E T H E R
Demande d'utilisation de ressources

28 novembre 2013

TITRE DU PROJET

PROMISE
Pérennisation des données sols, modernisation des interfaces et des outils de gestion des flux de données pour les stations d'observations.

RESPONSABLE DU PROJET

Franck Gabarrot
Contact : Franck.Gabarrot@univ-reunion.fr
Tél : +262 262 93 86 09

LABORATOIRE DU PROPOSANT

Observatoire des Sciences de l'Univers à la Réunion - UMS 3365
Université de la Réunion,
Bâtiment S4B,
15, avenue René Cassin
97715 Saint-Denis Cedex 9

SOUTIENS DU PROJET

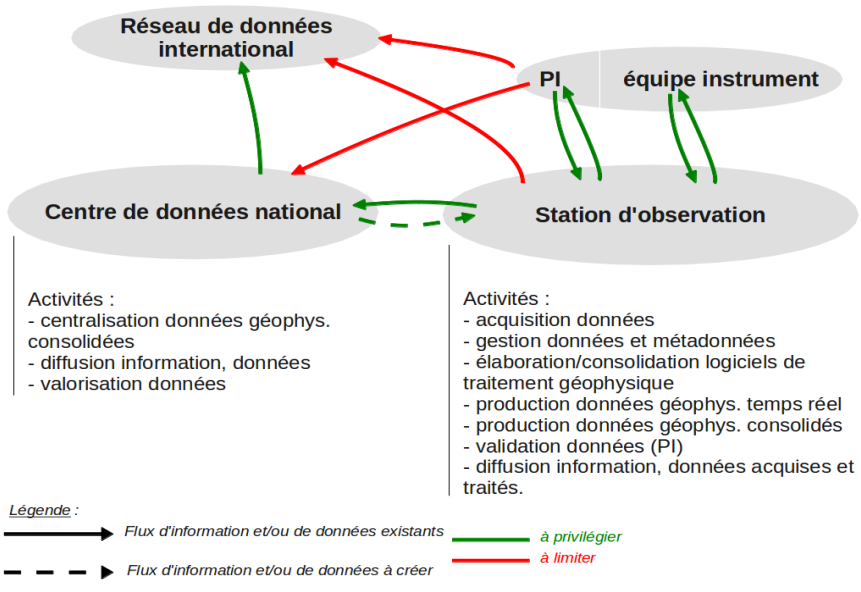
- Jean-Pierre Cammas, directeur de l'OSU-Réunion (CNRS & Université de la Réunion) et responsable de l'OPAR.
- Philippe Keckhut, coordinateur national du SO NDACC France et responsable scientifique de l'OHP.
- Martial Haeffelin, coordinateur du SOERE ROSEA, directeur du SIRTA.

CONTEXTE DU PROJET ET POSITIONNEMENT

Les SNO ont un mandat de délivrer des données géophysiques de qualité. Le succès de cette tâche sur les durées visées (décennies) et en intégrant les standards internationaux dépend d'un bon enchaînement des étapes successives depuis l'acquisition des données par l'instrument jusqu'à son utilisation géophysique. Aujourd'hui le flux des données n'est pas parfaitement maîtrisé, pas suffisamment consolidé et pas normalisé entre les différents sites de mesures et les services d'observations atmosphériques. Même si aujourd'hui les SNO remplissent leurs objectifs en étant à peu près à jour sur la fourniture de données géophysiques, notamment via ETHER concernant le NDACC, les données brutes ne sont pas ni pérennisées (et même certaines sont perdues) ni suffisamment documentées.

Le réseau NDACC, inséré depuis 1991 dans un réseau international a de fait tardé à trouver un ancrage dans le cadre des infrastructures Européennes. Aujourd'hui, le processus est enclenché via les projets européens d'infrastructure ARISE et NORS. Le premier valorise la synergie entre les différents réseaux reposant sur la notion de super-sites nœuds de réseau, indispensable à une structuration nationale. Le second projet correspond à une contribution à Copernicus concernant la

mise à disposition des données en temps quasi réel. Ces évolutions importantes mettent clairement en évidence les difficultés décrites ci-dessus et entrainent un surcroit de travail sur la gestion de flux de données du fait que les outils standards n'ont pas été déployés faute de ressources. Le SOERE ROSEA dans lequel le NDACC est impliqué pour une partie, montre qu'il existe les mêmes difficultés sur les autres super-sites français, mais que la structuration ROSEA permettrait de mutualiser les efforts, d'automatiser et de standardiser les outils, et d'améliorer l'interface avec le centre d'archivage ETHER. Les moyens du SOERE AURORE déployés dans le cadre ICARE suivent une même logique de standardisation bien que cette initiative ne concerne pas en amont la collecte de données dans les sites. Il y a donc un besoin urgent de coordination pour optimiser et rationaliser les efforts mais également un soutien au moins ponctuel pour effectuer par les super-sites un saut quantitatif sur le transfert des données et de la standardisation des outils.



Echanges entre les différents acteurs et activités respectives du centre de données national et de la station d'observation.

Les difficultés des stations par rapport à l'enjeu de la pérennisation des données acquises

Les stations d'observation sols sont un des fournisseurs primaires de données géophysiques pour l'observation de la Terre. Ces stations ont pour objectif d'assurer des acquisitions de qualité, c'est-à-dire de réaliser des mesures fiables (la donnée) et d'enregistrer l'information nécessaire pour retrouver les conditions d'acquisition et les moyens d'exploiter la donnée (la métadonnée). Ces données et métadonnées acquises sont les éléments de base qui vont être ensuite traités, validés, formatés, retraités, etc ... pour finalement être mis à disposition de la communauté au travers de produits élaborés distribués et valorisés par les centres de données nationaux. Mais quid de la pérennité des données acquises ?! Dans la pratique, les stations d'observation se débrouillent avec cette tâche, sans méthodologie commune, sans outils adéquats et avec des moyens informatiques de stockage parfois peu fiables et surtout dont la pérennité n'est pas assurée. Avec une archive de 10 ou 20 ans il est déjà difficile de traiter de façon uniforme avec un code de traitement unique l'archive d'un instrument. Tout simplement parce que l'on ne retrouve pas l'historique des conditions d'acquisition ou de configuration de l'instrument, ou encore parce qu'on ne retrouve pas les outils de lecture des données et la documentation du format. Alors on peut se poser la question de ce qu'il en sera dans plusieurs dizaines d'années. Et même sans parler de cette perte de l'historique de l'acquisition, le manque de standardisation dans la gestion de données et le manque d'outils rendent difficile l'exploitation de la synergie entre différents capteurs et différentes équipes.

Les difficultés des stations par rapport à la gestion des flux de données et à l'enjeu de l'alimentation des centres de données nationaux

Pour la personne en charge de la gestion des données station, assurer la gestion des flux de données d'une station d'observation sans les outils adéquats de transfert de données, de contrôle d'intégrité des fichiers, de gestion d'anomalies et d'automatisation de traitements équivaut à explorer quotidiennement le mythe de Sisyphe. Les conséquences directes des difficultés de gestion des flux de données sont :

- La perte d'information et une quasi absence de démarche qualité : on assure le minimum comme on le peut !
- Des équipes extérieures qui viennent installer des instruments et qui déploient leurs propres gestion des flux : on ne récupère pas certaines données acquises dans les stations.
- Des réponses très longues aux sollicitations de mise en place de nouveaux flux de données pour répondre aux attentes des réseaux internationaux ou des efforts d'organisation nationaux.

Sans les outils adéquats, il est impossible de déployer des flux de données systématiques et fiables, ce qui rend l'interface avec les centres de données difficile.

Les stations d'observation comme centre d'expertise sur la gestion des données sols

Une station d'observation n'est pas un centre opérationnel. Elle y tend avec la volonté de mise en place de flux de données organisés et systématiques vers des centres de données extérieurs, mais elle reste un lieu autour duquel on expérimente des méthodes d'acquisition, on met au point et on améliore des algorithmes de traitement, etc. On se doit donc d'assurer en même temps des services quasi-opérationnels sur la gestion des flux de données, et permettre de la souplesse sur l'arrêt et la reprise de ces flux, sur l'évolution des logiciels de traitement, sur les possibilités d'interaction avec la donnée offerte aux responsables scientifiques des instruments (PI), etc. Définir des outils en adéquation avec les besoins des stations demande une expertise que l'on ne peut trouver que chez les acteurs quotidiens de cette gestion de données. On positionne donc au travers de ce projet les stations d'observation comme centre d'expertise sur la gestion des données sols, avec un rôle de pouponnière similaire à celui des centres d'expertise scientifique tels que définis dans les pôles nationaux ETHER ou encore ICARE. C'est le positionnement actuel du groupe de travail données (GTD) dans le SOERE ROSEA, groupe de travail qui réunit l'expertise sur la gestion des données des principales stations d'observation françaises.

L'évolution des réseaux de données internationaux : des nouveaux défis à relever au travers d'une synergie centre de données - station

Afin de faciliter l'utilisation et l'intercomparaison de données géophysiques de sources différentes (satellites ou sols), un effort particulier a été déployé ces dernières années par la NASA, l'ESA et le réseau NDACC pour définir et déployer un standard pour l'écriture de métadonnées : le format GEOMS (Generic Earth Observation Metadata Standard[1]). Le projet NORS[2] est un exemple de mise en œuvre de cet effort côté NDACC afin de démontrer la pertinence des données NDACC pour le contrôle qualité des produits du service GMES pour le suivi de la composition de l'atmosphère et du climat (MACC-II). Nous en avons fait l'expérience, et malgré tout l'aspect positif qu'il en ressort, sans outils adéquats pour la gestion des données et des métadonnées, il est difficile de supporter ce type d'effort sur le long terme. Au travers des projets mis en place, que cela soit au niveau national ou international, il apparaît évident que ce type de démarche visant à faciliter l'utilisation et l'accès temps réel à des données sol de qualité est plébiscité. Le projet européen ACTRIS en est un autre exemple pour notre communauté. Pour y répondre il faudra que la gestion des données et des

[1] Retscher, C., De Mazière, M., Meijer, Y., Vik, A.F., Boyd, I., Niemeijer, S., Koopman, R.M., Bojkov, B., The Generic Earth Observation Metadata Standard (GEOMS), 2011, http://avdc.gsfc.nasa.gov/PDF/GEOMS/geoms-1.0.pdf.

[2] De Mazière, M., Network of Remote Sensing Ground-Based Observations for GMES Atmospheric Service. MACCII KO, Reading, 2012, http://www.gmes-atmosphere.eu/internal/meetings/maccii_assembly_reading/plenary/NORS4MACCII.pdf.

métadonnées des stations d'observation sol se modernise et que l'interaction station – centre de données se fluidifie et se normalise.

OBJECTIFS DU PROJET ET ACTIONS ASSOCIEES

Pour répondre aux enjeux actuels et futurs des stations d'observations et du NDACC France, nous avons définis trois objectifs avec des actions associées, l'ensemble ayant été réfléchi pour être compatible avec les travaux et les développements menés par le SOERE ROSEA. Les travaux et développements menés dans ce projet ont aussi vocation à s'intégrer avec ceux du SOERE ROSEA dans le cadre d'une future Infrastructure de Recherche Atmosphère.

Les objectifs du projet

1. **Initier une réflexion et un recensement des besoins sur la pérennisation des données des stations d'observation** : à l'heure actuelle, qui a pour mission d'assurer la pérennisation des données acquises et conserver l'information des conditions d'acquisition et de formatage pour en assurer l'exploitation sur les générations futures ? Quelles sont les données à pérenniser ? Quelles vont être les difficultés techniques et organisationnelles à surmonter ? Quelles actions serait-il bon de mener ? Quelles solutions peuvent apporter les centres de données comme ETHER ? … Une réflexion qu'il est important de mener autour des centres de données en y invitant notamment le GTD ROSEA, représentatif de l'expertise station sur la gestion des données et représentatif de la dynamique actuelle.
2. **Consolider les interfaces entre le centre de données, la station d'observation et le PI** :
    - **ETHER-station** : mettre à disposition systématiquement les données de modèles pour les stations. Avec l'objectif de centraliser dans les centres de données les besoins de données auxiliaires communs à toutes les stations et leur en faciliter l'accès.
    - **Station-PI** : développer et déployer des bases de données et des sites web stations. Avec l'objectif de fournir une visibilité sur les mesures et leur niveau de traitement en temps quasi réel et permettre au PI de valider en ligne les données.
    - **Station-ETHER**: fluidifier le flux des données vers le centre de données, construire les bases de données station afin de permettre une interrogation à distance et de faciliter le transfert des données dès leur disponibilité.
3. **Consolider la gestion des flux de données des stations d'observation** :
    - Développer et déployer des logiciels pour la gestion des flux de données dans les stations.

Les actions associées

- *Action 1 et livrable* : ***recensement des besoins sur la pérennisation des données***

Etablir un forum de discussion sur les besoins autour de la pérennisation des données des stations d'observation. Les échanges aboutiront à l'écriture d'un rapport de recensement des besoins mettant en avant les besoins bien sûr, les actions souhaitées et les pistes des solutions envisagées et envisageables pour répondre à cet enjeu de pérennisation très long terme des données.

- *Action 2 et livrable* : ***mise à disposition des données Arletty***

Mettre en place les applications nécessaires afin de mettre à disposition sur un serveur à ETHER les données Arletty au dessus de chaque station d'observation. Et cela en temps quasi-réel (dès que les données du centre européen sont accessibles) avec pour finalité d'alimenter en données auxiliares les traitements temps quasi-réel dans les stations, et avace la possibilité d'accéder à l'archive de ces données pour les re-traitements.

- *Action 3 et livrables* : ***développement des logiciels stations***

Développer et déployer des logiciels station pour la gestion des flux de données et développer les bases de données stations et leurs interfaces web. On souhaite mettre à disposition de tous des codes

opérationnels et documentés, et déployer une solution de développement collaboratif. L'ensemble des logiciels développés avec la documentation associé prêts à être employés seront livrés à ETHER qui pourra en assurer la perennité et l'évolution.

DESCRIPTION DETAILLEE DU PROJET

Les intervenants sur chaque action

	ACTION 1 Pérennisation Données station	ACTION 2 Données Arletty	ACTION 3 Logiciels stations
OPAR (F. Gabarrot)	Coordination	Suivi	Coordination
	Réalisation		Réalisation
OHP & DDU (M. Thetis)	Intervention		Expertise
ETHER (R. Bodichon)	Coordination	Coordination	Expertise
	Réalisation	Réalisation	Suivi
GTD ROSEA (C. Boitel)	Intervention		Expertise

L'organisation du recensement des besoins sur la pérennisation des données station (action 1)
Cette action va venir en complément et s'appuyer sur les travaux menés par le GTD ROSEA sur le recensement des données reliés aux instruments. Un forum de discussion va être mis en place selon deux modes :

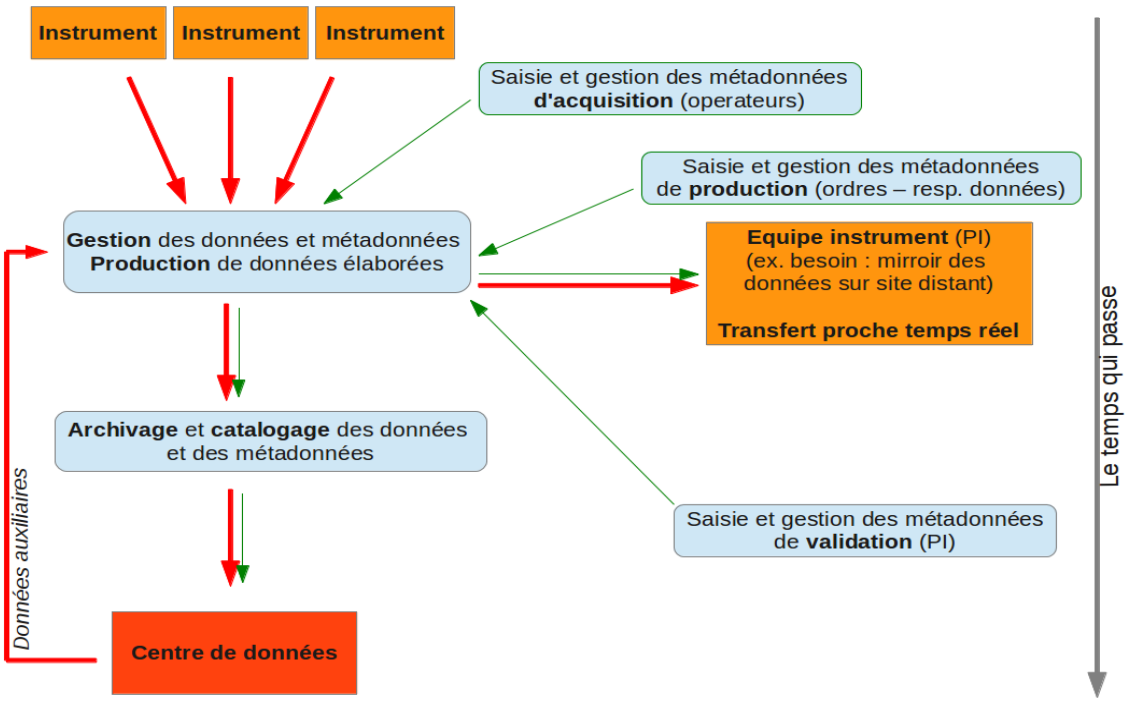
- Trois réunions sur les deux ans du projet pour ressembler physiquement le forum de discussion composé au minimum des représentants techniques des stations d'observations OPAR, OHP, DDU, des représentants d'ETHER, du GTD ROSEA (et d'ICARE au travers du GTD ROSEA). La diffusion de la tenue de ce forum sera large afin que tout autre acteur national le souhaitant puisse participer aux discussions. Les réunions pourront également être couplées à d'autres rendez-vous des groupes de travail données ROSEA ou ORAURE par exemple.
- Un forum de discussion virtuel sera mis en place en s'appuyer sur la plateforme de développement collaboratif déployée pour les développements des logiciels station (forum de discussion + wiki pour le recensement des informations).

L'OPAR et ETHER assureront conjointement la coordination de ces forums de discussion et l'écriture du document final. Michèle Thetis (LATMOS) et Christophe Boitel (SIRTA) seront sollicités pour associer aux débats la communauté la plus large et représentative possible.

Les fonctionnalités attendues pour la mise à disposition des données Arletty (action 2)
Les stations d'observation utilisent des voix et des moyens variés pour rapatrier et conserver les données de modèles atmosphériques, en particulier celles du centre européen. Ces données sont utilisées quotidiennement comme données d'initialisation dans les logiciels de traitement et notamment les logiciels de traitement lidar. Il s'agirait d'apporter un service de mise à disposition de ces données souhaité par l'ensemble des PIs des lidars du NDACC-France sur les trois sites principaux (OPAR, OHP, DDU). Ce service pourrait s'étendre par la suite à d'autres stations d'observation françaises.
Pour chacun des sites, un accès quotidien aux données Arletty par FTP aux échéances 0hTU, 6hTU, 12hTU, 18hTU est souhaité. Il est également souhaité qu'une archive long terme directement accessible par FTP de ces données soit assurée par ETHER.

Les fonctionnalités attendues pour les logiciels station (action 3)
Le schéma ci-dessous illustre le flux optimal des données et des métadonnées dans les stations d'observation depuis l'instrument jusqu'à la livraison au centre de données des données élaborées et validées. Les fonctionnalités nécessaires à la gestion des flux de données sont illustrées par des

cadres bleus. C'est pour assurer ces fonctionnalités, faciliter et automatiser les différentes tâches sous-jacentes, que nous devons développer des outils logiciels opérationnels et en adéquation avec le fonctionnement des stations d'observation.



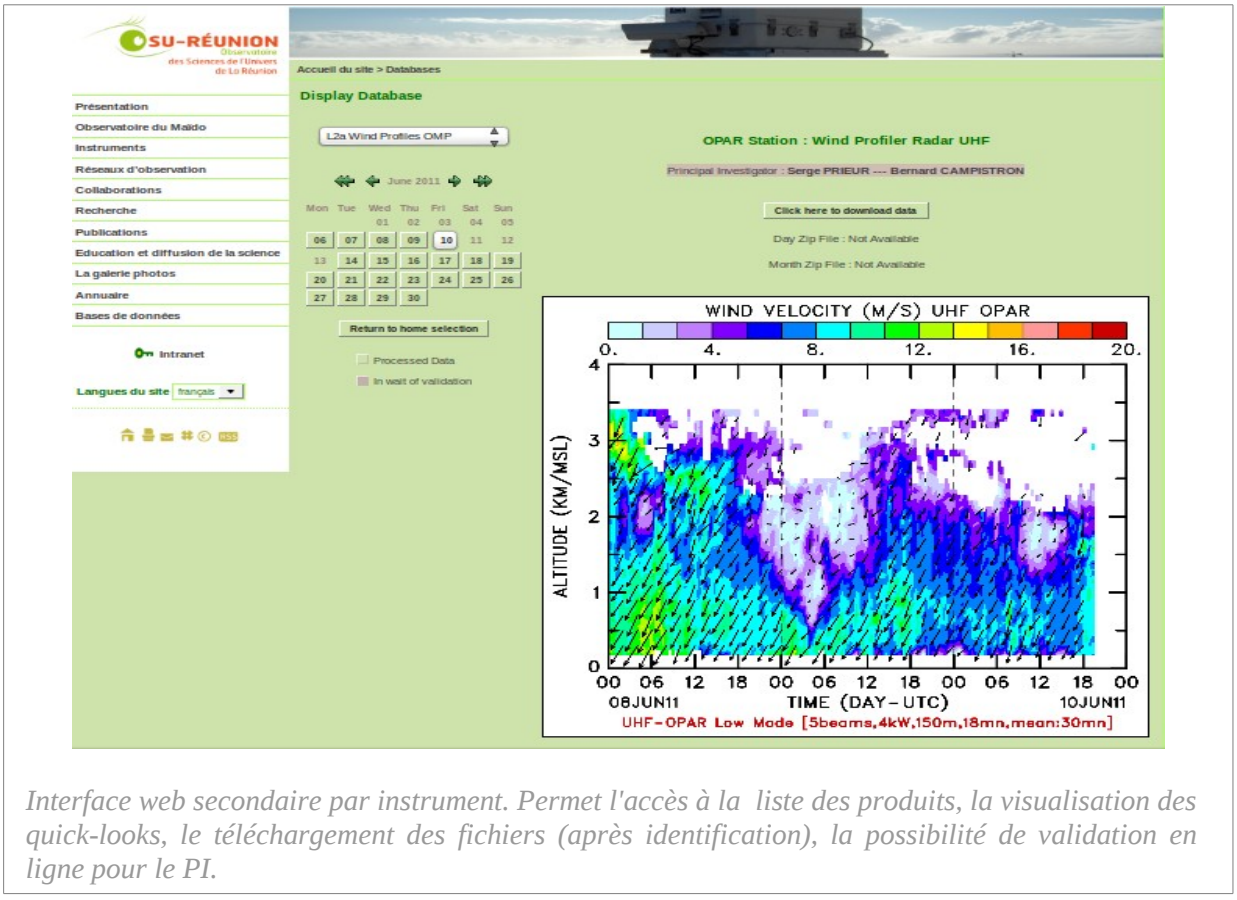
L'architecture fonctionnelle pour la gestion du flux des données et des métadonnées. Le flux des données est représenté par des flèches rouges, celui des métadonnées par des flèches vertes.

- Un logiciel pour la gestion des flux et l'automatisation des tâches : il s'agit de développer un logiciel type automate qui contrôle la présence de nouveaux fichiers de données issus des instruments et des nouveaux fichiers de métadonnées, qui contrôle leur intégrité, qui les organise selon une logique « produit » dans une base de fichiers et qui renvoie ces produits en temps quasi-réel selon différents protocoles (email, FTP, etc) à différents endroits si nécessaire (base de fichiers station, base de fichiers PI, etc). Ce logiciel aura pour rôle d'exécuter tout logiciel tiers de traitement de données afin de produire des données géophysiques élaborées et/ou des indicateurs de qualité des données (intercomparaison systématique des valeurs de différents instruments par exemple).
- Un logiciel pour l'archivage, le catalogage et la mise à disposition des données et des métadonnées : il s'agit de développer un logiciel capable d'ingérer des données organisées, de vérifier leur intégrité et de les archiver dans un espace disque redondé ad hoc. Ce logiciel devra assurer un catalogage courant des données (mode fil de l'eau) et/ou un catalogage rétrospectif (mode scanneur d'archive). Le catalogage devra être dynamique sur certains aspects, en particulier sur le statut et le niveau de traitement des données : évolution des données en accès privé vers accès retreint ou restreint vers public ; données traitées en temps réel, données traitées consolidées et évolution en données traitées consolidées validées, etc. La base de données catalogue devra être opérable à distance pour les besoins des centres de données ou des PI. En particulier, une interface internet permettra la « validation en ligne » des données. L'interface internet principale permettra aux utilisateurs (1) d'être renseigné sur la disponibilité des données ainsi que sur leur niveau de traitement par date, (2) de visualiser les quick-looks, (3) de télécharger les données en fonctions de leur droit (identification des utilisateurs et différents niveaux d'accès au téléchargement des données).
- Un logiciel pour la saisie des métadonnées : il s'agit de développer un logiciel type application distante pour la saisie de formulaire (saisie par accès web par exemple avec possibilité d'utiliser un client lourd pour les cas hors connexion). Les champs à saisir du

formulaire ainsi que le nommage du fichier devront être configurables lors de l'exécution du logiciel selon le choix de l'utilisateur (type de fiche, nom de l'instrument, etc). Lors de la saisie, ces champs seront enregistrés dans un fichier au format standard type XML et ce dernier sera stocké dans un répertoire spécifique afin d'être envoyé vers la production pour exploitation ou pour être archivé.

Un projet construit sur une expérience et des prototypes

Nous proposons de porter ce projet en collaboration avec l'équipe d'ETHER fort de notre expérience de ces dernières années pendant lesquelles nous avons élaboré et testé des prototypes, spécifié le besoin et des solutions techniques plus robustes, échangé dans le cadre de réseaux nationaux ou internationaux (NDACC-Fr, groupe de travail NDACC international sur l'algorithmie lidar accueilli à l'ISSI, GTD ROSEA). La volonté est ici de se donner les moyens d'aller au-delà du prototypage et d'aller également au-delà du simple échange dans les collaborations techniques nationales : proposer notre expertise en développement, notre expertise métier et les solutions techniques envisagées, puis, de façon communautaire, définir les solutions techniques définitives pour lesquelles nous assurerons le développement des premières versions des logiciels. Voici l'exemple des prototypes des interfaces web qui permettent d'accéder à l'heure actuelle aux bases de données catalogues de l'OPAR et de l'OHP :
https://opar.univ-reunion.fr/spip.php?rubrique4&lang=fr
http://sosgm.obs.uvsq.fr/spip.php?rubrique4&lang=fr



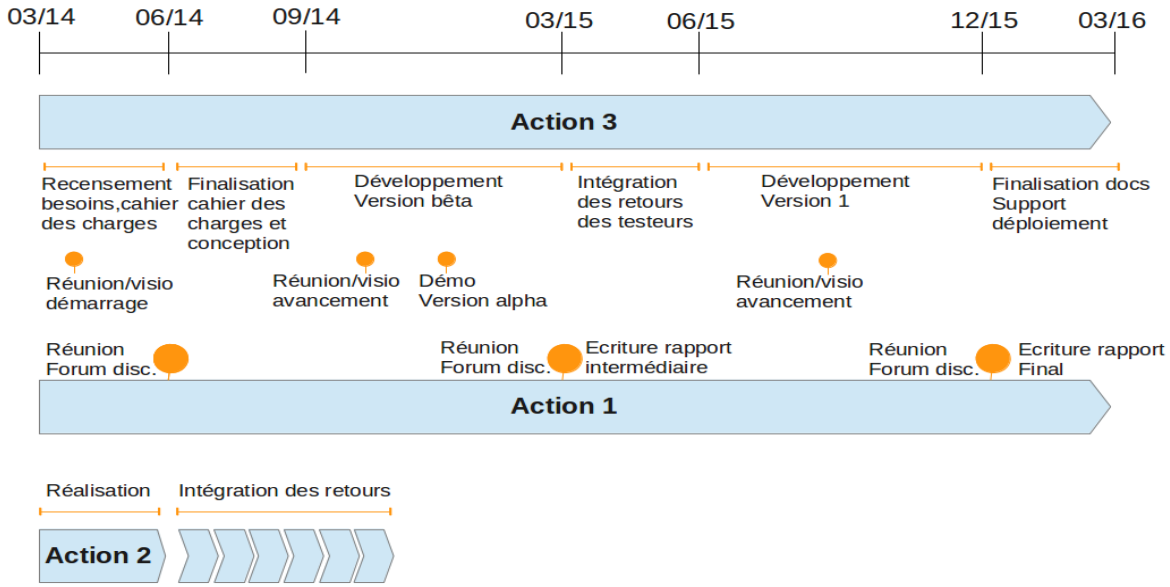
Interface web secondaire par instrument. Permet l'accès à la liste des produits, la visualisation des quick-looks, le téléchargement des fichiers (après identification), la possibilité de validation en ligne pour le PI.

La gestion des développements des logiciels station

On souhaite mettre l'accent sur l'accessibilité au projet au plus grand nombre. Cet aspect nous paraît essentiel pour recenser les besoins et anticiper les difficultés des utilisateurs : accès au code source au jour le jour, accès à jeux tests, documentation technique et utilisateurs et diffusion de l'information au travers d'un wiki. Une plateforme de développement collaboratif ouverte avec identification des utilisateurs sera mise en place à l'OPAR, ceci afin d'offrir à tous la possibilité de

rejoindre le projet en participant aux développements et d'essayer de déployer de nouvelles façon de collaborer entre les stations.

Calendrier de réalisation



MOYENS DONT DISPOSE LE PROPOSANT ET QUI SERONT AFFECTÉS À LA RÉALISATION DU PROJET

Moyens humains :
Franck Gabarrot (OPAR, OSU-Réunion UMS3365), IE CNRS, à 50 % sur le projet : gestion du projet, réalisation des actions 1 et 3, suivi de l'action 2.

Moyens informatiques :
Infrastructure informatique de l'OSU Réunion pour la mise en place d'une plateforme de développement collaboratif et infrastructure informatique de l'OPAR pour la validation des logiciels.

DEMANDE DE MOYENS

Moyens humains :

- Support d'un personnel ETHER pour la réalisation de l'action 2, pour participer à la coordination de l'action 1 et pour fournir son expertise pour l'action 3.
- Support pour un CDD de 2 ans pour la réalisation des actions 1 et 3 travaillant au centre d'expertise OPAR (UMS 3365) en lien avec les autres centres d'expertises (SIRTA-IPSL, LATMOS).

Profil du CDD : ingénieur de recherche issu d'une école d'ingénieur, profil mixte développeur avec des compétences en systèmes et réseau et systèmes d'information.

Missions :
2014

- 2 missions Réunion-Paris (2 personnes) : action 1 : lancement du forum de discussion ; action 2 : retours sur la réalisation ; action 3 : bilan sur le recueil des besoins et la conception des outils. Coût : 6k€.
- 2 missions Paris-Réunion (2 personnes = 1 personne ETHER + 1 personne ROSEA) : action 1 : préparation réunion intermédiaire du forum de discussion ; action 3 : recueil de

l'expertise ETHER et ROSEA en station avec démonstration des outils en version alpha : Coût : 6k€.

2015

- 2 missions Réunion-Paris (2 personnes) : action 1 : réunion intermédiaire du forum de discussion ; action 2 : présentation des logiciels en version bêta. Coût : 6k€.
- 2 missions Réunion-Paris (2 personnes) : action 1 : réunion finale du forum de discussion ; action 2 : présentation des logiciels en version bêta et support installation et utilisation. Coût : 6k€.

Total de la demande de moyens pour 2 années: 140k€

- **CDD IR : 100k€**
- **Missions : 24k€**
- **Frais d'organisation des 3 réunions : 3k€**
- **Frais de gestion (CNES et UR) : 13k€**

Visa du Responsable du Projet

Franck Gabarrot

Visa du Directeur de Laboratoire

Jean-Pierre Cammas