

Le partage de données et l'interopérabilité au sein de l'Observatoire Virtuel

Pierre Le Sidaner

DIO / Observatoire de Paris
61 Av. de l'Observatoire
75010 Paris

Albert Shih

DIO / Observatoire de Paris
61 Av. de l'Observatoire
75010 Paris

Résumé

Le partage de données et l'interopérabilité au sein de l'Observatoire Virtuel est possible grâce à la définition de standards d'échange, de formats et de métadonnées de description.

L'Observatoire Virtuel astronomique (OV) est un projet international démarré en 2001 ayant pour but de définir ces standards d'interopérabilité.

La définition de ces prérequis pour faire fonctionner un système interopérable dans lequel les utilisateurs peuvent accéder de manière transparente à toutes les données est propre à l'éco-système de la recherche. On peut néanmoins montrer une démarche qui est très semblable à celle que l'on trouve au sein du W3C ou de l'Open Geospatial Consortium (OGC).

Le succès rencontré par ce projet et les possibilités qu'il offre ont permis à quelques disciplines qui interagissent avec le monde de l'astronomie de construire à leur tour un système semblable.

Mots-clefs

interopérabilité, définition de standards, protocole d'échange

1 Introduction

L'Observatoire Virtuel astronomique (OV) est un projet international démarré en 2001 ayant pour but de définir les standards d'interopérabilité entre les données en astronomie.

L'exposé va s'axer sur la méthode employée pour fédérer une communauté, décrire les standards de protocoles et de formats, et construire des métadonnées pour que cette interopérabilité fonctionne. Nous présenterons le contexte technique indépendant de la discipline et décrirons le mode de fonctionnement, pour terminer sur une étude critique.

La construction du partage de données dans l'Observatoire Virtuel est basée sur l'adoption, par chaque fournisseur de données, de standards qui permettent d'acquérir les données mais surtout d'en avoir une description suffisante pour être en mesure de les *cross-corréler*, bien qu'elles proviennent de différentes sources (observatoires et archives) situées dans différents pays (France, USA, Japon, Brésil, etc.).

Je détaillerai aussi l'organisation fédératrice de l'OV baptisée IVOA (International Virtual Observatory Alliance) [1] et son mode de validation des standards. Le format des données et des métadonnées est basé sur XML et les protocoles d'accès sur HTTP. Un protocole spécifique permet aux applications de communiquer via des messages codés avec XML-RPC.

Enfin je présenterai ce qui fait le succès du projet et les résultats que l'on peut obtenir, en comparant notre approche avec les choix faits dans d'autres domaines comme la cartographie terrestre et le standard OGC [2].

2 Présentation de ce qu'est l'OV

L'Observatoire Virtuel est une collection d'archives de données qui utilisent l'Internet pour former un gros centre de données possédant des collections distribuées un peu partout au sein d'autres Observatoires. L'OV astronomique s'organise en projets nationaux regroupant les centres de recherche et de données. Ces initiatives ont créé une instance internationale, l'International Virtual Observatory Alliance (IWOA), qui fédère 18 pays et agences (ESO et ESA). Elle structure le développement de standards. Cela impose un modèle décentralisé dans lequel les acteurs se partagent les responsabilités.

Le but est de rendre l'ensemble du système d'accès aux données et aux services transparent pour l'utilisateur et qu'ultimement il ait l'impression que ce n'est pas plus complexe que de ne s'adresser qu'à une archive avec tous les outils de visualisation et de traitement simples associés.

La spécificité de l'OV est de promouvoir un système collaboratif dans lequel : chaque contributeur (centre de données) apparaît au même niveau, les standards sont définis de manière collégiale et dont l'unité fédératrice (l'IWOA) n'ait pas de fond propre ni de pays de rattachement. Les procédures de normalisation sont transparentes et accessibles à tous.

Ce projet lancé en 2001 a permis de définir des standards d'accès aux données en les élargissant à des données de plus en plus complexes. Il a commencé à décrire l'accès à des codes de simulation et des programmes de physique théorique. Beaucoup de procédures de normalisation ont été héritées du W3C. En France, l'Observatoire de Paris est, avec le CDS, l'un des gros contributeurs.

Le succès du projet a poussé certains membres à étendre ce concept à des disciplines voisines que sont la planétologie et la physique solaire ou encore la physique atomique et moléculaire.

3 Petit tour d'horizon sur le fonctionnement.

Tout d'abord le choix d'interopérabilité : le modèle de l'OV consiste à demander à chaque fournisseur de données de se conformer au standard et d'ajouter une couche d'interopérabilité au dessus de son archive. Cela implique d'imposer à une archive, soit d'installer un framework soit de développer cette couche à partir de bibliothèques existantes au dessus de sa base de données.

La structure passe facilement à l'échelle supérieure car elle ne demande pas de développement chaque fois qu'un nouveau fournisseur apparaît. Par contre la mise en place au départ est beaucoup plus lourde que celle du modèle simple constitué à partir d'un point central et d'API pour accéder aux différentes archives.

En ce qui concerne le niveau d'interopérabilité : l'OV a défini que l'utilisateur devait voir la structure comme une archive unique, c'est à dire qu'une machine soit capable de comparer des données provenant de plusieurs sources et de les afficher ou les *cross corrélér*. Pour que cela soit possible, il faut d'un part que les données soient décrites avec suffisamment de précision pour permettre la comparaison, et d'autre part que le format et la terminologie des métadonnées soient standardisés pour qu'un programme informatique puisse les comprendre.

Pour arriver à un tel degré d'interopérabilité il faut alors définir un certain nombre de standards :

- un ou plusieurs protocole d'accès, qui s'appuient sur HTTP REST (GET/POST) avec des possibilités synchrones ou asynchrones (pour les requêtes complexes et longues). L'important est de définir le modèle et la grammaire, c'est-à-dire la façon d'interroger et d'interagir ainsi que les paramètres physiques de la requête. Les protocoles OV construisent leurs requêtes en prenant par défaut toutes les données (l'équivalent d'un *select **), puis des filtres sont associés à la requête pour limiter la quantité de données retournées par le service. Selon le type de données, ces filtres peuvent avoir des formes différentes : se limiter à une zone du ciel, à une bande de fréquence, une résolution, etc. Les protocoles d'accès aux données agissent souvent en deux temps car les volumes de données sont conséquents et l'utilisateur ne souhaite que rarement tout charger. Dans un premier temps l'utilisateur fait sa requête et ne reçoit que les métadonnées associées à des url de téléchargement des données pour informer l'utilisateur sur le contenu et la taille des données proposées. Les métadonnées qui constituent la description du jeu de données doivent permettre à l'utilisateur de se faire une première opinion sur ces données donc de faire le choix de les télécharger ou pas. Des informations comme la résolution spatiale ou spectrale, le champ de vue, l'instrument d'acquisition, sont primordiales.

- un format d'échange standardisé : lorsqu'on interroge un service ou que l'on souhaite recevoir une faible quantité de données. Le format choisi par le VO est la VOTable [3]. C'est un fichier XML au sein duquel la verbosité a été réduite (par rapport au XML classique).

- des métadonnées standardisées. Les métadonnées sont la clef de l'interopérabilité, C'est grâce à ces informations qu'on va pouvoir comparer automatiquement deux jeux de données. Il faut qu'une machine puisse reconnaître que deux colonnes provenant de deux jeux de données, représentent les mêmes grandeurs physiques. Il est donc primordial que chaque fournisseur de données parle le même langage. Des métadonnées sont présentes : lors des requêtes comme on l'a vu précédemment, mais aussi dans les données pour caractériser les colonnes d'une table, les spectres, les images ou les cubes de données. La sémantique de ces métadonnées a subi une évolution majeure. Initialement chaque grandeur était décrite par un mot. Pour limiter le vocabulaire, la deuxième phase a été la combinaison de mots séparés par des points ou des point-virgules qui compose une grammaire que l'on nomme Unified Content Descriptors (UCD). Chaque UCD est construit par une suite de mots décrivant l'élément du plus large au plus précis. Par exemple « em.line.Lyalpha » représente une mesure de raie d'hydrogène lyman alpha. Construit à partir d'une souche « em » mesure du spectre électromagnétique, combiné en « em.line », désignation d'une raie atomique principale, puis Lyman Alpha, raie caractéristique de l'hydrogène très importante en astronomie. Ces UCDs sont primordiaux dans la *cross corrélation* de tables pour comparer des colonnes.

- un annuaire centralisé ou « registry » qui répertorie l'intégralité des services conformes aux standards du VO. Les services sont enregistrés avec l'url d'accès, la description des collections, du fournisseur, des contacts, etc. Les annuaires sont répliqués en utilisant le standard d'échange des bibliothèques (OAI-PMH). La technologie d'interrogation (web services SOAP combiné avec un langage de requête propre ADQL 1.0) est en train de migrer vers un standard plus simple (REST et Table Access Protocol).

Les outils de visualisation : la standardisation des accès permettant de rapatrier simplement les données, ces étapes sont maintenant intégrées aux outils de visualisation courants. Le développement et la maintenance d'un client de visualisation de spectre, d'image, de catalogue est très coûteux en temps humain. Ces outils sont le fruit de la compétence d'une personne ou d'une équipe qui se spécialise dans un type de données (image ou spectre ou catalogues et cubes). L'utilisateur souhaite pouvoir utiliser l'ensemble de ces briques comme faisant partie d'une boîte à outils d'où l'idée de permettre à chacun de ces programmes de dialoguer (échanger des données) avec les autres. Si un visualisateur d'image récupère un jeu de données contenant des spectres et des tables de données il sera capable à la demande de l'utilisateur de les transférer automatiquement à un client de spectre ou gestionnaire de tables de données. L'étape suivante a été de permettre aux pages web d'interagir de la même façon avec les clients de visualisation. Le protocole de dialogue inter-applications mis en place est composé d'un langage extrêmement simple. Chaque application s'enregistre sur un hub et peut communiquer avec quelques mots de grammaire ou envoyer des données reçues par tout ceux qui sont connectés à ce hub. C'est le Simple Application Messaging Protocol (SAMP) [4].

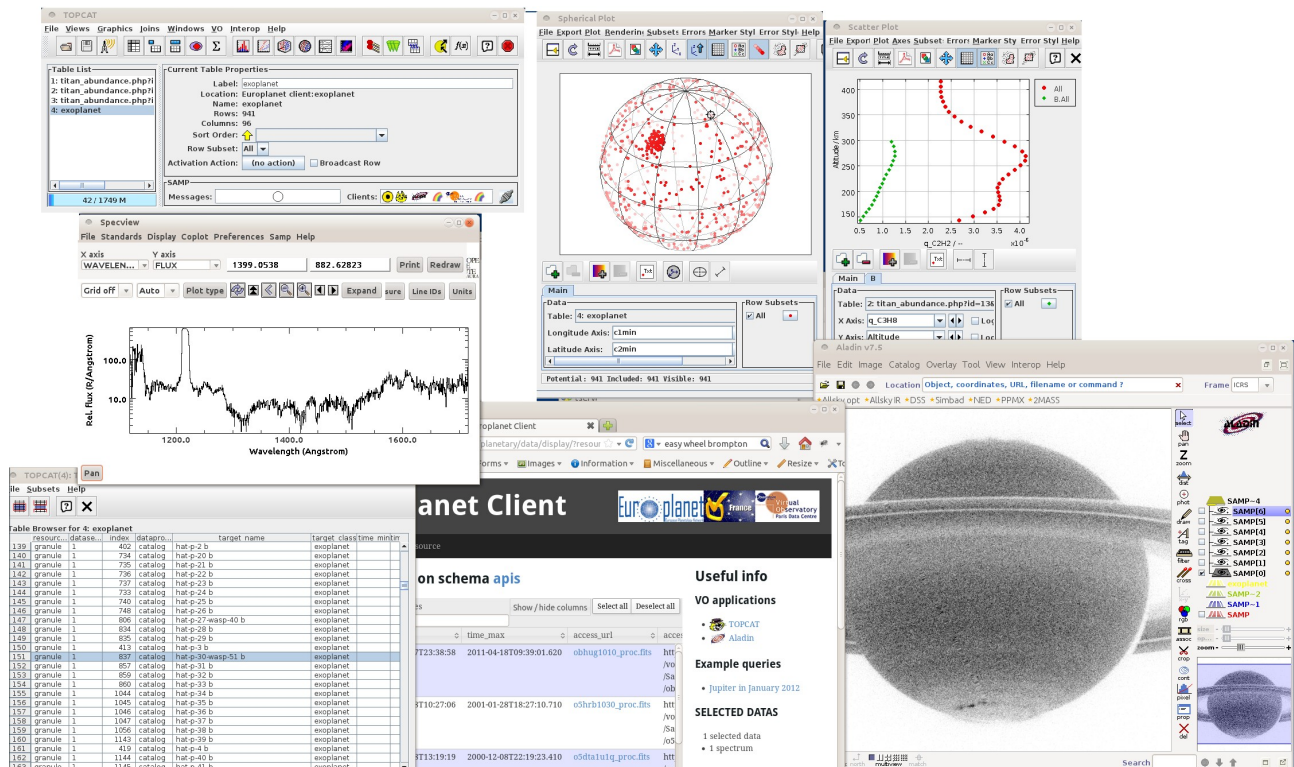


Figure 1 - Quelques clients de visualisation de l'OV SAMP compatibles

4 La chaîne de standardisation

L'IVOA est une organisation ouverte à tous et en charge de définir des standards reconnus. Chacun est donc libre d'y proposer un standard. Afin de conserver une cohérence d'ensemble et que chacune de ces propositions puisse être discutées et prise en compte également. Il a été nécessaire de définir une chaîne de validation des propositions. C'est ce processus long mais indispensable qui est décrit ci-dessous.

L'IVOA regroupe 18 pays comprenant chacun plusieurs observatoires, compte tenu du nombre et de la diversité des participants il est donc important de créer une structure coordinatrice qui définit les règles et les procédures. Pour être efficace l'IVOA est organisé en groupes de travail thématiques, les moyens d'échange étant essentiellement un wiki et des listes de diffusion. Chaque personne qui souhaite participer à un groupe ou plusieurs est la bienvenue, les thèmes clefs sont : data model, data access layer, grid & web services, registry et semantics. Ces groupes ont pour but de rassembler les compétences et le bonnes volontés pour répondre aux besoins des utilisateurs et faire avancer les standards. Si les moyens de communications électroniques sont prépondérants il est indispensable de se réunir (en présentiel) régulièrement pour créer des liens plus humains et trancher les points de discussions bloqués. Ces colloques ont lieu deux fois par an (meeting Interop).

Pour conserver une cohérence d'ensemble entre les travaux des différents groupes, l'IVOA se dote d'un Technical Coordination Group (TCG), et pour gérer l'organisation globale et la prospective il y existe un groupe Exécutif et un « chairman ». L'ensemble des personnes qui composent ces structures ont des mandats à durée déterminée non renouvelables sauf cas très exceptionnel.

Les standards suivent une procédure de labellisation copiée sur le modèle du W3C avec des phases de « draft, proposed recommendation, recommendation » suivant la figure 2

IVOA Document Standards Process

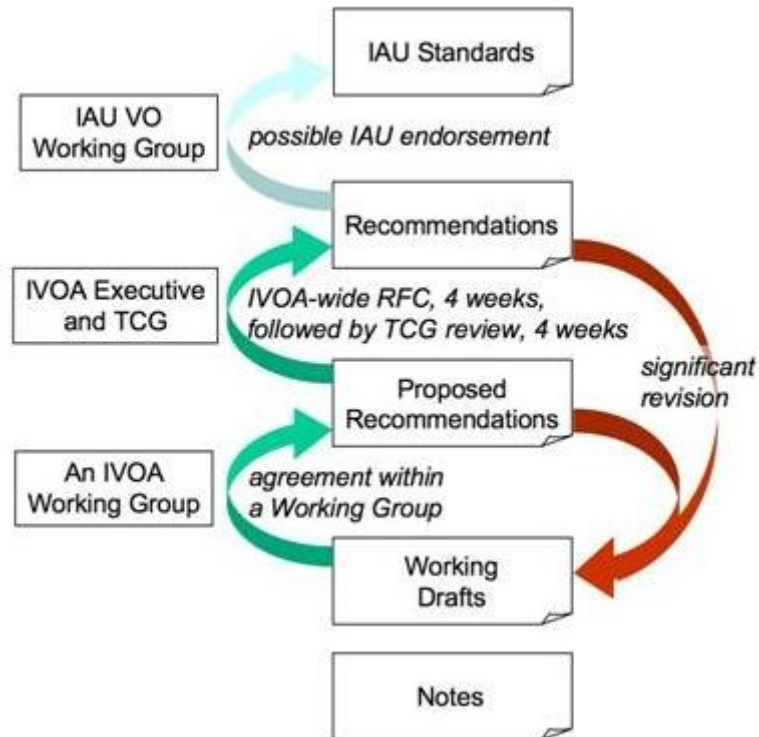


Figure 2 - Chaîne de validation des standards de l'IVOA

Le processus de validation commence par une proposition postée sur la liste de diffusion d'un des groupes de travail. Ce document prend le statut de « working draft ». Pour aller plus loin il faut deux implémentations indépendantes de la proposition de standard ainsi qu'une librairie si besoin et d'un validateur si ça s'applique.

Si le groupe de travail s'est mis d'accord sur le document, celui-ci devra être publié sur le wiki avec une période de commentaire « request for comment » d'au moins deux semaines. Ensuite le responsable fait escalader le processus et passe le statut en « Proposed recommendation ».

Les responsables et co-responsables des groupes de travail doivent afficher leurs commentaires durant la période de revue de 4 semaines. Ensuite l'exécutif entérine le changement de statut vers recommandation.

A chaque étape il peut y avoir des prolongements, des blocages voire des demandes de modification suffisamment importantes pour que le projet modifié refasse une partie du circuit de validation. Finalement le standard est adopté par l'International Astronomical Union (IAU) qui est le groupement de la communauté qui entérine beaucoup de standards, notamment le nom des étoiles et des objets du système solaire.

En conclusion, le modèle de standardisation est ouvert et accessible à tous ce qui rend les standards bien adaptés aux besoins de la communauté. Ce modèle est collégial, on assiste parfois à des luttes d'influence pour que le processus mis en place et développé par une équipe ou un projet soit préféré à celui d'un autre. Notons que l'utilisation de la langue anglaise dans tous les échanges et les documentations favorise nos voisins anglais et américains.

5 Le succès du système

L'OV a commencé par donner accès à des données de base d'astronomie (images, spectre, catalogue) avec des protocoles simples. Même si ces protocoles ont évolué au cours des versions, ils constituent une façon simple de publier des données dans l'OV. Mais comme les scientifiques ont toujours besoin de plus de précision pour faire de la fouille de données, les protocoles d'accès se complexifient au cours du temps. De plus on souhaite étendre le type de données mis

à disposition, en ajoutant de la photométrie, des cubes de spectres ou d'images, des codes en ligne, des données issues de simulations et de la physique théorique, etc.

Avec l'arrivée des télescopes robotisés type LSST capables de détecter plusieurs millions d'événements par nuit, l'OV est à la recherche d'un moyen d'envoyer un flux d'alerte semblable à un flux RSS.

Tous ces nouveaux défis font évoluer l'écosystème de l'OV, complexifient ses standards et le rendent aussi plus riche et ouvert à de nouvelles extensions.

L'OV regroupe toujours plus de services et un certain nombre de gros centres de données utilise ce modèle de diffusion comme référence d'accès à leur archive. Le nombre croissant de possibilités offertes induit aussi des besoins de consolidation d'infrastructure informatique notamment pour la fouille de données qui attaque les bases de données par des champs non indexés.

L'OV est devenu incontournable pour les gros centres de données astronomiques et il est maintenant intégré comme protocole de base d'accès à l'archive. Il devient la façon naturelle de promouvoir les données et de permettre la fouille de données, en supplément l'utilisation de SAMP permet aussi aux portails web classiques d'interagir avec l'ensemble des outils OV. Publier ses données dans l'OV c'est les rendre plus visibles, plus accessibles donc donner du crédit à l'équipe qui les a acquises.

La communauté scientifique qui utilise l'OV dans ses recherches ne croît pas aussi vite qu'on le voudrait car le prérequis pour construire des pipelines de fouille de données évolués n'est pas simple. Les utilisateurs de l'OV sont souvent des scientifiques avec une forte compétence en développement informatique. Si l'on parle de succès c'est que le volume de données augmente continuellement, que les nouveaux projets pensent d'entrée à intégrer la publication dans l'OV de leurs données. Enfin l'OV commence à être utilisé nativement par quelques équipes scientifiques pour mettre leurs données à disposition en interne, même si ces données ne sont pas publiques, tout simplement parce les outils pour comparer ces données avec celles existantes sont disponibles et éprouvés, ainsi que les outils de visualisation comme Topcat [5], Aladin [6], vospec [7], specview [8].

6 L'émulation de l'OV astronomique

L'envie de profiter du travail fait par la communauté d'astronomes pour avoir des accès simplifiés aux données a touché d'autres communautés que j'appellerais voisines au sens où elles interagissent avec l'astronomie. L'astronomie est le seul domaine de la physique réunissant autant de disciplines : mécanique, relativité, mécanique des fluides, électromagnétisme, physique atomique et moléculaire, etc.

Parmi les projets européens FP7 qui se sont achevés en 2012, plusieurs d'entre eux avaient pour objectif de construire un OV dans ces disciplines en utilisant les expériences de l'IVOA. On peut citer notamment la physique atomique et moléculaire qui a repris les techniques de l'OV pour développer un réseau similaire appelé VAMDC « Virtual Atomic and Molecular Data Centre »[9]. Un grand nombre de partenaires du projet provenaient de l'IVOA et ils ont recyclé les développements fait dans ce cadre. Il a fallu néanmoins utiliser et étendre un modèle de données qui provenait de la communauté de physique atomique et moléculaire et adapter les protocoles et les modes d'accès.

La planétologie avec le programme Europlanet[10] propose d'utiliser l'infrastructure IVOA tout en ayant développé des métadonnées, des formats de fichiers et un modèle de données spécifiques à cette communauté.

7 Conclusion

La spécificité des données d'astronomie et des sciences de l'univers est que même si elles sont très coûteuses à acquérir, elles n'ont pas de valeur marchande. Les acteurs de l'OV sont donc uniquement des observatoires ou des instituts de recherche, il n'y a pas de partenaires industriels. Qu'en est-il des autres modèles structurés ? Un exemple un peu différent d'acteurs qui utilisent un certain nombre de données similaires est l'« Open Geospatial Consortium » (OGC). Les besoins de standardisations géographiques ont été primordiaux pour l'armée, les industriels, les organismes de recherche mais aussi les individus. Pour poursuivre l'initiative, l'OGC, qui a utilisé le travail fait par le « Geographic information system » (GIS), a adopté un modèle standard ouvert et associatif, même si des logiciels clients ou serveurs peuvent être sous licence privée.

Dans le cas de l'IVOA il s'agit bien d'un modèle associatif sans fond propre et même sans financement par les contributeurs. Les standards sont ouverts, les partenaires sont des instituts de recherche.

La quantité de données en astronomie est colossale tant par les missions spatiales que par les acquisitions depuis le sol. L'arrivée des très grands radio-télescopes, des télescope robotisés ou des données à très hautes énergies multiplie la taille des archives et leur nombre. L'utilisateur ne peut pas connaître l'existence de chaque jeu de données ni le mode d'accès propre à chaque fournisseur. L'émergence de l'OV permet d'identifier, de rendre accessible et comparables entre elles les données provenant de différentes sources. L'OV ne fait pas le travail à la place des scientifiques, mais il leur fait gagner un temps considérable et il assure que la description minimale nécessaire à l'exploitation des données est présente. Si les standards sont loin d'être parfaits, le modèle d'interopérabilité et l'adoption de ces standards restent une référence enviée par d'autres disciplines.

Bibliographie

- [1] <http://www.ivoa.net>.
- [2] <http://www.opengeospatial.org/>
- [3] <http://www.ivoa.net/documents/VOTable/>
- [4] <http://www.ivoa.net/documents/SAMP/>
- [5] <http://www.star.bris.ac.uk/~mbt/topcat/>
- [6] <http://aladin.u-strasbg.fr/>
- [7] <http://www.sciops.esa.int/index.php?project=SAT&page=vospec>
- [8] http://www.stsci.edu/institute/software_hardware/specview
- [9] <http://www.vamdc.eu>
- [10] <http://voparis-europlanet.obspm.fr/>