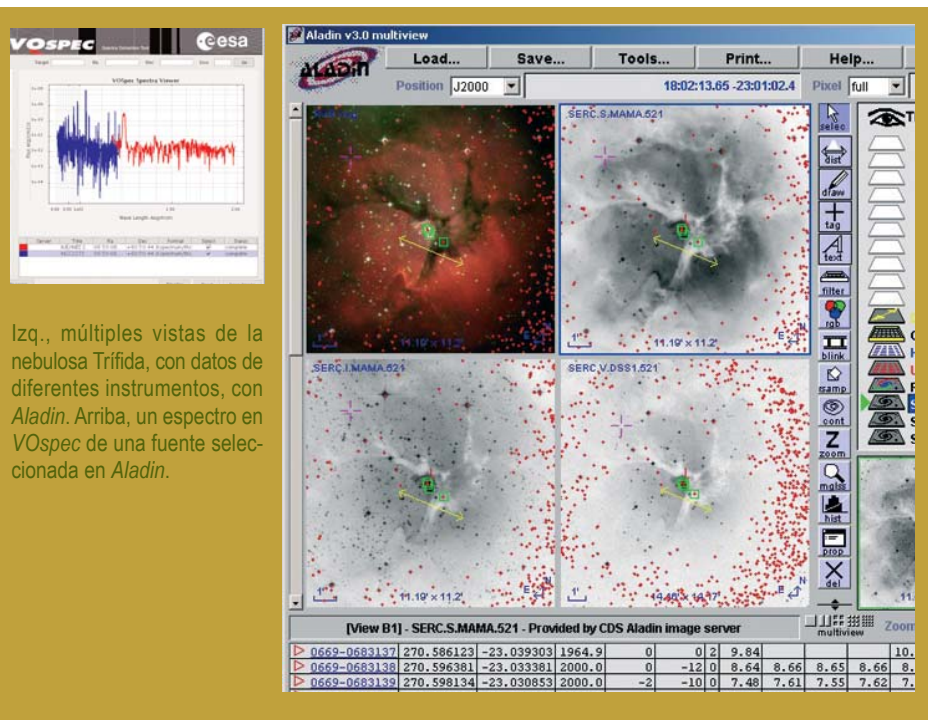
El OV pretende utilizar los avances disponibles en minería de datos, enriqueciéndolos con las descripciones detalladas proporcionadas por los modelos de datos, así como la unificación de formatos, para crear herramientas de minería de datos específicas de la astrofísica.

Un ejemplo es la misión *Gaia*, que pretende catalogar cerca de mil millones de estrellas de nuestra galaxia, y obtener características espectrales para todas ellas, a partir de los datos obtenidos por un satélite específico que se espera lanzar en 2011. Ese trabajo no se puede realizar manualmente, y se utilizarán técnicas de minería de datos para clasificar las estrellas automáticamente. El entrenamiento previo se está realizando con los datos de misiones anteriores, y se podrá refinar con los propios datos de la misión.

5. Actualización automática y registro de datos

Con todo lo que hemos descrito hasta ahora,

Izq., múltiples vistas de la nebulosa Trífida, con datos de diferentes instrumentos, con *Aladin*. Arriba, un espectro en *VOspec* de una fuente seleccionada en *Aladin*.



podemos establecer una infraestructura de OV que permita explotar datos conocidos por toda la comunidad. Pero si alguien publicase datos nuevos sobre uno o más objetos, tendríamos que descubrirlo por nuestra cuenta y añadir manualmente esa publicación a nuestro programa de análisis.

Eso se evita con los registros de recursos del OV, de modo que un programa que use el OV pueda preguntar, por ejemplo, por datos en ultravioleta (UV) y recibir una lista de todos los servidores que almacenan datos UV y que se pueden consultar. Para mayor disponibilidad, existen múltiples registros que se comunican entre sí de modo que, en caso de mal funcionamiento de uno, se pueda utilizar cualquier otro de forma indistinta.

El OV en acción

Ilustraremos el uso del OV con *Aladin*, una herramienta creada por el Centro de Datos de Estrasburgo (CDS) como prototipo de herramienta OV, capaz de recopilar información de cualquier instrumento o base de datos que forme parte del OV.

En una sesión típica, indicaríamos el nombre de un objeto astronómico, y gracias a un protocolo de OV se obtendrían automáticamente sus coordenadas.

A partir de las coordenadas y un radio de búsqueda, se obtendrían imágenes del objeto tomadas con diferentes clases de luz -visible, infrarroja, o UV, por ejemplo-, e identificarían diferentes condiciones físicas. También podríamos comparar imágenes del mismo objeto, tomadas en diferentes épocas, para comprobar su evolución. Así podríamos ver, por ejemplo, la expansión de las capas

de una estrella tras una explosión de supernova.

A continuación, podríamos obtener las zonas de esa galaxia para las que existen datos espectrales, y superponer sus datos con otro programa, denominado *VOspec*, creado por la Agencia Espacial Europea (ESA). Los espectros nos proporcionan información sobre la abundancia de ciertos elementos, sus temperaturas, y otros parámetros físicos.

Aladin también permite crear programas capaces de identificar los puntos de mayor luminosidad coincidentes entre instrumentos, de calcular tamaños de objetos a partir de una imagen o muchas otras tareas que antes implicaban la inspección visual por parte de un astrónomo. *VOspec*, por su parte, puede calcular a partir de un espectro la temperatura a la que se encuentra una estrella, o la del disco de polvo que la rodea; y puede incluso actualizar el modelo teórico con el que realiza el cálculo.

El poder relacionar datos de posición en imágenes con los de espectros en una base de datos o un modelo teórico, y establecer la relación entre programas diferentes, es posible gracias a la estandarización establecida por el OV, tanto para los datos en sí (XML), como para las relaciones entre datos de diferentes tipos (modelos de datos), y la forma de acceso a datos de diferentes clases (protocolos de acceso). Esta estandarización es la que nos permite simplemente elegir la región del cielo que necesitamos estudiar, y apuntar nuestro OV como un instrumento más, que accederá a los datos disponibles de forma automática y siempre actualizada.