



RESIF

Cahier des charges

Version 1.0
30 mars 2018

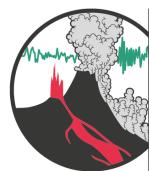
Institut de Physique du Globe de Paris

SiQaCo

Cahier des charges fonctionnel pour un outil d'automatisation de la validation des données

OBJET	Ce document contient le cahier des charges pour le développement d'un nouvel outil de validation de données sismologiques à usage des observatoires volcanologiques, sismologiques et de GEOSCOPE de l'IPGP, ainsi que les autres observatoires du réseau RESIF
--------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

DATE	30 mars 2018
STATUT	Première Version



**Observatoires
volcanologiques
et sismologiques**
INSTITUT DE PHYSIQUE DU GLOBE DE PARIS



Olivier Geber, Arnaud Lemarchand, Constanza Pardo, Jean-Marie Saurel

30 mars 2018

Table des matières

1 Introduction	2
1.1 Objet de ce document	3
1.2 Public visé par ce document	3
1.3 Références	3
1.3.1 Sites Web	3
1.3.2 Documents de Référence	4
1.4 Proposition de nom	4
2 Description Générale des Observatoires	5
2.1 Contexte	5
2.1.1 Organisation des flux de données au sein de l'IPGP	5
2.1.2 Technologies	6
2.1.3 Description générale de l'Observatoire Volcanologique du Piton de la Fournaise (OVPF)	7
2.1.4 Description générale de l'Observatoire Volcanologique et Sismologique de la Guadeloupe (OVSG)	9
2.1.5 Description générale de l'Observatoire Volcanologique et Sismologique de Martinique (OVSM)	10
2.1.6 Description générale du réseau West Indies (WI)	11
2.1.7 Description générale du réseau GEOSCOPE	13
2.1.8 Centralisation des Données et Méta-données	14
2.1.9 Accès aux données	15
2.1.10 Accès aux méta-données	16
2.2 AEQC : l'outil actuel de validation des données	16
2.2.1 Utilisation	16
2.2.2 Limites (retour d'expérience des observatoires)	16
3 Analyse des contraintes et limitations	18
3.1 Ressources infrastructures et réseaux	18
3.2 Qualité des données	18
3.3 Récupération des données depuis différentes sources	19
3.4 Sources des méta-données	19

3.5 Délais de récupération	19
3.6 Utilisation des formats SEED & StationXML	20
4 Expression des besoins	21
4.1 Principales fonctionnalités du logiciel	21
4.2 Fonctionnalités liés à l'administration du logiciel	21
4.3 Interface graphique & Visualisations	22
5 Description du produit	23
5.1 Utilisateurs du produit	23
5.2 Modularité	23
5.3 Métriques	23
5.4 Traitement des données	24
5.4.1 Fonctionnement général	24
5.4.2 Outils utilisés	24
5.4.3 Traitement des discontinuités	24
5.4.4 Gestion des traitements	25
5.4.5 Gestion des alarmes	27
5.5 Contrôle Qualité	27
5.5.1 Vérification des méta-données	27
5.5.2 Validation croisée des méta-données et des séries temporelles	27
5.5.3 Validation des séries temporelles	27
5.5.4 Modification du label de qualité des données (contrôle qualité effectué)	28
5.6 Visualisation de l'état de l'archive finale	28
5.7 Modifications manuelles de l'archive finale	28
5.8 Supervision de SiQaCo	28
5.9 Archive finale	28
6 Cas d'utilisation	29
6.1 Premier cas d'usage	29
6.2 Deuxième cas d'usage	30
6.3 Troisième cas d'usage	31
Appendix	32
A Glossaire	32
A.1 Noms	32
A.2 Abréviations	32
B Document de travail	33
C Validations complémentaires depuis l'archive finale	34

Historique des modifications

Version	Date	Auteur.e(s)	Modifications
1.0	30 Mars 2018	Olivier Geber, Arnaud Lemarchand, Constanza Pardo, Jean-Marie Saurel	1 ^{re} version

Chapitre 1

Introduction

Les observatoires volcanologiques et sismologiques de l'IPGP acquièrent des données sismiques en temps réel depuis des stations qui peuvent avoir un enregistrement local. Ces données en temps réel servent principalement à la surveillance, pour répondre aux autorités locales (risque d'éruption volcanique, séismes ressentis et alerte au tsunami par exemple), et sont par la suite archivées aux observatoires et à l'IPGP. Ces données archivées servent elles aux besoins de la recherche scientifique.

Or, à cause de la fragilité inhérente aux transmissions hertziennes (VSAT, Wifi, radio) utilisées, il est nécessaire de récupérer des données sur les stations afin de boucher les trous provoqués par la rupture occasionnelle des communications temps-réel, ce qui peut générer des données incomplètes ou même erronées. De plus, bien que les interventions multiples sur les stations soient documentées dans un système d'information interne aux observatoires (WebObs), ce sont des opérations manuelles qui peuvent avoir des répercussions dans les méta-données et engendrer des erreurs. Or, pour pouvoir être distribuées et utilisées ultérieurement, les données et méta-données archivées doivent être les plus "propres" possible : cela signifie qu'il faut boucher les trous et traiter les recouvrements dans les données, ainsi que les erreurs possibles dans les méta-données, afin que les données puissent être validées.

Ainsi, plusieurs outils ont été développés pour traiter automatiquement les trous et les recouvrements dans les données et pour gérer la redemande de données. Cependant, la pérennité de ces outils est trop dépendante de leur développeur initial. De plus, ils ne sont pas facilement généralisables.

Les données issues de ces procédures sont ensuite validées à Paris à travers toute une chaîne de vérification aujourd'hui lancée manuellement par des étudiant.e.s dans le cadre d'une formation de l'école doctorale de l'IPGP. Là aussi, ces vérifications s'appuient sur des outils maison développés il y a quelques années.

Nous souhaitons désormais développer un outil logiciel généralisable, dans un premier temps, à l'ensemble des stations du réseau des observatoires volcanologiques de l'IPGP et du réseau Géoscope, collaboratif, modulaire et qui pourra évoluer, par l'utilisation d'un système de plugin, pour accueillir de nouveaux types d'équipement, de format de données etc. Il pourra par la suite être utilisé par d'autres centres de données intéressés. Ce logiciel permettra d'automatiser la validation des données à la fois aux observatoires et au centre de données de l'IPGP, ce qui en accélérera la mise à disposition et la distribution aux utilisateurs.

1.1 Objet de ce document

Ce présent document a été réalisé à partir de l'expression des besoins des différents observatoires volcanologiques et sismologiques de l'IPGP, ainsi que du réseau GEOSCOPE, et des retours sur les outils utilisés jusqu'à présents, afin de présenter l'idée générale du développement qui sera effectué, quelles fonctionnalités nous souhaitons lui donner et comment le logiciel a été pensé, c'est à dire quelle sera sa forme.

Dans un premier temps, nous allons présenter les différents observatoires et les différents flux de données, afin de rappeler le contexte du développement, puis nous discuterons des outils qui ont été utilisés jusqu'ici. Nous exposerons ensuite les besoins exprimés par les différents observatoires, puis les différentes contraintes et limitations à prendre en compte.

Une fois validé, ce cahier des charge servira à la rédaction d'un dossier de conception qui présentera l'architecture de la solution proposée et les différents outils que nous utiliserons pour développer le logiciel.

1.2 Public visé par ce document

Ce document s'adresse principalement aux personnes travaillant aux observatoires volcanologiques et sismologiques de l'IPGP et à GEOSCOPE, mais aussi aux autres organismes intéressés par ce développement, et plus particulièrement aux personnes en charge de la récupération et de la validation des données provenant des stations sismologiques. Ce sont eux qui valideront ce présent cahier des charges qui permettra de définir le déroulement et les objectifs du développement logiciel qui en découlera.

1.3 Références

1.3.1 Sites Web

GEOSCOPE	geoscope.ipgp.fr
Volobsis	volobsis.ipgp.fr
RESIF	resif.fr
Centre de données de l'IPGP	centrededonnees.ipgp.fr
FDSN - Federation of Digital Seismograph Networks	fdsn.org
USGS- U.S. Geological Survey	usgs.gov
IRIS - Incorporated Research Institutions for Seismology	iris.edu
The Mini-SEED library	github.com/iris-edu/libmseed
ISPAQ - IRIS System for Portable Assessment of Quality	github.com/iris-edu/ispaq
Seismology Software and Manuals	iris.edu/manuals
L'outil Wave Form Catalog	orfeus-eu.org/data/eida/Web Services/wfcatalog
Outil de monitoring et de visualisation de données	grafana.com

1.3.2 Documents de Référence

L'un des documents de référence pour le projet SiQaCo est le manuel SEED qui décrit le format général d'échange des données. Ce manuel est distribué par IRIS à l'adresse : fdsn.org/seed_manual/SEEDManual_V2.4.pdf où il est possible de télécharger la dernière version (2.4) du manuel.

Le schéma du format StationXML se situe quand à lui à l'adresse suivante : fdsn.org/xml/station

1.4 Proposition de nom

Un premier choix de nom pour ce projet est "SiQaCo" pour *Seismic Quality Control*

Chapitre 2

Description Générale des Observatoires

2.1 Contexte

2.1.1 Organisation des flux de données au sein de l'IPGP

VOLOBSIS est le portail de distribution des données des 3 observatoires volcanologiques et sismologiques de l'IPGP. Ces observatoires opèrent des stations sismologiques courte-période analogique, courte-période numérique 3 composantes, moyenne-bande numérique et large-bande numérique. Ces données sont utilisées pour le suivi de l'activité volcanique et tellurique régionale, ainsi que l'alerte aux tsunamis.

Il a 4 réseaux de stations et 3 observatoires :

- Le réseau GL (Guadeloupe), localisé sur l'archipel de Guadeloupe, qui envoie ses données à l'OVSG
- Le réseau MQ (Martinique), localisé en Martinique, qui envoie ses données à l'OVSM
- Le réseau PF (Piton de la Fournaise), localisé à l'Île de la Réunion, qui envoie ses données à l'OVPF
- Le réseau WI (Western Indies), localisé sur l'ensemble des petites Antilles, opéré par l'OVSG et l'OVSM

GEOSCOPE est un réseau global de stations sismologiques large bande :

- Le réseau G (Géoscope), localisé sur l'ensemble de la planète, opéré par l'IPGP et l'EOST

Ces réseaux envoient leurs données en temps réel à l'IPGP, et sont validées à T - 1 an. Lorsque les données sont complétées et validées, elles sont redistribuées à la communauté et vers les centres de données de RESIF et de l'IRIS DMC.

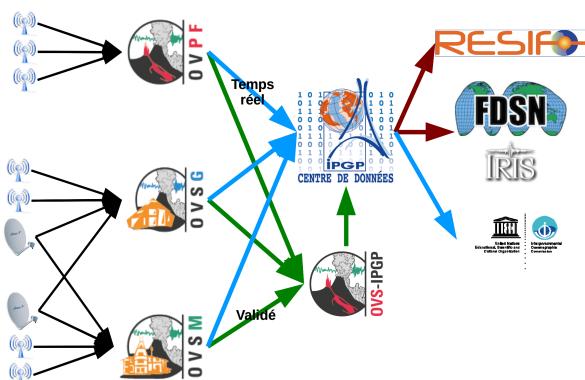


FIGURE 1 – Distribution des données – Source : Nœud A VOLCANO : bilan et perspectives (Poster), 2017

RESIF est "un équipement national d'excellence pour l'observation et la compréhension de la Terre interne. C'est un instrument ambitieux permettant à des disciplines comme la géodésie, la sismologie, la gravimétrie, d'acquérir de nouvelles données de qualité et ainsi de progresser dans notre compréhension de la dynamique de notre planète.

Réparti sur l'ensemble du territoire français, il permet de mesurer l'activité du sol sur des échelles de temps allant de la fraction de seconde à la décennie. En cette période de profonde évolution, RESIF est un outil de recherche de pointe qui aide ainsi à une meilleure identification des risques et des ressources naturelles, afin de mieux les gérer.

Cet instrument s'intègre aux dispositifs européens et mondiaux d'instruments permettant d'imager l'intérieur de la Terre dans sa globalité et d'étudier de nombreux phénomènes naturels. " (*Source : resif.fr, dernière connection le 21 Mars 2018*)

2.1.2 Technologies

Les observatoires de l'IPGP utilisent des capteurs analogiques et numériques.

Les capteurs analogiques envoient un signal continu qui est numérisé à l'observatoire. Ce signal est transmis par modulation/démodulation de fréquences, il est continu et numérisé à l'observatoire. Il n'y a pas de buffer en station, et le signal existe en toutes circonstances. Cela veut dire qu'en cas de panne, le signal ne sera que du bruit mais il existera toujours. De plus, comme ce signal est multiplexé lors de la numérisation avec les autres signaux analogiques des autres stations du réseau, cela veut dire que tous les signaux du réseau sont synchrones entre eux. Enfin, les stations analogiques n'ont que des capteurs verticaux une composante CP-Z.

Les capteurs "numériques" (i.e. le numériseur est présent sur la station) sont numérisés en station. Ces numérisateurs sont équipés de buffers. Les capteurs sont soit les capteurs large bande (Guralp ou Nanometrics) ou bien des capteurs courte période. Les numérisateurs sont :

- des Geosig (qui n'ont pas de buffer interne, buffer sur PC Fox)
- des Q330 de Quanterra (pas de buffer interne, buffer sur Baler)
- des Q330s de Quanterra (avec buffer interne)
- des Taurus de Nanometrics (numérisent en SEED, l'arborescence est en SDS uniquement avec les derniers firmwares)
- des Centaur de Nanometrics (qui numérisent directement en SEED, mais dont l'archive n'est pas sous la forme SDS - Seiscomp Data Structure, cf. Annexe A.2)
- des Libra II de Nanometrics (qui utilisent le VSAT, ce qui optimise le transfert de données temps réel)

2.1.3 Description générale de l'Observatoire Volcanologique du Piton de la Fournaise (OVPF)

Carte du réseau

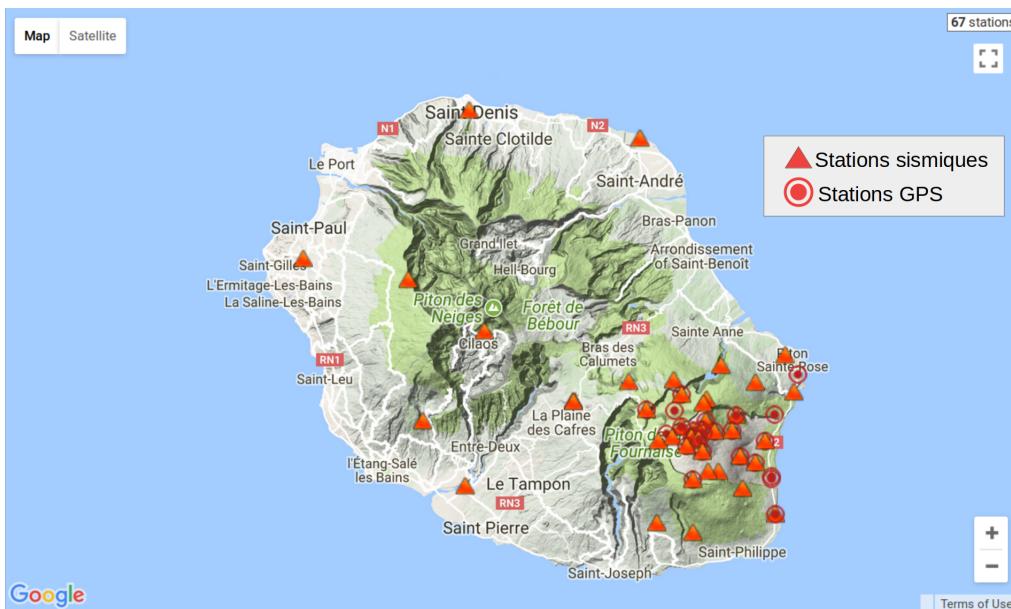


FIGURE 2 – Carte des stations de l'OVPF – *Source : site web de Volobsis*

Le réseau PF comprend 67 stations, donc 25 stations GPS permanent et 42 stations sismologiques, pour un total de 179 canaux.

Flux de données

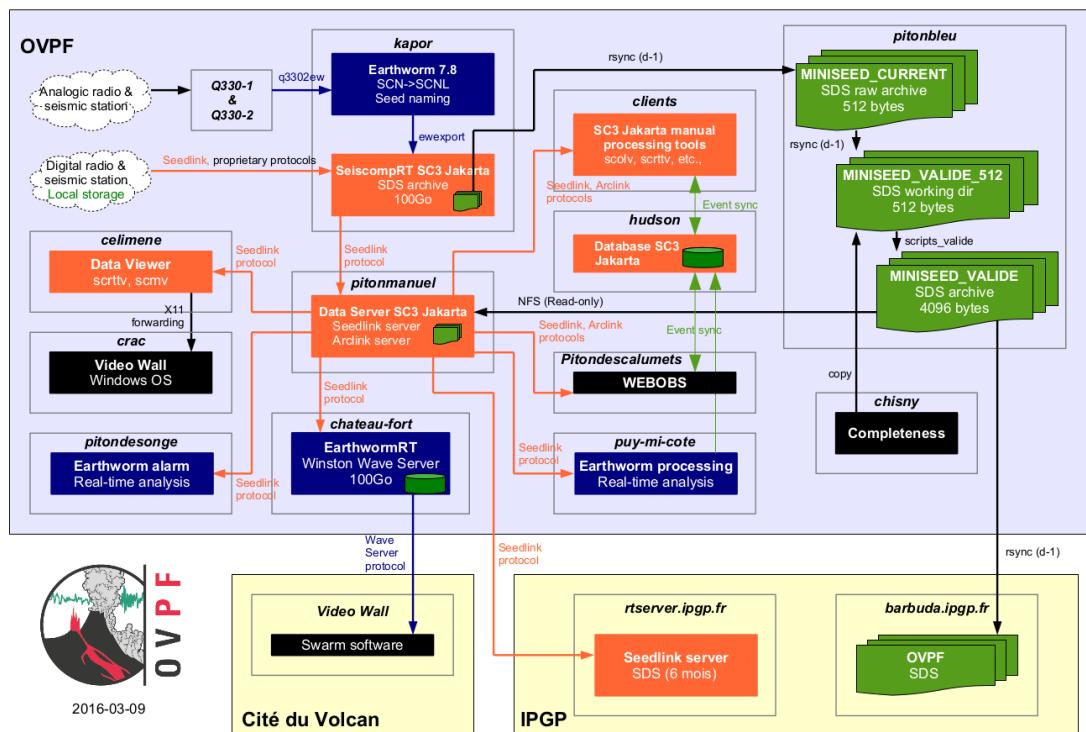


FIGURE 3 – Flux de données pour l'OVPF – Source : OVPF, 2018

Traitements des discontinuités des données

Les trous de données de l'OVPF sont en général dus à des pannes (station, transmission) qui sont de fait durables : plusieurs heures à plusieurs jours/mois, pas de trous de "secondes". La récupération des données se fait donc par fichiers de journées entières, ce qui ne pose généralement pas de problème de ressources. En pratique, il s'agit d'un script journalier qui effectue 3 passes et scrute une période passée pour l'archiver. Ce script est indépendant du type de la station, il y a des paramètres par défaut qui sont chargés depuis un fichier de config. Le rapatriement des données se fait préférentiellement à distance plutôt que sur place en station.

Les trous dans les données sont visualisés manuellement via un scan qui fait le bilan des miniseed, c'est à partir de ce scan que l'on va commander les récupérations. Il existe également un script qui permet de visualiser les trous où les données n'existent pas (en cas de panne de courant par exemple). Si aucune récupération n'a été faite, les fichiers ne sont pas découpés de minuit pile à minuit pile.

Dans les cas des stations avec numériseurs Quanterra, il y a un Baler (Baler 14D ou Baler 14F selon les cas) où sont enregistrées les données du numériseur. Ces données sont stockées sous forme de fichiers miniseed multiplexés (contenant plusieurs canaux) en blocs de 4K et d'une durée de 4h heures (environ). Les données stockées dans les baler sont récupérées à l'OVPF, si besoin, par la commande "scp".

L'outil AEQC n'était pas satisfaisant pour l'OVPF, qui a dû faire des re-développements.

2.1.4 Description générale de l’Observatoire Volcanologique et Sismologique de la Guadeloupe (OVSG)

Carte du réseau

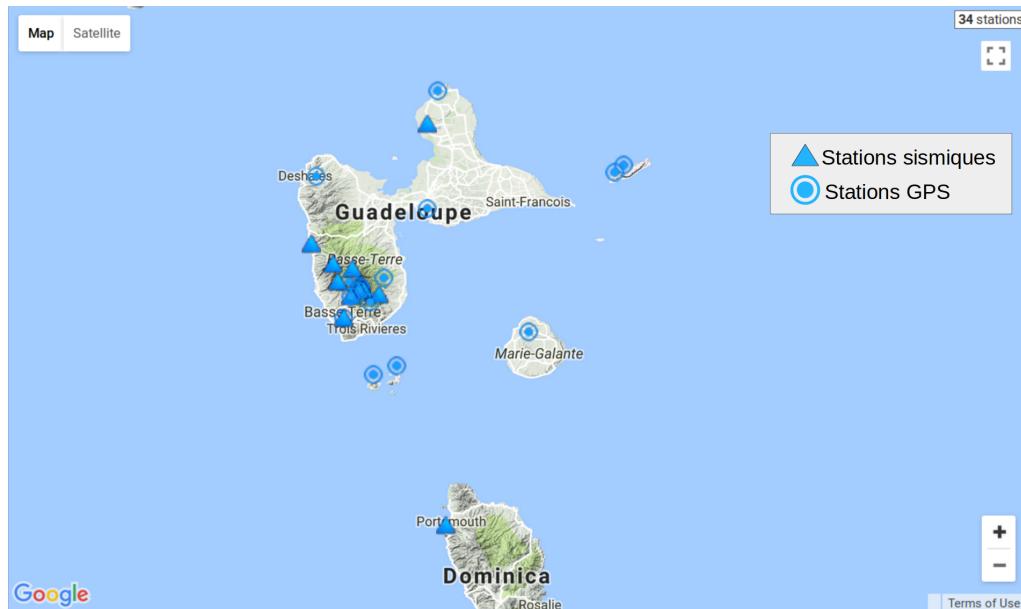


FIGURE 4 – Carte des stations de l’OVSG – *Source : site web de Volobsis*

Le réseau GL comprend 34 stations, donc 18 stations GPS permanent et 16 stations sismologiques, pour un total de 112 canaux

Flux de données

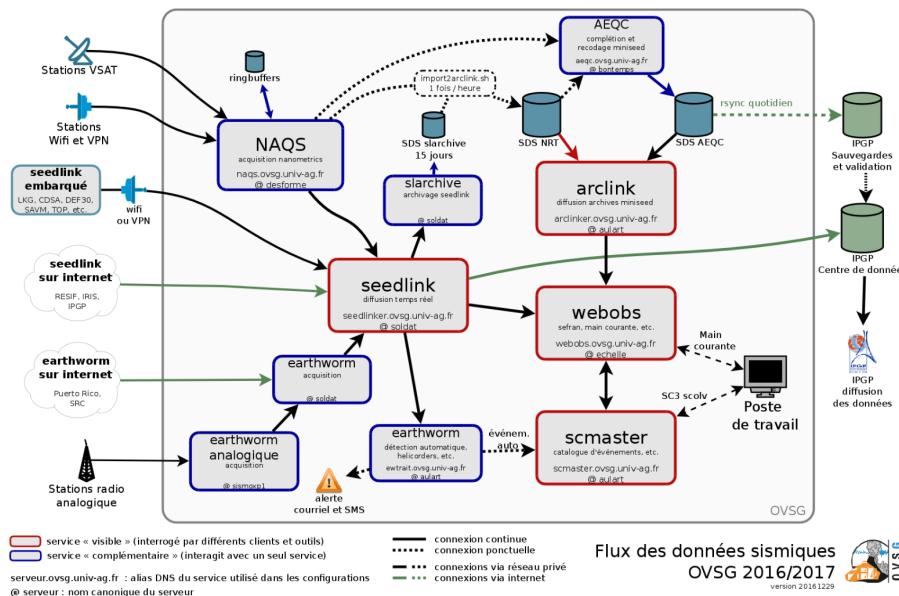


FIGURE 5 – Acquisition des données à l’OVSG – *Source : OVSG, 2017*

Traitement des discontinuités des données

L'OVSG utilise l'outil AECQ pour le traitement des discontinuités des données. Contrairement à l'OVPF, les trous dans les données sont plus fréquemment dus à des problèmes de transmission et au protocole de transmission de Nanometrics qu'à des pannes : on a donc, en règle générale, des trous de quelques secondes. Cependant, en cas de panne sur une station, il peut, de la même manière qu'à l'OVPF, y avoir des trous sur plusieurs journées.

La transmission des données vers le NAQS est basée sur le réseau VSAT, dont la bande passante (BP) est limitée, et le protocole traite les données dans le désordre (si on a un trou dans la transmission des données, le protocole VSAT va le traiter, mais pas immédiatement). Or le protocole seedlink lui traite les données dans l'ordre. Ces deux protocoles ne se comprenant pas, la donnée du trou va rester dans le NAQS et il faut attendre qu'un outil aille la chercher. AEQC a, par exemple, un délai de 4 jours pour récupérer les données (AEQC tourne 2 fois : une première fois, pendant 2 jours, afin d'effectuer les corrections sur les données qui seront placées dans une arborescence tampon, puis une seconde fois, pendant 2 jours à nouveau, afin de déplacer les données corrigées depuis le tampon des données corrigées vers l'archive finale).

La demande de traitement des discontinuités via AEQC se fait en parallèle sur plusieurs fichiers de configuration, mais à l'intérieur de ces fichiers le traitement est séquentiel. Les requêtes faites par l'outil import2arclink, qui a été développé par l'OVSG pour reconstruire toutes les heures l'archive SDS d'un jour J, sont de 1 requête par station (au lieu de 1 par canal avec AEQC, ce qui divise le nombre de requêtes par 6), ce qui ne surcharge pas le NAQS et permet de reboucher rapidement les trous.

2.1.5 Description générale de l'Observatoire Volcanologique et Sismologique de Martinique (OVSM)

Carte du réseau

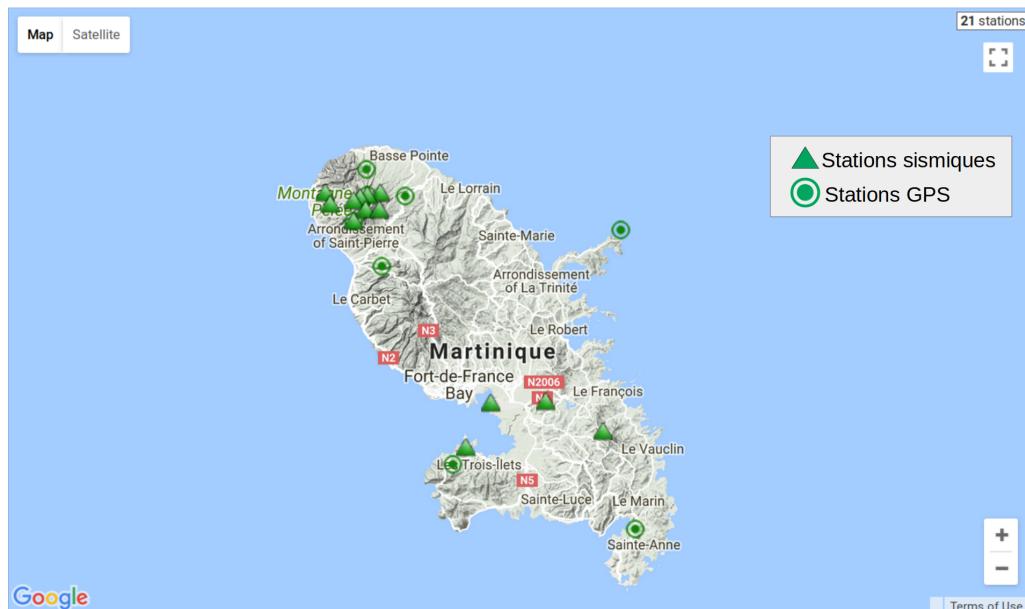


FIGURE 6 – Carte des stations de l'OVSM – Source : site web de Volobsis

Le réseau MQ comprend 21 stations, donc 8 stations GPS permanent et 13 stations sismologiques, pour un total de 72 canaux.

Flux de données

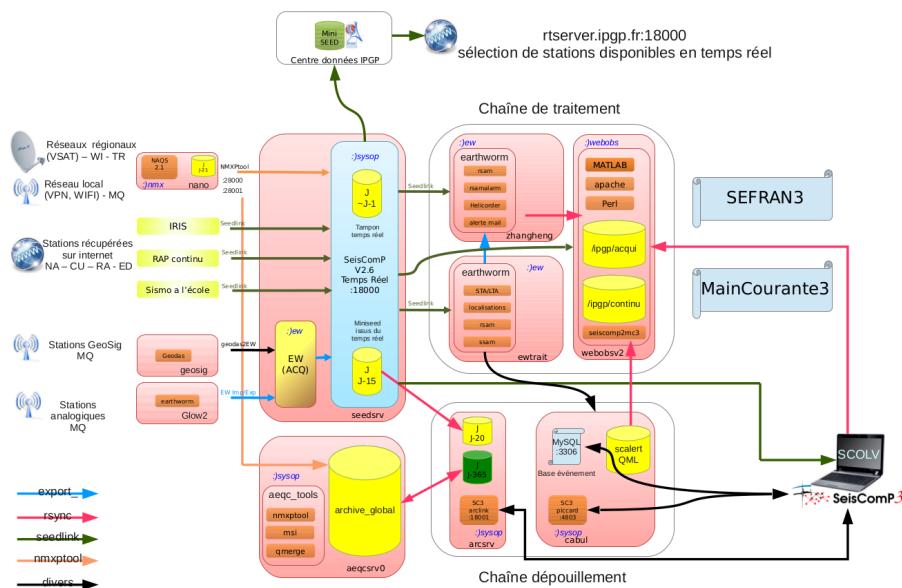


FIGURE 7 – Acquisition des données à l'OVSM – *Source : OVSM, 2013*

Traitements des discontinuités des données

Idem que pour OVSG, sauf pour l'outil import2arclink qui est propre à l'observatoire de Guadeloupe.

2.1.6 Description générale du réseau West Indies (WI)

Carte du réseau

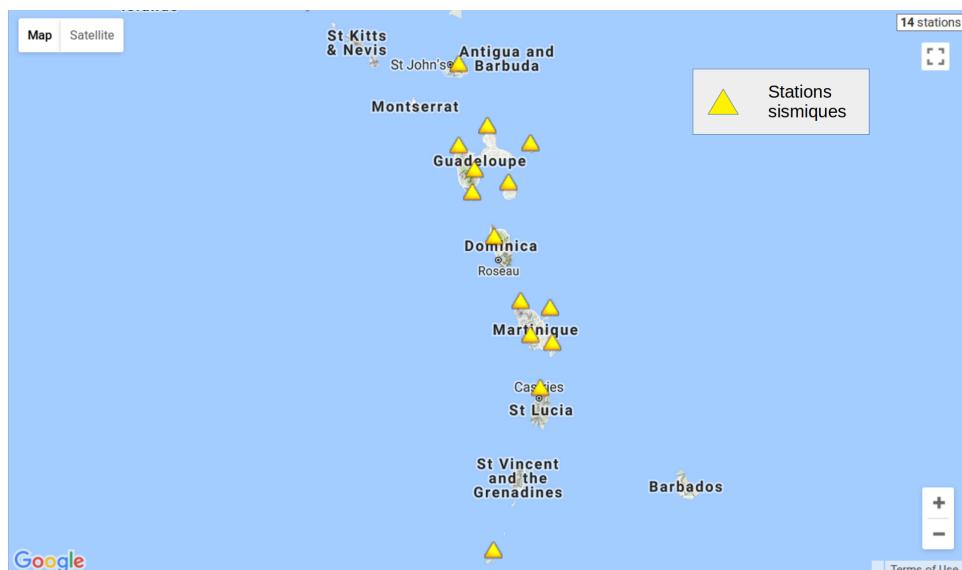


FIGURE 8 – Carte des stations des WI – *Source : site web de Volobsis*

Le réseau WI comprend 14 stations sismiques, pour un total de 112 canaux.

Flux de données

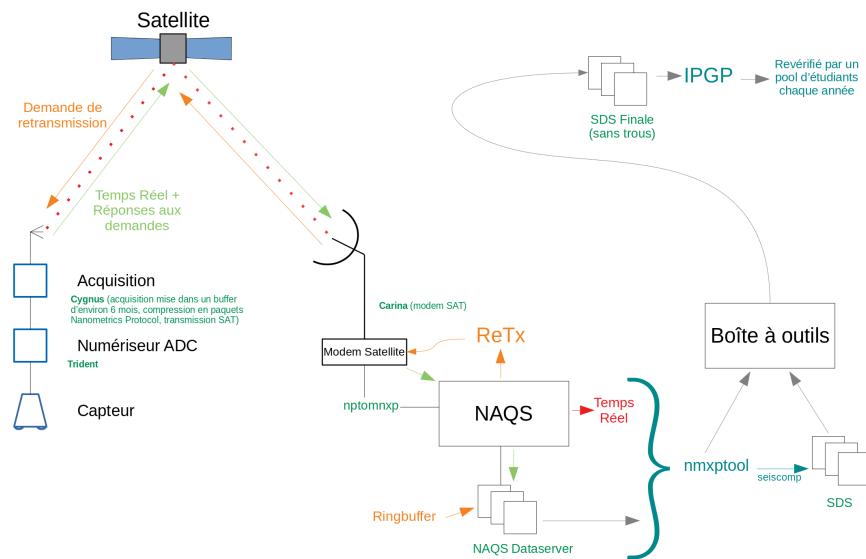


FIGURE 9 – Schéma de la chaîne d'acquisition pour WI

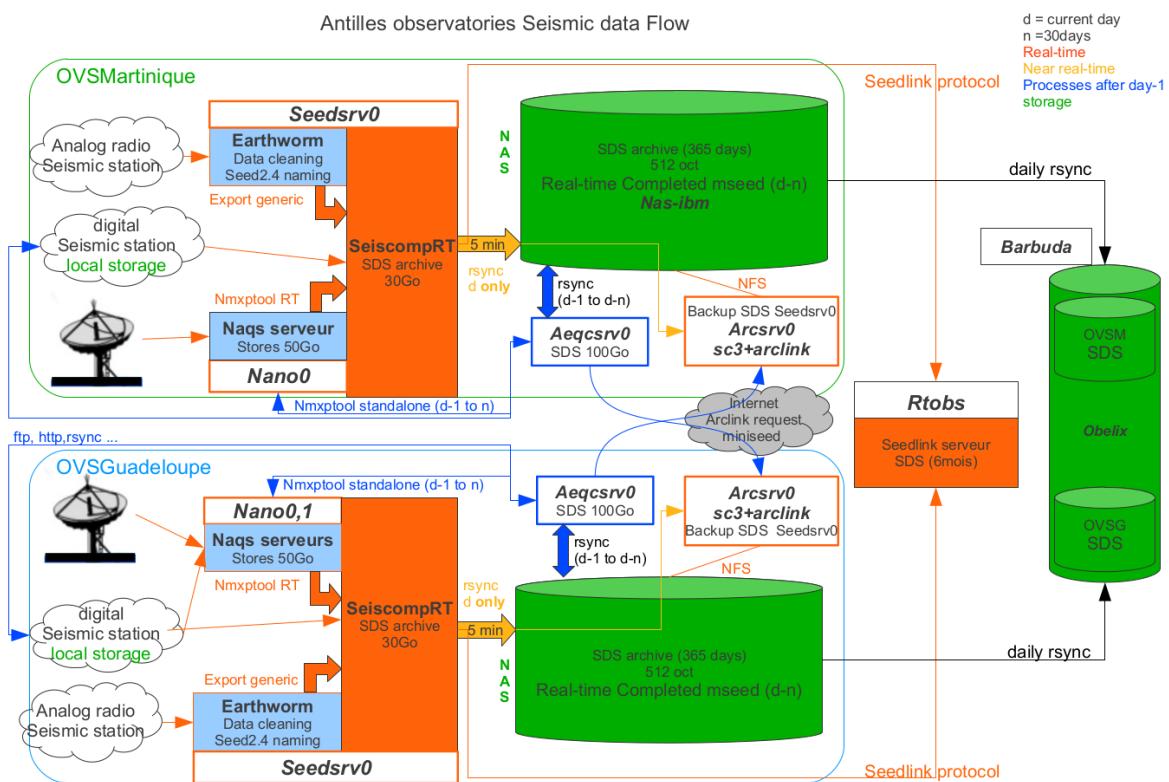


FIGURE 10 – Flux de données entre les stations des Antilles – Source : OVSM, 2013

Le réseau WI utilise le système Libra II de Nanometrics

cf. Figure 13

2.1.7 Description générale du réseau GEOSCOPE

Carte du réseau

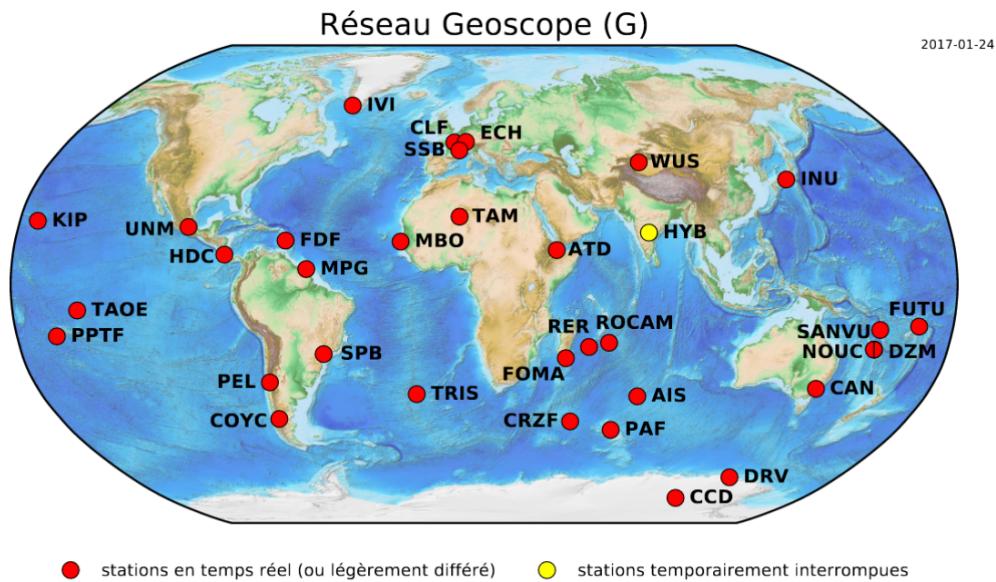


FIGURE 11 – Liste des stations du réseau Géoscope – *Source : geoscope.ipgp.fr*

Le réseau GEOSCOPE comprend aujourd’hui 34 stations en fonctionnement qui envoient leur données vers le centre de données de l’IGPG. L’EOST s’occupe des 5 stations des Terres Australes et d’Antarctique, ainsi que de la station ECH dans les Vosges. Le CEA/DASE est en charge des stations TAOE et DZM. L’IU/GSN s’occupe des stations TRIS et KIP.

L’ensemble des stations Géoscope est généralement équipé d’un sismomètre STS1 3 composantes ou d’un capteur STS2 complété par un accéléromètre, ainsi que d’autres capteurs. Les canaux des stations Géoscope sont H (sismomètres), éventuellement N (accéléromètres), K (température), D (Pression) et M (position de la masse).

Les numériseurs sont des Quanterra Q330HR sauf pour les stations du CEA.

L’information manquante est demandée à l’archive RT dans seiscomp (et pas au numériseur en station).

Seule la station INU (au Japon) est connectée directement au numériseur, il n’y a pas de seiscomp box en station.

Si une station a deux sources (seiscomp box et Baler par ex.) on interroge en priorité la seiscomp box pour récupérer des données.

Dans les cas des stations IU/GSN, les données sont rapatriées via les Web Services de la FDSN.

Sur certaines stations Geoscope, il y a 2 capteurs installés qui mesurent la même chose, où il existe donc 2 flux validés (de locID différents).

Flux de données

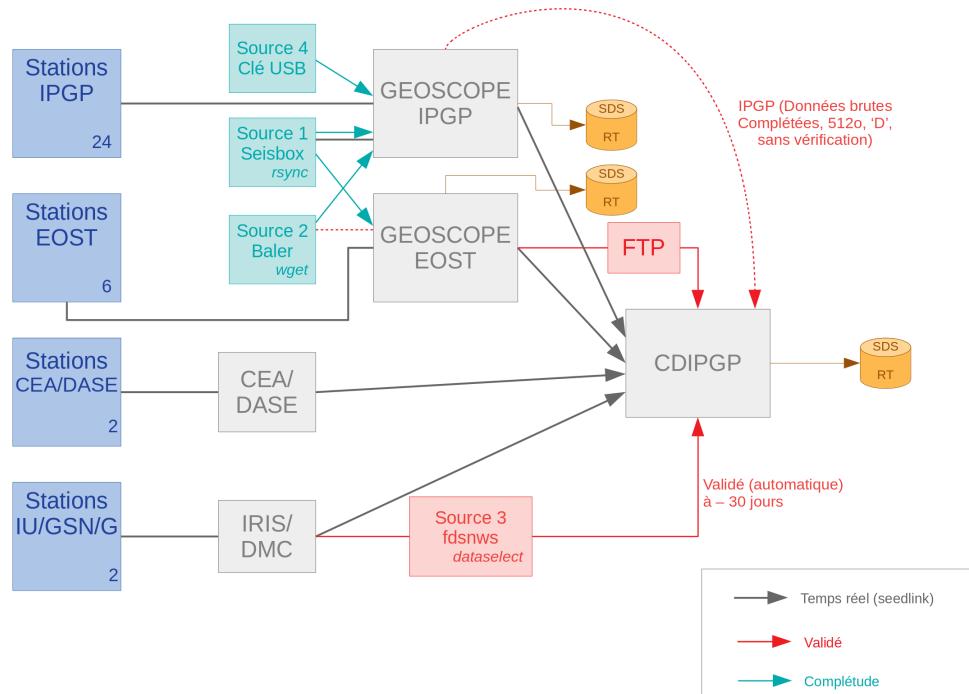


FIGURE 12 – Flux de données pour le réseau Géoscope

2.1.8 Centralisation des Données et Méta-données

Observatoires Volcanologiques et Sismologiques

Les données acquises en observatoire sont synchronisées tous les jours dans une baie de stockage primaire (Keldix) à Paris. Cette synchronisation est répliquée sur une baie de travail (ntap0svc) accessible aux utilisateurs. C'est en particulier sur cette dernière que l'environnement de validation des données et méta-données s'appuie pour accéder à une archive de travail.

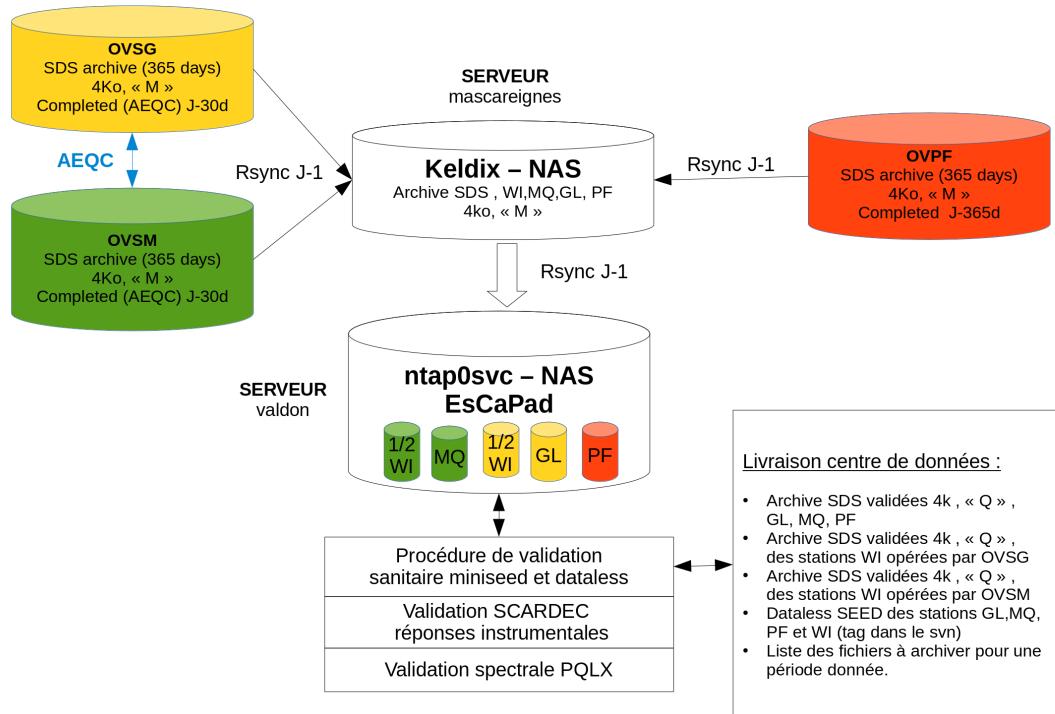


FIGURE 13 – Validation de données pour les stations des OVS – *Source : OVS, 2018*

Les méta-données des stations sismologiques sont des fichiers dataless SEED organisés par stations. Elles sont centralisées dans le SVN de l'IPGP qui permet d'estampiller des versions validées et livrées au centre de données pour archivage.

2.1.9 Accès aux données

Les données des 5 réseaux gérés par les observatoires volcanologiques et sismologiques, ainsi que par GEOSCOPE sont disponibles au centre de données de l'IPGP.

Deux flux de données sont disponibles en parallèle :

- Les données en temps réel, disponibles en miniseed via le protocole seedlink (développé par GFZ)
- Les séries validées, ie. méta-données et séries temporelles. Les méta-données sont disponibles aux formats stationXML et dataless SEED. Les séries temporelles en miniseed.

Ces données sont disponibles via les Web Services FDSN standards à partir d'un URI unique au centre de données : eida.ipgp.fr/fdsnws/.

Les 2 Web Services suivants sont mis en place :

- fdsn-dataselect : eida.ipgp.fr/fdsnws/datalog/1
- fdsn-station : eida.ipgp.fr/fdsnws/station/1

Les méta-données sont aussi disponibles sur : centrededonnees.ipgp.fr/metadata/

Les observatoires volcanologiques et sismologiques ont aussi un portail d'accès aux données : volobsis.ipgp.fr/

Ce portail permet d'accéder à la description des réseaux et des stations, à l'inventaire des données dispo-

nibles, d'utiliser une interface d'aide à la construction des requêtes pour accéder aux métadonnées des stations (au format StationXML ou Dataless SEED), et aux enregistrements sismiques pour les réseaux GL, MQ, PF et WI.

L'observatoire GEOSCOPE a, de même, son portail d'accès aux données : geoscope.ipgp.fr

2.1.10 Accès aux métadonnées

L'accès aux métadonnées se fait à partir du Web Service suivant : centrededonnees.ipgp.fr/metadata/.

2.2 AEQC : l'outil actuel de validation des données

2.2.1 Utilisation

L'outil AEQC est utilisé dans les observatoires de Martinique et de Guadeloupe, pour récupérer les données depuis les NAQS dataserver afin de compléter les données présentes aux observatoires. Pour les stations qui n'ont pas de NAQS dataserver, l'outil AEQC effectue la correction des recouvrements ainsi que la compression en 4Ko des données. Il est également sensé synchroniser les données du réseau WI (qui sont envoyées parallèlement vers ces 2 observatoires) chaque jour pour le jour précédent.

À l'observatoire de Guadeloupe, un second script import2arclink permet de compléter toutes les heures l'archive SDS pour un jour J pour les stations NAQS en utilisant la commande *nmxptool*.

2.2.2 Limites (retour d'expérience des observatoires)

L'outil AEQC, bien qu'il soit nécessaire dans la chaîne de validation de données sismologiques et volcanologiques, présente plusieurs limites.

En premier lieu, les développements de l'outil dépendent de leur développeur initial (M. El Madani Aissaoui). De plus, cet outil est complexe et difficile à mettre en place. Il consomme également beaucoup de ressources, ce qui sature les infrastructures des observatoires utilisant cet outil.

Le délai de traitement de la donnée est trop long : il faut 4 jours pour reboucher le trou dans l'archive finale servie par arclink. Cela provient de plusieurs paramètres, dont la limitation physique du Naqs dataserver : on ne peut pas traiter beaucoup de demandes simultanément (par exemple une trentaine) sinon il sature. L'outil AEQC définit une limite de 10 requêtes par canal et par jour pour la Guadeloupe. Par ailleurs, le nombre de connexions simultanées au dataserver est de 4 par défaut.

Les fichiers log générés sont souvent d'une taille > Mo, il faut chercher l'information d'un problème en utilisant la commande *grep*, ce qui n'est finalement pas pratique.

L'outil utilise plusieurs arborescences SDS pour compléter les trous dans les données (afin de ne pas perdre de données en cas de plantage), ce qui participe à sa grande consommation de ressources. Ces 4 arborescences sont les suivantes :

- Un répertoire d'entrée
- Un répertoire de travail (ARCHIVE WORK)
- Un répertoire buffer (CORRECTION BUFFER) qui permet, en cas de crash de la machine ou du logiciel, de pouvoir revenir en arrière sans perdre de données
- Un dernier répertoire qui contient l'archive finale avec les données complétées. De plus, il n'y a pas de nettoyage automatique des répertoires tampon, conduisant à une consommation d'espace croissante et une nécessité de mettre en place des mécanismes de nettoyage.

Dans le cas de l'OVSG et de l'OVSM, le répertoire d'entrée est en fait le répertoire de sortie, il n'y a donc que 3 arborescences SDS.

De plus, l'outil AEQC n'analyse que la donnée d'entrée (compressée en 512o) et pas les données de sortie.

La configuration du logiciel (via fichiers de configuration), ou le fait de devoir lancer des scripts de types différents qui ne sont pas dans AEQC (pour compléter l'outil) afin de rapatrier plus rapidement des données –en cas de séisme ressenti par exemple– ne sont pas très conviviaux. Ainsi, cela a conduit l'OVSG à développer un script “maison” afin d'avoir un bouchage des trous au bout d'une heure.

En plus des stations Nanometrics transitant par le Naqs Server, les observatoires de Martinique et Guadeloupe opèrent un certain nombre de stations de technologie différente dont les données transitent par Seedlink sur le wifi et ne passent pas par le Naqs. Il s'agit notamment de numériseurs Taurus et Centaur, dont les données stockées sur place sont accessibles par Web Services, requêtes URL ou rsync. Cependant, la version d'AEQC capable d'effectuer des requêtes de données sur ces stations est une branche différente de celle capable d'effectuer des requêtes sur le Naqs server. Il n'y a donc pas d'automatisme pour reboucher les trous de ces stations et cela est fait manuellement.

La synchronisation des données du réseau WI, effectuée par AECQ sur les 2 observatoires des Antilles (cf. Figure 13) n'est pas robuste : il arrive d'avoir des données différentes entre les deux observatoires après la synchronisation, ou des erreurs.

Enfin, AEQC n'intègre pas de visualisation du traitement des discontinuités.

Chapitre 3

Analyse des contraintes et limitations

Le développement du projet SiQaCo devra suivre plusieurs contraintes qui ont étées définies par les utilisateur.rice.s.

3.1 Ressources infrastructures et réseaux

Afin de ne pas interférer avec les données en temps réel, il faudra faire attention à l'utilisation des ressources : il faudrait éviter de multiplier les tampons et les archives temporaires, tant en nombre qu'en volume, de façon à limiter le besoin de ressources pendant le traitement.

Pour cela, il faudra pouvoir décider si l'on souhaite rapatrier une ou plusieurs journée(s) de données ou seulement une tranche de données. La durée du trou de données devra aussi être prise en compte : on pourra mettre une taille maximale de récupération (pour éviter de récupérer à distance des trous de plus de 2 mois, par exemple).

Le rapatriement de données doit pouvoir être effectué en parallèle sur plusieurs stations, mais il faudrait également pouvoir décider d'un ordre de priorité des stations dont on récupère les données.

L'outil doit pouvoir s'adapter en fonction des ressources locales et distantes. Pour cela, il faudra prendre en compte la charge système minimale d'une requête. De plus, il est parfois préférable de faire une seule requête à plusieurs : il vaut mieux demander à récupérer 1 fois 1 à 3 journée(s) que 40 fois 1 seconde.

Il a également été proposé d'intégrer un outil qui analyse automatiquement les limitations de la bande passante (qui existe déjà dans AEQC). Cependant, il peut arriver que la limitation provienne des équipements et de la charge système, plutôt que de la bande passante, donc il faudrait également définir les limites des ressources distances (CPU, etc.). De plus, la bande passante peut beaucoup varier sur une période courte, donc il peut être difficile de l'évaluer. Ce type de module pourra donc faire l'objet d'un développement futur (ie. il faudra pouvoir l'intégrer facilement par la suite). En attendant, le mieux serait de définir une limitation à priori, en fonction des machines.

Il ne faut pas perdre de vue que le protocole arclink est bientôt obsolète. Il faut utiliser les Web Services de la FDSN à la place.

3.2 Qualité des données

En ce qui concerne la qualité des données, plusieurs problèmes se posent :

Il faudrait, dans la mesure du possible, ne toucher que le minimum de blocs miniseed possible (soit les blocs adjacents). Il faut éviter de décompresser puis recompresser le fichier complet à chaque fois. A chaque décompression/recompression, on génère en effet un nouveau vecteur temps et de potentiels nouveaux blocs minised. Et comme le miniseed de source ne contient le temps et la valeur complète de l'échantillon que pour

le premier échantillon du bloc, on court le risque d'altérer la donnée plus loin dans le fichier.

Dans le cas des dataserver de Nanometrics, le paquet seedlink présent sur le numériseur n'a aucune raison d'être identique à celui envoyé en temps réel.

Il faut pouvoir récupérer les données manquantes (si on traite par tranche de journée) sans avoir à décompresser puis recompresser les blocs miniseed.

À l'EOST, pour les données du réseau FR, on vérifie, en se connectant à une seiscomp Box, la différence entre les données récupérées et le temps réel par un checksum via rsync sur des paquets de 1 ou 2 Mo : si le bloc est différent de celui de destination alors on copie.

Il faudrait pouvoir savoir si on a bien récupéré toutes les données. Par exemple, on voit un trou de données dans un ensemble sismomètre + accéléromètre, on rapatrie les données depuis le numériseur, mais on ne récupère que les données du ou des sismomètre(s) : il faut pouvoir voir que l'on a pas rapatrié les données d'accéléromètres. Pour cela, on pourrait déclarer les canaux attendus sur chaque station dans la configuration.

3.3 Récupération des données depuis différentes sources

La récupération des données se fait depuis différentes sources, et s'il faut garder à l'esprit qu'elles peuvent évoluer (changement de matériel, ajout de nouveaux formats), voici à l'instant de notre développement les différentes sources depuis lesquelles nous allons récupérer les données. On peut les classer selon deux catégories :

Sources locales

Par sources locale nous entendons les sources accessibles directement par le logiciel SiQaCo .

Ces sources locales peuvent être alimentées automatiquement (Temps Réel, SDS) ou par une action manuelle après récupération des données directement à la station (PC Fox, Baler etc.).

Sources distantes

Par sources distantes, il s'agit des sources de données que l'on interroge par un protocole d'accès à distance :

- Depuis le NAQS dataserver de Nanometrics
- Depuis le Web Service FDSN dataselect
- Depuis les numériseurs Nanometrics Centaur
- Depuis les numériseurs de Quanterra Q330 (Q330S, Q330 et Baler)
- Depuis les numériseurs Taurus de Nanometrics
- Depuis une station Guralp

3.4 Sources des métadonnées

Les métadonnées utilisées par SiQaCo peuvent se trouver sur différentes sources :

- Un Dataless SEED par station
- Une requête depuis le Web Service FDSN station
- Un fichier stationXML (non prioritaire)

3.5 Délais de récupération

En cas d'occurrence d'un séisme ressenti, les observatoires antillais ont besoin d'accéder rapidement aux données les plus complètes et valides possibles pour pouvoir localiser et caractériser le séisme en question. L'outil doit donc permettre aux utilisateur.rice.s de paramétriser des délais de récupérations de données inférieurs à la journée.

3.6 Utilisation des formats SEED & StationXML

Les données sismiques proviennent de différents types d'équipements (accéléromètres, sismomètres, etc.), numérisés par des différents types de coffrets de numérisation qui eux-mêmes sont développés par différents constructeurs, et utilisent donc souvent des formats de données propres au fabricant.

Depuis 1988, la FDSN (*International Federation of Digital Seismograph Networks*) a développé un format standard d'échange de données sismiques : le SEED (*Standard for the Exchange of Earthquake Data*) qui est aujourd'hui utilisé par la plupart des réseaux sismologiques du monde. L'ensemble des stations des observatoires de l'IPGP, et du réseau RESIF transmettent et archivent leurs données au format miniseed, et les méta-données sont elles actuellement produites au format dataless SEED.

Les données collectées en temps réel sont transmises par blocs de 512 octets, tandis que les données sont archivées par blocs de 4096 octets.

Le format dataless SEED tend à disparaître au profit du format StationXML qui est le nouveau standard de méta-données pour la communauté internationale de sismologie.

Le format de données SEED, bien que standardisé, est voué à évoluer dans le futur. Ainsi, notre logiciel sera développé en se basant sur des données au format miniseed et des méta-données aux formats dataless SEED et StationXML, et l'archive finale des données validées par SiQaCo se fera sous la forme d'une arborescence SDS mais il ne devra pas être figé sur ces standards.

Chapitre 4

Expression des besoins

Ce paragraphe a pour but de synthétiser les demandes faites par les différents utilisateur.rice.s.

4.1 Principales fonctionnalités du logiciel

Le projet SiQaCo a pour vocation de développer un logiciel dont la fonctionnalité première est de créer et de mettre à jour une structure finale de données et de méta-données validées à partir de différentes sources de données et de méta-données. La seconde fonctionnalité du logiciel sera d'effectuer le contrôle qualité sur les données et méta-données.

Le logiciel comprendra donc deux grands axes de traitement des données et des méta-données : une première partie qui visera à compléter les données, et qui ira modifier, s'il le faut, les données. La seconde partie quand à elle comprendra l'ensemble des algorithmes nécessaires à la validation des données et des méta-données (ie. le contrôle qualité).

4.2 Fonctionnalités liés à l'administration du logiciel

Les administrateurs informatiques ont demandé que l'outil développé soit synthétique, simple à utiliser par les opérateur.trice.s, avec des rapports d'erreurs faciles et rapides à analyser, et qu'il soit collaboratif et évolutif.

Afin de rendre cet outil pérenne, il a été jugé important que son développement ne soit pas figé, qu'il soit possible à tous d'ajouter des couches (afin d'intégrer de nouveaux équipements, de nouveau format de donnée ou de nouveau type d'arborescence par exemple), et l'utilisation d'un outil de gestion de version (git, svn ...) est essentiel.

- Développer un logiciel déployable sur un serveur unique avec une interface d'installation et de mise à jour
- Développer un logiciel dans un environnement collaboratif
- Le logiciel doit intégrer plusieurs niveaux de verbosité et mode de déboggage pour tracer des erreurs de fonctionnement
- Les messages inhérents au fonctionnement du logiciel sont "loggés" dans des fichiers ou dans les log système avec une gestion de l'espace disque qu'ils occupent selon des critères classiques de taille et d'ancienneté afin de pas saturer les espaces disque disponibles.

4.3 Interface graphique & Visualisations

Le développement du projet SiQaCo intégrera un outil de visualisation de l'état du traitement des discontinuités des données par réseau/station/canal. En effet, cela permettra aux utilisateur.rice.s d'analyser rapidement et facilement l'état de l'archive pour les différents réseaux/stations/canaux, pour savoir quelles sont les demandes de retransmission à effectuer, et de prendre connaissance du résultat des analyses d'erreurs et des métriques de l'archive finale.

Chapitre 5

Description du produit

5.1 Utilisateurs du produit

Ce produit sera principalement utilisé par les opérateur.rice.s dans les observatoires de l'IPGP (GEO-SCOPE et observatoires volcanologiques) dans le cadre du traitement des discontinuités et du contrôle qualité des données afin de les valider. L'outil devra permettre de traiter rapidement et de manière robuste les discontinuités, ainsi que de faire le contrôle qualité des données et de visualiser les erreurs sur les données et méta-données, tout en laissant aux utilisateur.rice.s la main sur le paramétrage de l'outil. Cet outil pourra également être utilisé par d'autres opérateur.rice.s de la communauté RESIF.

5.2 Modularité

Afin de satisfaire à l'évolution des formats de données et de méta-données, et dans un soucis d'adaptation à de nouvelles technologies (coffrets de numérisation, protocoles de communication avec les stations etc.), le logiciel sera développé de façon modulaire. Ainsi, chaque étape du traitement des données et des méta-données ainsi que de leur validation sera développé comme autant de "modules" qui seront appelés par la chaîne de traitement principale. L'accès aux données se faisant depuis différentes sources, chaque source aura un "module" associé qui permettra au logiciel d'interagir avec elle. De plus, la structure des données, que nous définirons par défaut comme "SDS", devra elle aussi pouvoir évoluer ; elle sera donc associée à un module spécifique.

Dans l'optique de développements futurs, le logiciel sera pensé de telle manière à ce qu'un nouveau module puisse être intégré facilement à la chaîne de traitement et de validation des données et des méta-données.

Le logiciel doit permettre de définir des groupements de stations qui partageront les mêmes paramètres de configuration.

Enfin, le logiciel devra fonctionner en continu, et les traitements seront automatiques. Cependant, il y aura la possibilité de rajouter des requêtes manuelles, telles que l'ajout d'une nouvelle source "offline" par exemple.

5.3 Métriques

Le logiciel SiQaCo utilisera un ensemble de métriques et d'erreurs à partir desquelles seront effectuées les vérifications sur les données et les méta-données ainsi que les décisions pour le traitement des données.

Les erreurs seront qualitatives (problèmes de compression, problème de datation par exemple), les métriques seront quantitatives (valeur moyenne du signal, nombre de trous par jours...). Les erreurs et les métriques utilisées seront paramétrables par l'utilisateur, afin que ce dernier puisse adapter le logiciel aux spécificités de ses stations et jeux de données. De plus, l'utilisateur.rice devra avoir la possibilité d'acquitter des erreurs (dans le cas où l'erreur n'en est pas une).

Une fois de plus, il ne sera pas question de développer des outils permettant de mesurer ces métriques : en effet, il existe déjà deux outils disponibles que nous utiliserons. Il s'agit des outils ISPAQ et WaveFormCatalog (*voir Références 1.3*).

Dans un premier temps, le logiciel SiQaCo utilisera les métriques suivantes :

- Valeurs moyenne, médiane, minimum et maximum du signal
- Détection de pics dans le signal
- RMS (*root mean square*)
- Nombre de trous et de recouvrements

Par la suite, on pourra également intégrer les métriques sur la symétrie, la latence de complétion (différence entre le temps où une requête est émise et le moment où elle est traitée), la vérification des spectres par rapport à un gabarit, l'inter-corrélation de différents canaux, l'orientation des canaux (lors de télésismes), la valeur du max STA/LTA... (cf. ISPAQ 1.3, l'ensemble des métriques est visualisable sur le site web).

5.4 Traitement des données

5.4.1 Fonctionnement général

Afin d'obtenir la structure finale la plus "propre" et complète possible, le logiciel va commencer par effectuer, de la manière la plus automatique possible, le traitement des discontinuités des données depuis plusieurs sources. Le logiciel devra intégrer une interface pour paramétriser les différentes sources de données d'entrées, les critères liés à la construction de la structure finale des données et enfin les caractéristiques de cette structure (nous détaillons ces métriques dans le paragraphe 5.3).

Pour cela, la première étape sera d'effectuer la surveillance des sources afin d'optimiser la construction d'une base ou d'une table de données qui permette de savoir où se trouvent les données : si elles sont disponibles en station, à l'observatoire, sur un buffer... Ce monitoring doit prendre en compte des informations émanant des utilisateur.rice.s car il semble difficile de confier au logiciel l'analyse de l'ensemble de l'environnement des sources de données qui en affecterait leurs comportements. Cette surveillance permettra également de diminuer le besoin en ressources, puisque cela permettra d'éviter d'envoyer plusieurs requêtes redondantes.

A partir de cette base de données, le logiciel, configuré par les utilisateur.rice.s, ira mettre à jour la structure finale en bouchant les trous dans les données et en traitant les recouvrements.

Enfin, à partir de l'analyse de l'état de la structure finale, le logiciel va créer et mettre à jour un système d'information qui crée et met à jour le statut de l'archive finale selon les critères et les métriques définis par l'utilisateur. Ce système d'information devra se présenter sous la forme d'un outil de visualisation (via une page web par exemple) qui permette à l'utilisateur de visualiser rapidement l'état de la structure finale (cf. 5.6).

5.4.2 Outils utilisés

Le traitement de données, dans le cas du format de données *miniseed* (cf. 3.6), s'effectuera à l'aide de plusieurs outils existants (les outils *qmerge* ou *dataselect* par exemple), il ne s'agit pas ici de développer de nouveaux outils. Cependant, comme le format *miniseed* est voué à évoluer, les outils utilisés pour ce traitement devront être facilement modifiables (cf. Modularité 5.2).

5.4.3 Traitement des discontinuités

L'analyse des discontinuités doit se faire par fichier mais également de manière globale sur l'ensemble de l'archive finale afin de ne pas manquer le traitement de certaines discontinuités qui peuvent exister sur une période de temps supérieure à la journée, en particulier au niveau du changement d'année.

Définition des sources

Nous avons vu (cf. Analyse des contraintes 3.3) que dans notre cas les données peuvent se trouver sur différentes sources. Ainsi, nous ne discriminerons pas les sources entre elles, mais l'utilisateur va avoir à définir la priorité des sources entre elles.

Chaque source sera associée à un module de récupération de données qui sera capable de communiquer spécifiquement avec elle et qui pourra éventuellement interroger l'état des données présentes sur cette source.

Traitement des trous

Selon la configuration du logiciel, qui sera effectuée par les opérateur.rice.s, le logiciel SiQaCo va venir compléter de manière optimale (cf. Gestion des traitements 5.4.4) l'archive finale à partir des différentes sources à un instant T.

S'il existe plusieurs sources, l'utilisateur leur aura défini un ordre de priorité, et le logiciel ira récupérer la donnée dans la première source (selon l'ordre des priorités) où se situe la donnée, puis s'il reste des trous dans l'archive finale, il ira lire la source suivante jusqu'à ce que le trou ait été traité au mieux.

Une fois la donnée récupérée, le logiciel fusionnera cette nouvelle donnée avec l'archive finale tout en veillant à ne pas générer de recouvrement (cf. Fusion des données 5.4.3).

Traitement des recouvrements

Dans le cas d'un recouvrement dans l'archive finale, il nous faut traiter 4 éventualités :

- En cas de *leap second*, si on voit un recouvrement d'une seconde le jour précédent ou suivant cette *leap second*, il faudra laisser la donnée.
- Selon la configuration, on vérifie si ce recouvrement n'existe pas dans la source prioritaire (cf. Définition des sources 5.4.3). Si c'est le cas, on remplace la période de temps de l'archive finale par les données de cette source prioritaire.
- Si ce recouvrement existe également dans la source prioritaire, le logiciel vérifie si ce recouvrement est deux fois la même donnée sur la même période de temps, dans ce cas il suffit de supprimer la moitié des données.
- Enfin si les deux données qui se recouvrent sont différentes sur un même intervalle de temps, on crée un trou.

Fusion des données dans l'archive finale

Le traitement des données se fera en modifiant le moins possible les blocs *miniseed* afin de ne pas introduire d'erreurs. Le logiciel ira donc insérer le ou les blocs dans l'archive finale afin de ne pas avoir à décompresser et recompresser les blocs existants, tout en veillant à ne pas introduire de recouvrement.

La structure finale pourra en outre comprendre des trous (mais pas de recouvrements).

5.4.4 Gestion des traitements

Récupération automatique des données

Selon la configuration du logiciel, qui sera effectuée par les opérateur.rice.s, le logiciel SiQaCo va venir compléter (trous) et corriger (recouvrements) automatiquement et de manière optimale (cf. Gestion des traitements 5.4.4) l'archive finale à partir des différentes sources. La fenêtre de temps à traiter et la fréquence de traitement devront être paramétrables par l'utilisateur.rice et pouvoir être différents pour chaque station (on pourra ainsi synchroniser les données d'une station A toutes les heures pour les données de la veille, et d'une station B tous les jours pour le jour courant par exemple).

Il pourra être défini une période de validité de la source au-delà de laquelle le logiciel ne pourra plus aller lui demander des données (dans le cas d'un buffer où les données y sont présentes durant un certain nombre de jours avant d'être écrasées).

Récupération manuelle des données

Si l'utilisateur.rice le souhaite, il ou elle pourra également venir "forcer" la récupération de données depuis une nouvelle source : on peut penser au cas où le logiciel SiQaCo aura effectué automatiquement le traitement des discontinuités sur un ensemble de données depuis les sources qu'il a jugé être les "meilleures" (cf. Analyse des sources 5.4.4) et construit l'archive finale, mais que l'utilisateur.rice a ensuite récupéré les données directement sur la station (carte SD, USB), et qui sont, sauf en cas de corruption des fichiers ou conversion des données à la volée (dans le cas des Taurus et Cygnus par exemple), les meilleures données disponibles possibles. A ce moment là, il devra être possible de "forcer" manuellement l'utilisation de ces données et de remplacer celles existantes (ou non) dans l'arborescence finale.

Analyse de l'archive finale

Cette fonctionnalité du logiciel ira analyser automatiquement, selon une fréquence définie par l'utilisateur, l'état de l'archive finale, à l'aide de différentes métriques (cf. Métriques 5.3), et ira mettre à jour une table d'état de l'archive finale.

Analyse des sources

Cette fonctionnalité du logiciel va analyser la disponibilité des sources et mettre à jour une table de l'état des sources .

Création des requêtes

Afin de répondre aux contraintes de ressources (cf. 3.1), le logiciel traitera les requêtes, demandées par l'utilisateur au sein d'une "pile" de traitement, afin d'éviter de réitérer une requête qui aura échoué, mais aussi de pouvoir gérer la priorité des tâches à effectuer.

Cette fonctionnalité ira donc créer les requêtes de récupération des données en fonction de l'état de l'archive finale, en fonction de l'heure courante, en fonction de l'état des sources, et en fonction des limitations en ressources locales et distantes (bande passante, CPU...). Il devra être donné la possibilité de créer manuellement une requête afin de forcer la récupération de données (cf. Récupération Manuelle 5.4.4).

De plus, afin de vérifier que l'on a bien récupéré l'ensemble des canaux d'une station (cf. Analyse des contraintes 3.2), il faudra définir a priori le nombre et le type de canaux de chaque station que le logiciel SiQaCo comparera avec les canaux reçus. Ainsi, s'il ne voit pas le fichier journalier d'un canal, alors ce module génère une nouvelle requête à traiter de la totalité de cette journée pour ce canal.

De plus ce module va analyser l'état des requêtes après traitement, afin de répéter un certain nombre de fois cette requête en cas d'échec, suspendre la requête en cas d'échecs répétitifs, supprimer la requête en cas de succès... par exemple.

Gestion de la pile

Cette fonctionnalité va analyser les différentes requêtes créées afin de les traiter de manière optimale : regrouper des requêtes similaires (demander une fois trois jours de données plutôt que trois fois une journée), exécuter les requêtes prioritaires en premier... Dans un premier temps, cette fonctionnalité va fonctionner de manière "*premier entré, premier sorti*".

Traitement des requêtes

Les requêtes seront traitées de la manière suivante : lecture des données depuis la source la plus prioritaire (cf. 5.4.3), traitement de la discontinuité (trou ou recouvrement : cf. 5.4.3 et 5.4.3), lecture de la seconde source prioritaire, traitement etc. jusqu'à ce que le traitement soit complet où que l'on ait traité toutes les sources.

Nettoyage

Cette fonctionnalité va mettre à jour l'état des requêtes après leur exécution (succès ou échec)

5.4.5 Gestion des alarmes

En cas de problème durant le rapatriement de données, s'il manque un canal par exemple ou si une station est en panne, le logiciel va envoyer un message d'alarme à l'utilisateur (via l'outil de visualisation, cf. 5.6).

Si une alarme est récurrente (dans le cas d'une station en panne), il devra pouvoir être laissé le choix à l'utilisateur de rendre silencieuse cette alarme, mais sans que cela arrête le traitement en cours.

5.5 Contrôle Qualité

La seconde fonctionnalité du logiciel SiQaCo est le contrôle de la qualité des données et des méta-données, qui est effectué une fois que les discontinuités ont été traitées. Cette partie du traitement n'ira pas modifier les données.

5.5.1 Vérification des méta-données

Dans un premier temps, le logiciel va effectuer la vérification des méta-données. Pour cela, une vérification de la mise à jour des méta-données sera effectuée, puis la vérification de l'ensemble des méta-données du réseau sélectionné.

Cette vérification générera un fichier de log contenant les erreurs possibles sur les méta-données, qui sera visualisé via un outil de visualisation (cf. 5.6). De plus, la vérification des méta-données va générer un fichier contenant la visualisation des réponses instrumentales.

De plus, les coordonnées des stations seront visualisées sur une carte afin de détecter d'éventuelles anomalies.

Par station, les vérifications seront :

- Conformité aux standards (Dataless SEED ou StationXML)
- Qu'il n'y a pas de recouvrement dans les époques des méta-données
- Que tous les canaux définis dans SiQaCo sont présents dans la méta-donnée
- Cohérence des dates dans les méta-données (canaux, stations, réseaux, commentaires)
- Vérification de la cohérence avec le nom du canal (fréquence d'échantillonnage, *dip* et *l'azimut*)

5.5.2 Validation croisée des méta-données et des séries temporelles

La seconde étape de la validation des données consiste à vérifier la cohérence des informations partagées entre les méta-données et les séries temporelles *miniseed*. De plus, tout échantillon doit avoir une méta-donnée unique.

5.5.3 Validation des séries temporelles

La troisième étape de validation est la vérification des séries temporelles. Ces vérifications sont les suivantes :

- Vérifier que les "NSLC" Network/Station/LocID/Channel y soient correctement renseignés

- Vérifier que le label de qualité soit celui attendu
- Vérifier qu'il n'y ait pas d'erreur lors de la décompression de la donnée
- Vérifier que la fréquence d'échantillonnage est homogène
- Vérifier que l'encodage est homogène et conforme à celui attendu
- Vérifier que la taille de blocs est conforme à celle attendue et homogène
- Vérifier que l'ordre des octets soit correct (en big endian)
- Vérifier les recouvrements qui sont toujours présents dans les données

5.5.4 Modification du label de qualité des données (contrôle qualité effectué)

Lorsque les données et les méta-données d'une période de temps définie ont été vérifiés, le logiciel ira modifier le label de qualité de toutes les données considérées comme validées de l'archive finale par celui indiqué dans la configuration.

5.6 Visualisation de l'état de l'archive finale

Le logiciel SiQaCo intégrera un outil de visualisation de l'état de l'archive finale et des métriques, où l'utilisateur pourra accéder rapidement au réseau/station/canal qu'il souhaite. L'utilisateur.rice pourra également consulter les logs générés par le système.

5.7 Modifications manuelles de l'archive finale

Dans certains cas, il se peut que l'utilisateur.rice veuille sélectionner une période pour une certaine station ou pour un certain canal qu'il veut supprimer. Cela peut arriver dans le cas où la vérification croisée des données avec les méta-données renvoie des erreurs (présence de données sans métadonnées par exemple), et usuellement lors d'une intervention, quelques données doivent être supprimées.

Ainsi, le logiciel SiQaCo va laisser la main à l'utilisateur.rice afin qu'il ou elle puisse :

- Supprimer un ou plusieurs fichier(s)
- Couper le début ou la fin d'un fichier afin de supprimer les données jugées non-valides
- Couper les données à l'intérieur d'un fichier (en gardant le début et la fin du fichier)

5.8 Supervision de SiQaCo

Le logiciel SiQaCo comprendra un outil de supervision sous la forme d'une interface qui permettra à l'utilisateur.rice :

- La gestion de la configuration
- La visualisation du travail en cours
- La visualisation du travail effectué (historique des requêtes)
- La visualisation des erreurs et des alertes
- La prise de contrôle manuelle sur les opérations en cours

5.9 Archive finale

L'archive finale des données se fera sous la forme d'une archive SDS avec des fichiers dont la taille des blocs, la compression et le label de qualité seront configurés par l'utilisateur.rice.

Cette archive sera persistante. Elle pourra être utilisée comme archive d'entrée pour des validations complémentaires.

Chapitre 6

Cas d'utilisation

Nous souhaitons à présent donner quelques exemples de configuration qui permettront de montrer comment l'outil s'adaptera aux différents observatoires de l'IPGP.

6.1 Premier cas d'usage

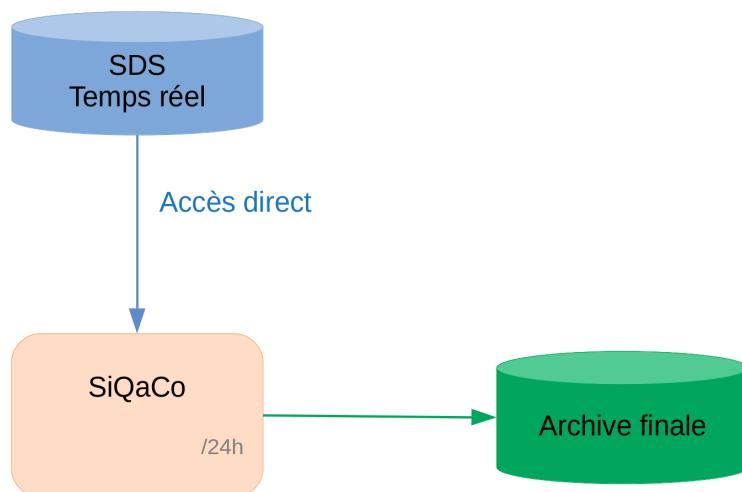


FIGURE 14 – Cas d'usage 1

Dans un premier cas d'usage, on imagine une archive finale qui n'est alimentée que par le temps réel. Le logiciel SiQaCo va analyser l'archive finale à J-1 et sera lancé toutes les 24 heures.

1. L'analyse de l'archive finale va dire qu'il y a un trou dans les données pour le jour J-1
2. Le logiciel SiQaCo va créer la ou les requêtes pour aller lire les données du jour J-1 avec *dataselect* sur le disque de données temps réel
3. Le logiciel SiQaCo ira ensuite traiter ces données : il va aller créer les données du jour J-1 dans l'archive finale
4. Dans le cas où les données ont un ou plusieurs trous, le logiciel ne va rien faire (vu qu'il n'y a qu'une source, il ne peut rien faire de mieux)
5. Dans le cas où il y a un ou plusieurs recouvrements, le logiciel va traiter ce ou ces recouvrement

6.2 Deuxième cas d'usage

Dans le cas du réseau West Indies (WI), l'archive de l'observatoire OVSM va alimenter l'archive finale de l'OVSG et vice versa, toutes les heures pour l'heure précédente (H-1). Nous nous retrouvons dans le cas de figure suivant, par exemple à l'OVSM.

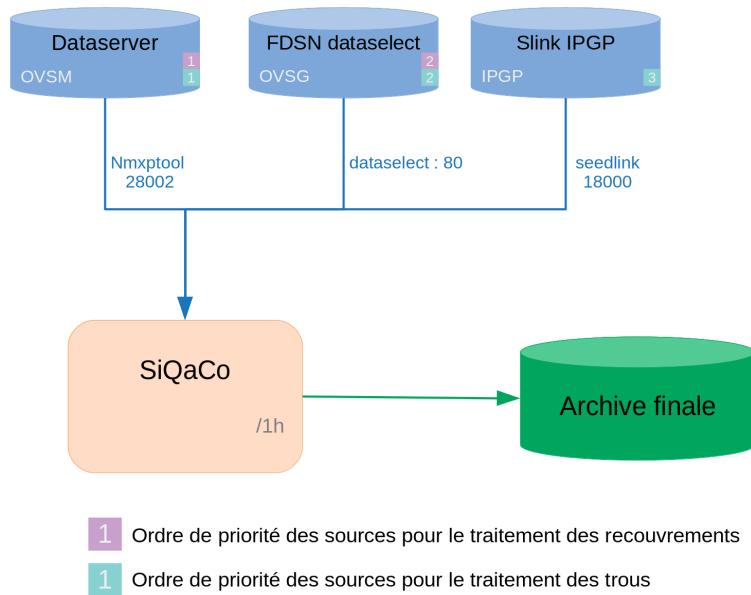


FIGURE 15 – Cas d'usage 2 : réseau WI

1. L'analyse de l'archive finale va dire qu'il y a un trou dans les données pour l'heure H-1
2. Le logiciel SiQaCo va créer la ou les requêtes pour aller lire les données de l'heure H-1 en fonction des priorités :
 - (a) via nmxptool sur dataserver
 - (b) via le Web Service FDSN dataselect sur l'archive de l'OVSG
 - (c) via une requête seedlink sur l'IPGP
3. Le logiciel SiQaCo ira ensuite traiter ces données : il va aller ajouter les données de l'heure H-1 dans l'archive finale
4. Dans le cas où les données ont un ou plusieurs trous, le logiciel va les traiter en allant lire la source suivante
5. Dans le cas où il y a un ou plusieurs recouvrements, le logiciel va traiter ce ou ces recouvrement

6.3 Troisième cas d'usage

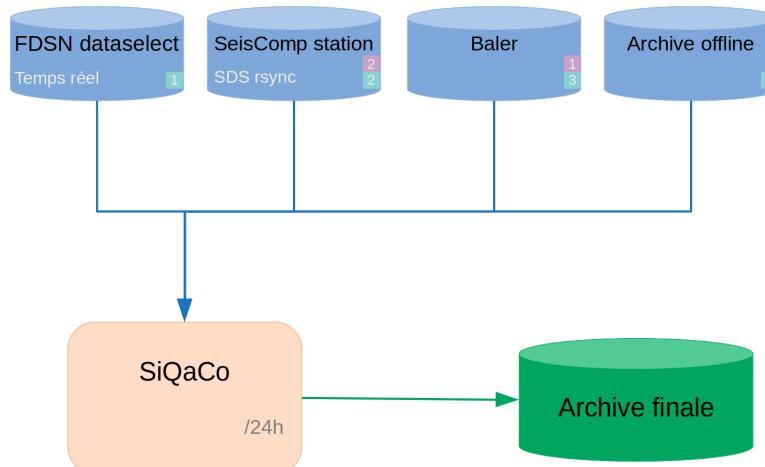


FIGURE 16 – Cas d’usage 3

Dans ce cas d’usage, le logiciel SiQaCo tourne toutes les 24 heures afin de corriger les discontinuités dans l’archive finale.

1. L’analyse de l’archive finale va dire qu’il y a un trou dans les données pour la journée J-1 et un recouvrement pour la journée J-3
2. Le logiciel SiQaCo va créer la ou les requêtes pour aller lire les données du jour J-1 pour boucher le trou, en fonction des priorités :
 - (a) via le Web Service FDSN dataselect
 - (b) via la SeisComp Box présente en station
 - (c) via une requête sur le Baler
3. Le logiciel SiQaCo va ensuite créer la ou les requêtes pour aller lire les données du jour J-3 pour corriger le recouvrement, en fonction des priorités :
 - (a) via une requête sur le Baler
 - (b) via la SeisComp Box présente en station
4. Le logiciel SiQaCo ira ensuite traiter ces données : il va aller ajouter les données pour boucher le trou de J-1 dans l’archive finale
5. Enfin, le logiciel va corriger le recouvrement pour le jour J-3

Annexe A

Glossaire

A.1 Noms

RESIF	Réseau Sismologique et Géodésique Français
IPGP	Institut de Physique du Globe de Paris
EOST	École et Observatoire des Sciences de la Terre (Strasbourg)
OVPF	Observatoire Volcanique du Piton de la Fournaise (Île de la Réunion)
OVSG	Observatoire Volcanologique et Sismologique de Guadeloupe
OVSM	Observatoire Volcanologique et Sismologique de Martinique
WI	West Indies : Réseau de l'arc des petites Antilles
FDSN	International Federation of Digital Seismograph Networks

A.2 Abréviations

SEED	Standard for the Exchange of Earthquake Data
SDS	Seiscomp Data Structure, du type " <SDSdir>/YEAR/NET/ STA/CHAN.TYPE/NET.STA.LOC.CHAN.TYPE.YEAR.JDAY"

Annexe B

Document de travail

Les parties 5 et 6 ont été rédigées en s'appuyant sur un schéma de développement que nous présentons ici. Il faut garder à l'esprit que ce schéma n'est pas figé, et que ce n'est qu'une ébauche du développement qui sera défini dans un futur dossier de conception. Voici ce schéma :

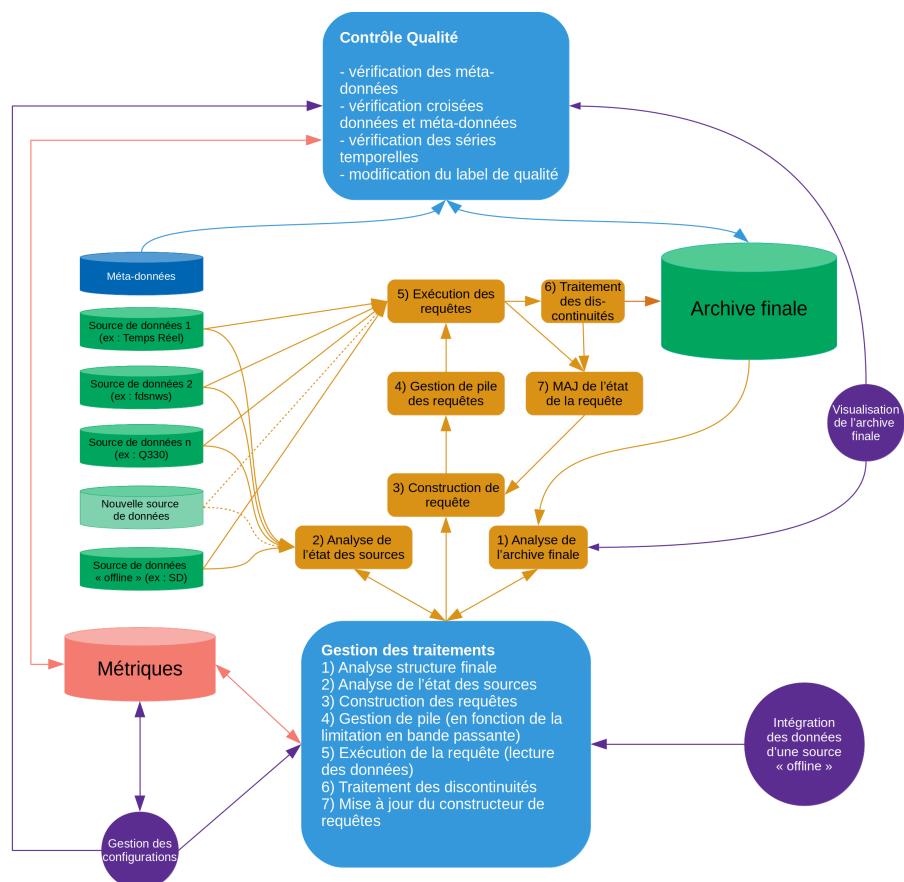


FIGURE 17 – Fonctionnalités du projet SiQaCo

Annexe C

Validations complémentaires depuis l'archive finale

Dans le cas des observatoires volcanologiques et sismologiques de l'IPGP, les données de l'archive finale seront ensuite validées par les doctorant.e.s de l'IPGP, qui procéderont à deux contrôles manuels supplémentaires.

Pour information, ces contrôles sont les suivants :

- **Comparaison avec des séismes synthétiques** Un volume SEED sera créé qui comprendra les 3 plus gros télé-séismes de l'année pour chaque réseau. Ces 3 télé-séismes seront définis en interrogeant la base de données de l'USGS pour des événements de magnitude comprise entre 6.8 et 8.2, à une distance comprise entre 40° et 85° du réseau. Ce volume contiendra une heure de données de chaque réseau pour chacun des trois télé-séismes. On utilisera ensuite la méthode SCARDEC d'inversion du tenseur de moment afin d'afficher la différence entre les données et un signal synthétique, ce qui permettra de visualiser des erreurs sur les données.
- **Contrôle qualité sur les spectres** Les données seront ajoutées à la base de données du logiciel PQLX afin d'effectuer le contrôle des spectres de chaque canal de chaque station.