

SITools2, un système d'accès aux données scientifiques web 2.0

Jean-Christophe Malapert
CNES
18 Av. Edouard Belin 31400 Toulouse Cedex 9

Hervé Ballans
IAS
Centre universitaire d'Orsay, Bât 120 - 121, 91405 ORSAY CEDEX

Chrystel Moreau
LAM
Pôle de l'Étoile Site de Château-Gombert - 38, rue Frédéric Joliot-Curie 13388 Marseille cedex 13

Résumé

Les communautés scientifiques d'astronomie (et d'autres disciplines) sont tenues de mettre en accès public leurs données expérimentales. Ceci a conduit à redévelopper un grand nombre de fonctionnalités communes ainsi que des services spécifiques. Notre objectif pour SITools2 est d'offrir ces mêmes services à travers une plateforme web "open source" conviviale. L'architecture de cette plateforme permet d'une part le développement et l'intégration rapide de nouveaux services par un mécanisme de plug-ins, et d'autre part le partage des bénéfices apportés par ces plug-ins entre les membres de la communauté. Le développement du projet est conduit par une méthode Agile depuis août 2010, dont on décrira les avantages et les contraintes. Aujourd'hui, le Laboratoire d'Astrophysique de Marseille et l'Institut d'Astrophysique Spatiale ont choisi d'utiliser SITools2 pour mettre à disposition de la communauté scientifique internationale les données spatiales hébergées dans leurs centres de données (respectivement CeSAM et IDOC), ce qui donne une idée de la maturité de la plateforme et de ses capacités d'évolution. Quelques exemples d'interfaces opérationnelles et de développements internes seront présentés.

Mots clefs

archivage, sitools2, système d'information, accès, données, web 2.0, REST, RIA

1 Introduction

L'archivage vise à collecter et à conserver l'ensemble des informations permettant de décrire la donnée à préserver dans le but de la restituer dans une forme compréhensible pour une communauté donnée. Dans le modèle OAIS (Open Archival Information System), une archive est décomposée en six entités fonctionnelles : l'insertion, le stockage, la gestion des données, l'administration, la planification de la préservation et l'accès. Actuellement, il existe un ensemble d'archives OAIS avec des implémentations diverses et variées. Chaque nouvelle implémentation conduit alors à redévelopper des services déjà existants dans d'autres archives.

Suite à ce constat, le CNES décide, en 2002, de créer le projet SITools afin de fédérer les laboratoires scientifiques à travers un outil d'accès aux données. Après neuf ans d'utilisation, les technologies et les besoins des archives ont évolués.

Basé sur l'identification de ces nouveaux besoins et sur le retour d'expérience des utilisateurs, le CNES décide en accord avec ses partenaires d'implémenter une nouvelle génération de l'outil, appelé SITools2. Ce projet est en cours d'implémentation depuis août 2010. Cette nouvelle version, open source, implémente la couche d'accès aux données ainsi que sa fonction d'administration. Elle permet également de s'affranchir des limites techniques rencontrées dans la première génération de l'outil et d'utiliser des standards reconnus pour faciliter l'interopérabilité des jeux de données. De plus, afin de pouvoir s'adapter aux nouvelles exigences rencontrées pendant la phase de réalisation et d'affiner les besoins des archives scientifiques, le développement de SITools2 est conduit depuis sa création par une méthode Agile. Aujourd'hui, le Laboratoire d'Astrophysique de Marseille et l'Institut d'Astrophysique Spatiale ont choisi d'utiliser SITools2 pour mettre à disposition de la communauté scientifique internationale les données spatiales hébergées dans leurs centres de données respectifs CeSAM et IDOC. Ces deux centres de données accompagnent le projet SITools2 depuis sa création.

Dans un premier temps, nous présenterons l'architecture de SITools2. Ensuite, quelques exemples d'interfaces opérationnelles et de développements internes seront présentés.

2 L'architecture

Durant les années 2009 et 2010, deux études de veilles technologiques ont été menées par le CNES : l'une concernant l'architecture REST (Representational State Transfer) et l'autre en rapport aux technologies RIA (Rich Internet Application).

L'objectif de la première étude était d'étudier le remplacement de la technologie SOAP qui n'était pas assez « scalable » pour SITools et qui souffrait de problèmes de performances liés au transfert des gros volumes de données entre le client et le serveur. Un autre objectif était de choisir la technologie la plus simple à utiliser afin de favoriser le développement de services spécifiques par les laboratoires scientifiques. L'objectif de la seconde étude était d'étudier l'apport des RIA dans nos systèmes d'informations. L'étude s'est focalisée, en particulier, sur la richesse des composants graphiques et la personnalisation de ceux-ci.

Suite à ces études, l'architecture REST et les RIA respectivement implémentée par les bibliothèques Restlet[1] et Ext-JS[2], ont été retenues pour l'implémentation de SITools2.

2.1 La plateforme

A partir des technologies REST et RIA, nous avons donc créé une plateforme web 2.0 basée sur une architecture client/serveur, capable de s'intégrer à une architecture existante dans un laboratoire, de s'adapter à un design de base de données existante (évitant ainsi la migration des données) et comportant déjà un certain nombre d'applications. Cette architecture nous permet aussi de découpler les parties clientes de la partie serveur. Par exemple, des développements, non décrits dans ce papier, sont en cours afin de connecter certains clients de valorisation de données pour l'Observation de la Terre à SITools2. La figure 1 représente une vue de haut niveau de SITools2 et de ses principales applications. Le client est composé de trois interfaces web : une interface d'administration en charge de configurer le système, un portail présentant l'ensemble des projets disponibles à travers une instance SITools2 et une interface projet partiellement paramétrable par chaque utilisateur. Le serveur est composé d'un ensemble d'applications regroupées en trois catégories (système, administration et

utilisateur) au sein d'une seule application « Registry » : A partir de ce « Registry », il est possible d'activer et de désactiver les services ainsi que de configurer les droits d'accès pour chaque service en fonction d'utilisateurs et/ou groupes d'utilisateurs.

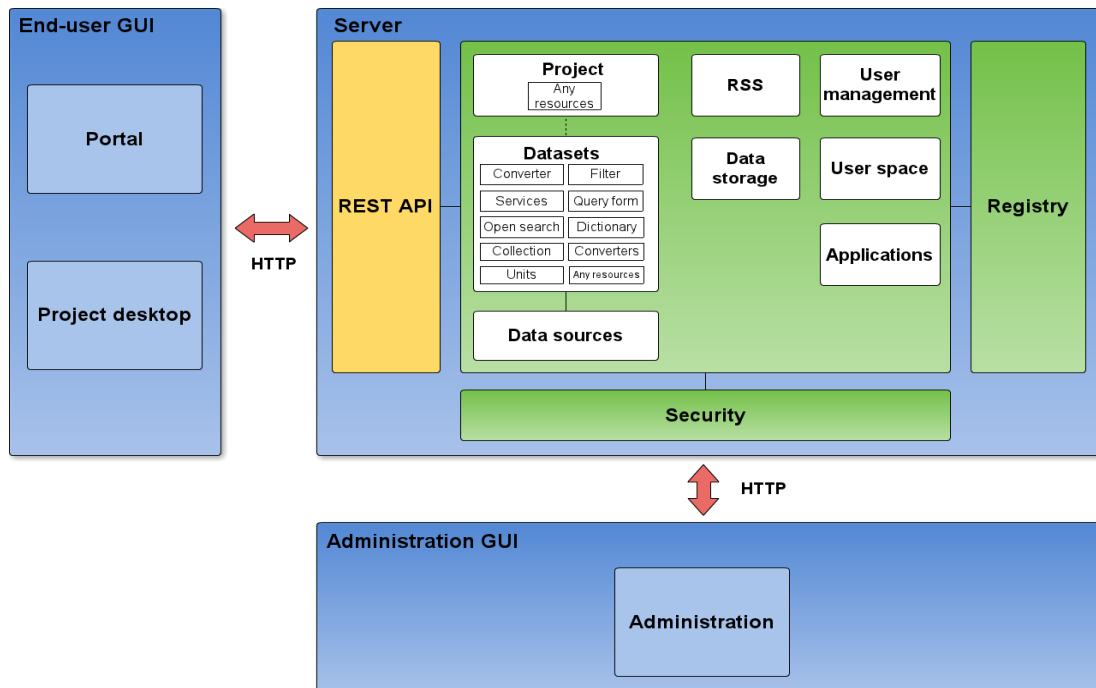


Figure 1: La plateforme SItools2

2.1.1 L'organisation du serveur par « couches »

La figure 2 décrit l'organisation du serveur en forme de « couches ».

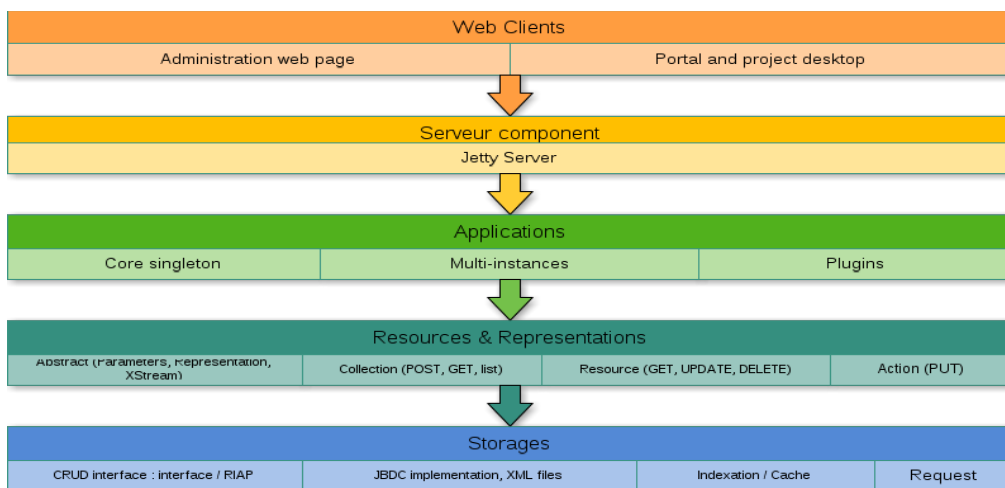


Figure 2: Vue de SItools2 en couches logicielles

- La couche « web clients »

Cette couche comprend plusieurs applications clientes, dont le tableau de bord d'administration, le portail et le bureau projet. Ces applications sont construites avec le Framework JavaScript Ext-JS. Il s'agit d'applications à part entière, dont l'architecture repose sur

le modèle MVC. Le modèle est représenté par des classes Store Ext-JS. La présentation est constituée par l'assemblage des composants graphiques « Form », « Grid », « Panel », ... Le rôle du contrôleur est joué par les classes et méthodes chargées du traitement des événements utilisateur, et d'échange avec le serveur.

- La couche « Serveur composants »

Dans la version initiale, SITools2 est composé d'un seul serveur intégrant l'ensemble des services d'administration, et d'utilisation des jeux de données. Dans une version ultérieure, nous aurons pour objectif de répliquer différents serveurs d'exposition de données pour assurer une meilleure montée en charge.

- La couche « Applications »

Cette couche rassemble des composants fonctionnels de haut niveau correspondant aux fonctionnalités majeures du système. Certaines applications indispensables au fonctionnement font partie du cœur du système. D'autres applications « optionnelles » peuvent se greffer par un mécanisme de plug-ins. Nous considérons aussi que l'exposition d'un jeu de données élémentaire constitue une application à part entière sur laquelle on peut appliquer la même stratégie de sécurité. Les applications sont des objets non volatiles, qui possèdent les informations nécessaires pour le traitement des requêtes par des ressources. Une application pourra ainsi posséder un modèle de données en cache, et la référence des objets des couches inférieures de persistance.

- La couche « Resource & representation »

Cette couche rassemble les classes chargées du traitement de chaque requête HTTP. Il s'agit plus précisément, de décoder la représentation fournie dans le corps de la requête HTTP, d'invoquer les traitements métiers adéquats et de restituer dans le corps de la réponse HTTP la meilleure représentation. Cette couche est fortement liée à la couche « Applications ». Elle y fait d'ailleurs appel pour obtenir son contexte et la référence vers les interfaces de la couche de persistance des objets qui l'intéressent.

- La couche « Storage »

Elle est chargée de restituer ou de modifier les objets de l'application. Chaque application possède sa propre interface de stockage. On pourra ainsi assembler le système en choisissant une implémentation de chaque interface. Par exemple, la gestion des inscriptions utilisateurs et la gestion des utilisateurs et des groupes reposent sur des stockages différents : le fichier XML pour les inscriptions, la base de données SQL pour les utilisateurs et les groupes.

2.1.2 Dépendance entre les ressources et les applications

Certaines ressources et applications de la plateforme peuvent être dépendantes les unes des autres. Afin de pouvoir synchroniser les dépendances entre celles-ci, un système de notification d'évènements a été mis en place dans SITools2. Cette caractéristique permet donc de configurer l'ensemble du système sans le redémarrer.

2.2 Le Framework de développement coté serveur

La plateforme SITools2 comporte les fonctionnalités communes d'un système d'accès aux données. Cependant, il est important de pouvoir aussi ajouter des applications spécifiques afin de s'adapter aux différents projets. C'est pourquoi nous avons donc mis en place des points d'extensions au niveau de la plateforme afin de pouvoir héberger des plug-ins permettant d'enrichir les fonctionnalités de celle-ci.

Afin de faciliter le développement de SITools2 et le déploiement de nouvelles applications par les développeurs, nous avons concentré l'ensemble des applications communes et l'ensemble des classes et des routines que les développeurs peuvent appeler ou étendre à partir de l'API dans un unique JAR. Pour ajouter de nouvelles fonctionnalités à la plateforme, les développeurs doivent donc créer un ou plusieurs JAR « extension » contenant les classes d'implémentation désirées de chaque plugin.

Ensuite, ce JAR est copié dans un répertoire particulier qui sera analysé au démarrage du serveur de SITools2. Le serveur ajoutera les différentes classes se trouvant dans ce répertoire au CLASSPATH de SITools2. Ces plug-ins sont alors découverts dans l'interface d'administration et prêts à être configurés pour utilisation.

La figure 3 représente les points d'extensions du serveur de la plateforme de SITools2.

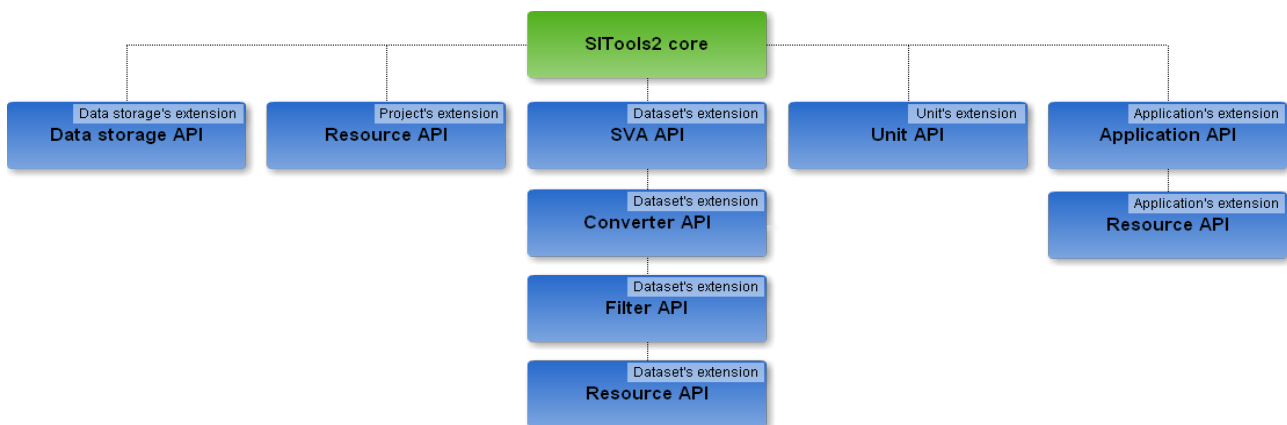


Figure 3: Points d'extension du serveur de la plateforme de SITools2

Une courte description de chaque point d'extension est présentée ci-dessous :

2.2.1 Les filtres et les convertisseurs

Les filtres et les convertisseurs sont des classes JAVA liées à la notion du jeu de données qui peuvent être exécutées les unes à la suite des autres. Ces classes sont paramétrables via l'interface d'administration et échangeables entre les instances SITools2.

Le filtre est un élément permettant de transformer l'information envoyée par les clients en une information compréhensible pour la source de données avec laquelle il interagit. Par exemple, un filtre peut effectuer une recherche géométrique sur un jeu de données.

Les convertisseurs permettent de transformer une information provenant d'une source de données en une autre information qui est accessible par les clients.

Ces deux mécanismes permettent à SITools2 de s'adapter à des bases hétérogènes sans les modifier

2.2.2 Les services à valeur ajoutée (SVA)

Les SVA sont des classes JAVA liées à la notion du jeu de données et paramétrables via l'interface d'administration. Ces classes peuvent utiliser aussi bien le mode synchrone ou asynchrone. Par exemple, les SVA peuvent permettre de commander des données vers un espace utilisateur, d'exporter des données dans un format particulier ou bien d'implémenter des services d'interopérabilité.

2.2.3 Le stockage

Une fois que les données sont sélectionnées par un utilisateur, elles peuvent être téléchargées par celui-ci. Dans le cas où les données ne sont pas publiques et que les données ont différents droits d'accès dans un même répertoire, l'autorisation de l'accès à un fichier doit être réalisée par un processus métier. Ce processus est intégrable dans SITools2 par l'intermédiaire d'une classe Java et permet donc une délégation de l'autorisation de l'accès à un fichier par une classe métier.

2.2.4 La conversion d'unité

Un processus de conversion automatique d'unité est actuellement implémenté dans SITools2. Ce processus est réalisé à partir de l'analyse dimensionnelle des unités incluses dans une dimension physique. Il est actuellement possible d'ajouter de nouvelles conversions d'unités qui n'ont pas la même dimension physique : par exemple, la conversion de la fréquence en longueur d'onde par l'utilisation de la constante de la vitesse de la lumière.

2.2.5 La ressource

La ressource est une extension qui permet d'ajouter une ressource REST à une application, à un jeu de données ou à un projet. Cette ressource peut être administrable par l'interface d'administration.

2.2.6 L'application

Des applications sont aussi ajoutables et paramétrables par configuration. Ces applications sont déconnectées de toute notion métier et permettent donc au développeur d'implémenter ce qu'il souhaite.

2.3 Le Framework de développement coté client

Du côté client, le Framework écrit en JavaScript est composé d'un noyau et d'extensions appelées « module d'application ». Chaque module peut être enlevé ou ajouté grâce au gestionnaire de modules. En plus de ces modules, il est possible de déposer de nouveaux composants graphiques et de nouvelles vues qui peuvent être utilisés par l'administrateur pour la configuration des formulaires de recherche ou pour le choix de la présentation des résultats.

3 Agilité

Un des objectifs du CNES était de faire participer les laboratoires scientifiques tout au long du développement de SITools2. Cette participation intervenait sur plusieurs points :

- le retour utilisateur (ergonomie, implémentation de telle ou telle fonctionnalité, test)
- la spécification de certaines user stories
- la priorisation des user stories après accord avec le « product owner »

Après un an de développement, le bilan est très positif grâce à une forte implication des différents acteurs (laboratoires scientifiques, « product owner », « scrum master » et l'équipe de développement). Parmi les points positifs, on peut citer un meilleur dialogue avec les différents acteurs du projet, des échanges d'idées, une intégration continue, une dynamique au niveau du développement, une meilleure transparence et un outil qui répond aux besoins. Parmi les points négatifs, on peut noter que la dynamique génère parfois du stress car l'équipe

de développement travaille en flux tendu, la documentation ne suit pas toujours les développements (dilemme dans la priorisation entre le fonctionnel, le test et la documentation) et le temps que le « product owner » et les laboratoires scientifiques doivent consacrer au projet est loin d'être négligeable.

4 Mise en production dans les centres de données

4.1 CeSAM

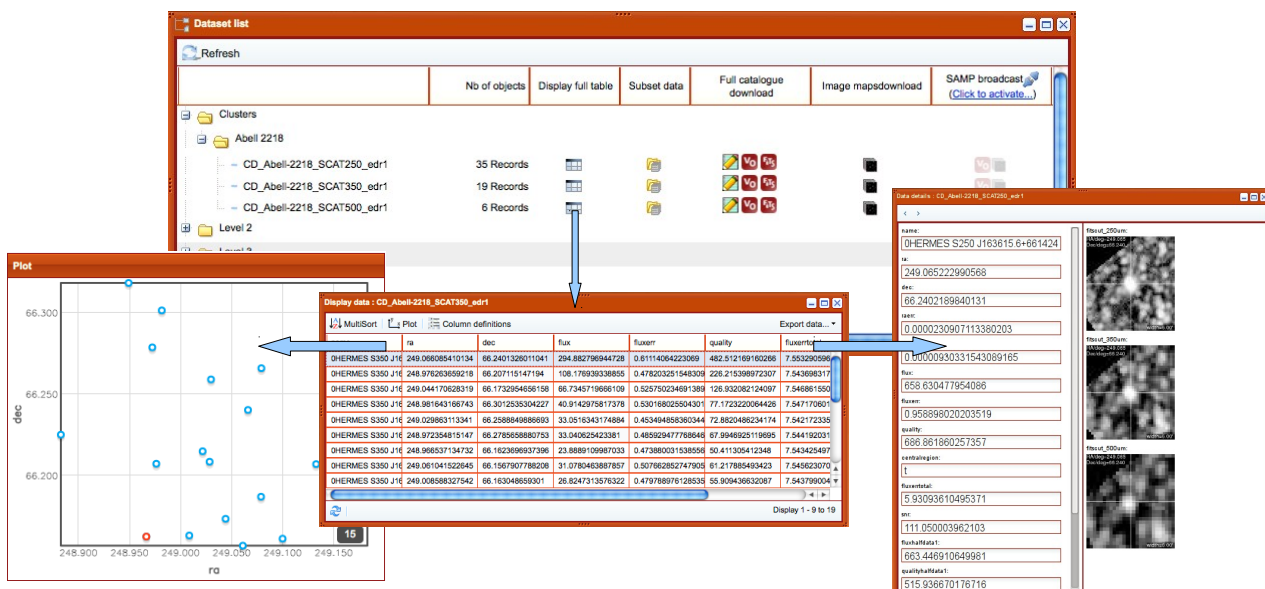
4.1.1 Présentation du CeSAM

Le Laboratoire d'Astrophysique de Marseille (LAM) a regroupé toutes ses activités liées à l'Informatique Scientifique au sein du Centre de données Astrophysique de Marseille (CeSAM), qui a la responsabilité de systèmes d'information pour mettre à disposition les données à valeur ajoutée des différents projets européens du CNES, de l'ESO, de l'ESA et des projets internationaux dans lequel le LAM s'implique. Le CeSAM a participé depuis l'origine à l'élaboration et au déploiement des systèmes SiTools avec le CNES (développement d'API, de modules d'interrogation, spécifications des besoins,...). SiTools2 permettant d'optimiser les investissements de développement et de déploiement pour les SI scientifiques, s'impose comme le middleware privilégié.

4.1.2 Les Fonctionnalités des systèmes d'information via SiTools2

Les systèmes d'information du CeSAM offrent différentes fonctionnalités au travers SiTools2 pour permettre :

- un téléchargement direct des données (catalogues et images) au format Votable et Fits ;
- une recherche des données par multi-critère, par recherche autour d'une position ;
- une visualisation des données (catalogues, images, spectres) ;
- une extraction des données au format Votable, Fits et Csv.



4.2 IDOC

4.2.1 Présentation de IDOC

Les données des missions spatiales en astronomie auxquelles participe l'IAS sont mises à la disposition des communautés nationale et internationale. Cette mise à disposition est faite par l'intermédiaire de son centre de données IDOC dont la fonction est de gérer, préserver, valoriser et promouvoir les informations hébergées, en offrant notamment un accès aisé à l'ensemble des jeux de données disponibles.

L'ensemble des archives de l'IAS est constitué de données hétérogènes provenant de missions spatiales d'observation et représente un volume de stockage global de plus de 200 To et plusieurs bases de données de métadonnées de gros volumes.

Parmi ces jeux de données : TRACE, SOHO, STEREO et SDO pour les missions d'observation du soleil ; Planck et Herschel pour les missions cosmologiques, CoRoT pour l'étude des étoiles lointaines et la recherche d'exoplanètes ainsi que l'instrument OMEGA de Mars-Express pour l'étude des sols martiens.

4.2.2 SITools2, outil générique et modulable

L'IAS avait contribué à la définition des objectifs de la première génération de l'outil et utilisait celui-ci depuis plusieurs années.

Le choix de confirmer SITools2 comme plateforme stratégique d'IDOC s'appuie sur les fonctionnalités étendues de la nouvelle version. Elle répond aux nouveaux besoins et permet également de s'affranchir des limites techniques rencontrées dans la première génération de l'outil. En outre, l'utilisation de standards reconnus facilite l'interopérabilité des jeux de données avec d'autres catalogues astronomiques externes (notamment avec la possibilité d'intégrer nativement les services web).

L'hétérogénéité des jeux de données implique un cahier des charges différent pour chaque projet de mise à disposition et ce à trois niveaux :

- la nature des données (métadonnées, données d'observation avec différents niveaux de correction, images, spectres, courbes de lumières,...) ;
- le type de traitement en ligne qu'on souhaite appliquer sur ses données. Outre les outils disponibles de base (export CVS, téléchargement d'une archive zip,...), SITools2 permet le développement de SVA pour l'ajout de fonctionnalités sans modifier la partie serveur ;
- l'aspect cosmétique de l'interface cliente (image de fond d'écran, nombre d'icônes, menu personnalisé, formulaire de recherche personnalisé,...).

La modularité de SITools2 permet de répondre à ces différents aspects et de mettre en place une interface d'accès ergonomique et personnalisée à chaque jeu de données.

5 La forge

SITools2 est disponible sous sourceforge <http://sitools2.sourceforge.net>. Deux « mailing lists » sont disponibles : une liste pour les développeurs (sitools2-developers@lists.sourceforge.net) et une liste pour les utilisateurs (sitools2-users@lists.sourceforge.net).

6 Conclusion

SITools2 est une plateforme web générique qui possède par défaut un grand nombre de fonctionnalités administrables. Toutes ces fonctionnalités permettent une utilisation immédiate

de la plateforme. Cependant, afin de s'adapter à chaque projet, des fonctionnalités supplémentaires peuvent être ajoutées et développées par l'intermédiaire des différents points d'extension du Framework. Tous ces éléments d'architecture ont permis aux centres de données du LAM et de l'IAS de personnaliser leurs systèmes d'information pour leurs différents projets. Parmi les fonctions non décrites dans cet article, le support d'OpenSearch, la gestion des fils RSS, la gestion des utilisateurs, la gestion des espaces de stockage pour l'utilisateur, le système de template pour la présentation des données, l'internationalisation, le design des URL, l'API d'accès à une source de données, la construction des formulaires de requête WYSIWYG sont autant de fonctionnalités qui permettent à SITools2 d'être une plateforme générique pour les systèmes d'information scientifiques.

Bibliographie

[1] Noelios Technologies, Restlet, <http://www.restlet.org>

[2] Sencha, Ext-Js, <http://www.sencha.com>