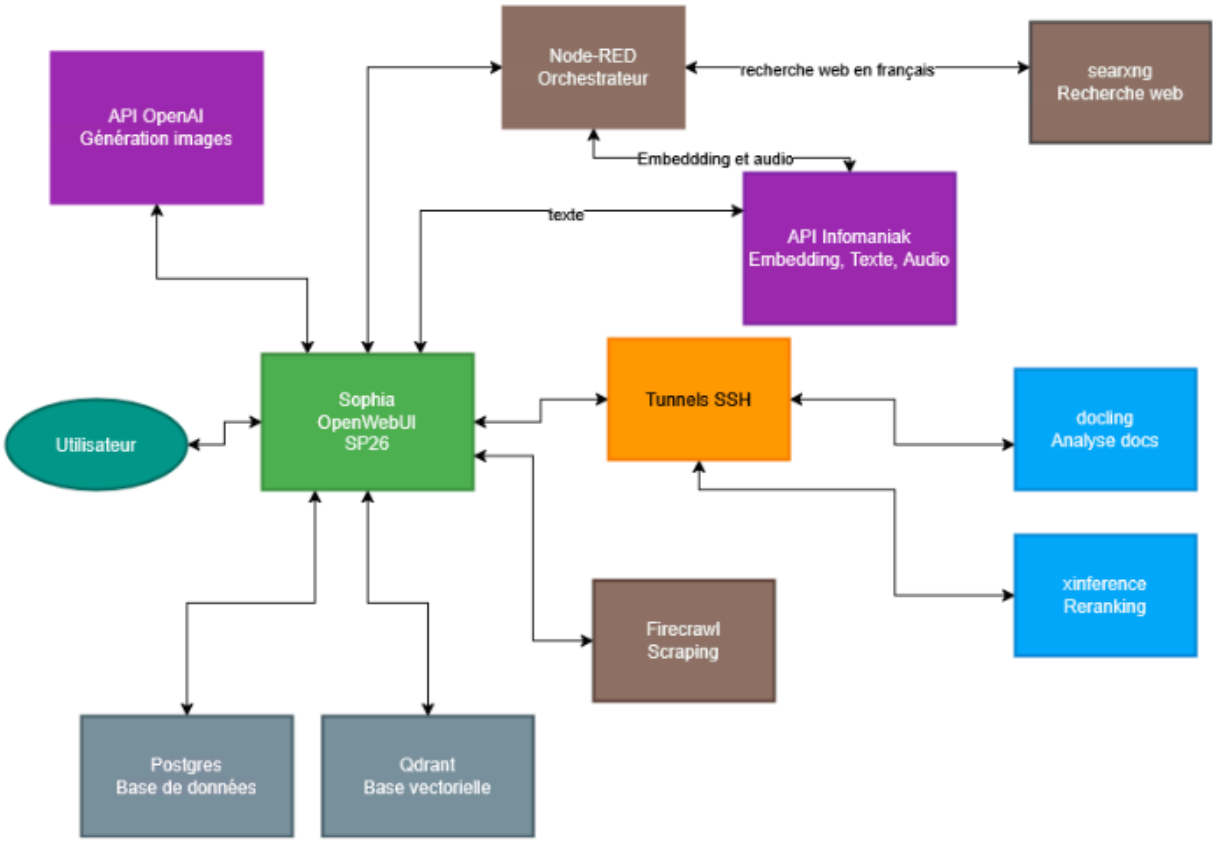
Sophia

Architecture générale

Schéma d'architecture générale de Sophia



Pour modifier l'image : https://onlyoffice.vincennes.fr/Products/Files/DocEditor.aspx?fileid=40912

L'architecture Today at 19:45

L'architecture est basée sur la documentation d'openwebui : https://docs.openwebui.com/.

Tableau récapitulatif des conteneurs - visualisation par Portainer

Conteneur / Service	Stack / Origine	Rôle	Port hôte (local)	Port conteneur	URL d’accès depuis Sophia
`openwebui`	Sophia	Interface principale	3002	8080	`https://sophia.vincennes.fr`
`qdrant`	Sophia	Base vectorielle (RAG)	6333, 6334	6333, 6334	Interne (via API locale)
`postgres`	Sophia	Base de données d’application	5437	5432	Interne
`sophia-nodered-openwebui-1`	Sophia	Automatisation, intégration des services	1886	1880	`http://sp26.vincennes.fr:1886`
`infomaniak-xinference-1`	Infomaniak (tunnel SSH)	Tunnel vers xinference (reranking)	9995	9998	`http://sp26.vincennes.fr:9995`
`infomaniak-docling-1`	Infomaniak (tunnel SSH)	Tunnel vers docling (analyse de documents)	5002	5001	`http://sp26.vincennes.fr:5002`
`infomaniak-portainer_agent-1`	Infomaniak (tunnel SSH)	Agent Portainer pour supervision distante	9002	9001	Interne (gestion Docker)
`searxng`	ia-app	Moteur de recherche web	8056	8080	Via Node-RED (recherche)
`firecrawl-api-1`	firecrawl	Analyse de pages web (scraping intelligent)	3003	3002	`http://sp26.vincennes.fr:3003`
`docling` (VM Infomaniak)	-	Analyse de documents (PDF, etc.)	5001	5001	Accès distant (via tunnel SSH depuis SP26:5002)
`xinference` (VM Infomaniak)	-	Modèle de reranking (`bge-reranker-v2-m3`)	9997	9997	Accès distant (via tunnel SSH depuis SP26:9995)
`portainer_agent` (VM Infomaniak)	-	Agent Docker pour supervision	9001	9001	Accès distant (via tunnel SSH depuis SP26:9002)

Réglages dans l’interface administrateur - visualisation par Openwebui

1. Menu Connexions

- **URL** : `https://sophia.vincennes.fr/admin/settings/connections`
- **Fonction** : Gestion des connexions aux modèles IA hébergés sur Infomaniak.
- **Détails** :
  - Affiche l'état de disponibilité des endpoints IA (texte, audio, image).
  - Aucune action de configuration possible directement : en cas de problème, contact avec le support requis.
  - Vérification de la connectivité aux API distantes.

2. Menu Documents

- **URL** : `https://sophia.vincennes.fr/admin/settings/documents`
- **Fonction** : Configuration du traitement des documents (analyse, embedding, reranking).
- **Paramètres** :
  - **Analyse de documents** :
    → URL : `http://sp26.vincennes.fr:5002/`
    → Connecté au service **docling** via tunnel SSH.
  - **Embedding** :
    → URL : `http://sp26.vincennes.fr:1886/`
    → Passerelle via **Node-RED**, qui relaye vers le modèle `bge_multilingual_gemma2` sur Infomaniak.
    → Utilisation de Node-RED pour limiter et réguler les appels API.
  - **Reranking (RAG)** :
    → URL : `http://sp26.vincennes.fr:9995/v1/rerank`
    → Connecté au modèle `bge-reranker-v2-m3` via **xinference** sur Infomaniak (tunnel SSH).

3. Menu Recherche web

- **URL** : `https://sophia.vincennes.fr/admin/settings/web`
- **Fonction** : Configuration de la recherche web et de l'analyse des résultats.
- **Paramètres** :
  - **Recherche** :
    → URL : `http://sp26.vincennes.fr:1886/search?q=<query>`
    → Passerelle via **Node-RED**, qui interroge **searxng**.
    → Node-RED assure la traduction de la requête en français si nécessaire.
  - **Analyse des pages** :
    → URL : `http://sp26.vincennes.fr:3003/`
    → Envoi des URLs récupérées vers **Firecrawl** pour extraction de contenu structuré.

4. Menu Interface utilisateur

- **Fonction** : Génération automatique de métadonnées pour améliorer l'expérience utilisateur.
- **Utilisation** :
  - Appels aux **API Infomaniak** pour :
    - Génération de **titres** pertinents.
    - Création de **tags** et de **mots-clés** pour la recherche.
  - Ces fonctionnalités enrichissent les réponses et les contenus générés.

5. Menu Audio

- **URL** : `https://sophia.vincennes.fr/admin/settings/audio`
- **Fonction** : Transcription de fichiers audio en texte.
- **Paramètres** :
  - **Transcription** :
    → URL : `http://sp26.vincennes.fr:1886/`
    → Utilisation de **Node-RED** pour gérer l'appel à l'API Infomaniak.
    → Le processus est asynchrone : Node-RED surveille la tâche et renvoie le texte une fois disponible.
  - Permet de traiter des fichiers audio longs via une file d'attente.

6. Menu Images

- **Fonction** : Génération d'images à partir de descriptions textuelles.
- **Paramètres** :
  - Appel direct aux **endpoints OpenAI** (ex: DALL·E).
  - Aucun traitement local ou intermédiaire : la requête est envoyée directement à OpenAI.
  - Les images générées sont intégrées dans les réponses de l'interface.

Vérification du bon fonctionnement

Voici une **checklist de vérification du bon fonctionnement** du système, suivie d'un **tableau récapitulatif des problèmes possibles en cas d'indisponibilité**.

Checklist de vérification du bon fonctionnement

Élément à vérifier	Procédure de vérification	Statut (✓/✗)
1. Containers dans Portainer	Accéder à l'interface Portainer et s'assurer que tous les conteneurs sont en état **"Running"** : - `openwebui`, `qdrant`, `postgres`, `sophia-nodered-openwebui-1` - `infomaniak-xinference-1`, `infomaniak-docling-1`, `infomaniak-portainer_agent-1` - `searxng`, `firecrawl-api-1 (et les containers de la stack firecrawl)`	
2. Interface Sophia (OpenWebUI)	Accéder à : `https://sophia.vincennes.fr` → Vérifier que la page charge correctement et que l'interface est interactive.	
3. Interface Node-RED	Accéder à : `http://sp26.vincennes.fr:1886` → Vérifier que l'éditeur s'ouvre et que les flux sont actifs.	
4. Firecrawl (API et message par défaut)	Accéder à : `http://sp26.vincennes.fr:3003` → Vérifier que la réponse "SCRAPERS-JS: Hello, world! K8s! "s'affiche.	

Élément à vérifier	Procédure de vérification	Statut (✓/✗)
5. Interface Docling	Accéder à : `http://sp26.vincennes.fr:5002/ui` → Vérifier que l'interface utilisateur de Docling est accessible.	
6. XInference et modèle de reranking	1. Accéder à l'interface XInference via le tunnel : `http://sp26.vincennes.fr:9995` 2. Vérifier que le modèle `bge-reranker-v2-m3` est **chargé et en cours d'exécution** dans la section `/reranker`.	
7. SearXNG	Accéder à : `http://sp26.vincennes.fr:8056` → Vérifier que l'interface de recherche s'affiche. → Effectuer une recherche test (ex: "test").	
8. Endpoint OpenAI	Depuis l'interface Sophia (Menu Images) : → Lancer une génération d'image. → Vérifier que la requête atteint OpenAI et retourne une image.	
9. Endpoints Infomaniak (via Node-RED)	Vérifier dans les menus Sophia : - **Connexions** : état des endpoints texte/audio/image - **Documents**, **Audio**, **Recherche** : que les appels passent via `http://sp26.vincennes.fr:1886` → Vérifier les logs Node-RED en cas d'erreur.	
10. Logs des conteneurs de tunnel SSH	Vérifier les logs des conteneurs suivants dans Portainer : - `infomaniak-xinference-1` - `infomaniak-docling-1` - `infomaniak-portainer_agent-1` → Rechercher des erreurs de connexion SSH, timeout ou refus. → S'assurer que la VM Infomaniak est joignable (ping ou test de connexion réseau).	

⚠ Tableau des problèmes possibles en cas d'indisponibilité

Service indisponible	Symptôme observé	Cause probable	Impact fonctionnel
Conteneur non démarré dans Portainer	Interface inaccessible, erreur 502/503	Conteneur crashé, mauvaise configuration, dépendance manquante	Blocage total du service associé
Sophia (OpenWebUI)	Page blanche, erreur de chargement	Problème réseau, conteneur down, certificat SSL expiré	Interface principale inutilisable
Node-RED	Éditeur inaccessible ou flux inactifs	Conteneur down, erreur dans un flux critique	Toutes les intégrations (recherche, audio, embedding) bloquées
Firecrawl	API ne répond pas ou timeout	Conteneur down, VM surchargée, erreur de scraping	Impossible d'extraire du contenu web pour les recherches
Docling	`/ui` inaccessible ou erreur 500	Tunnel SSH rompu, service Docling arrêté sur Infomaniak	Analyse des documents PDF/HTML impossible
XInference / bge-reranker-v2-m3	Modèle non chargé ou erreur 404/500	Modèle non démarré, mémoire insuffisante, tunnel SSH défaillant	Détérioration de la qualité des réponses (RAG)
SearXNG	Page de recherche inaccessible	Conteneur down ou configuration réseau incorrecte	Recherche web impossible via Sophia
Endpoint OpenAI	Génération d'image échoue	Clé API invalide, réseau bloqué, quota dépassé	Fonctionnalité d'image inactive
Endpoints Infomaniak	Erreurs dans les menus Sophia (Connexions, Audio, etc.)	Tunnel SSH rompu, VM éteinte, problème d'authentification	Toutes les fonctions IA distantes (transcription, embedding) inactives
Logs des tunnels SSH	Messages répétés : `Connection refused`, `Broken pipe`, `Failed to connect`	VM Infomaniak éteinte, firewall, clé SSH expirée, réseau instable	Perte de contrôle et de données vers les services distants

---

> **Recommandation** : Cette checklist peut être utilisée quotidiennement ou après chaque déploiement/mise à jour. En cas d'indisponibilité, consulter en priorité les **logs des conteneurs de tunnel SSH** et vérifier la **connectivité réseau vers la VM Infomaniak**.

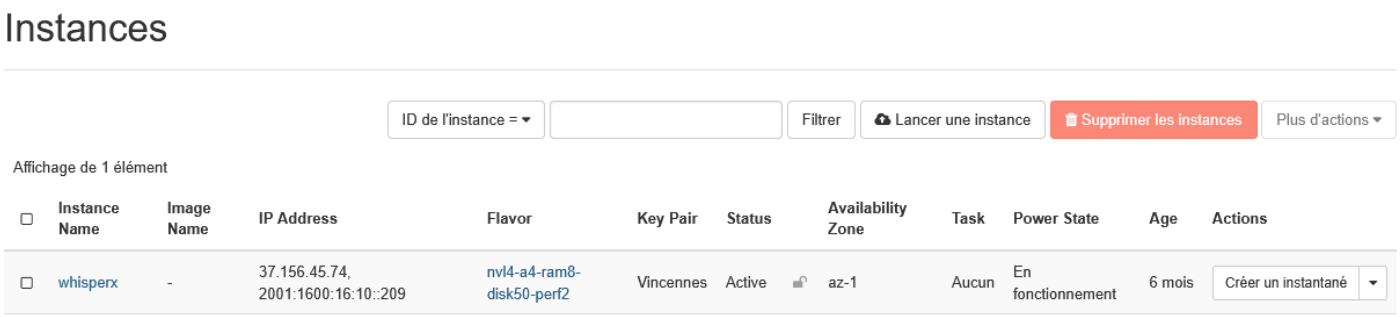
Redémarrage de la VM GPU sur Infomaniak

Aller sur https://api.pub2.infomaniak.cloud/horizon/project/

Les identifiants sont dans psono "Infomaniak openstack / Horizon"

Puis sur la page https://api.pub2.infomaniak.cloud/horizon/project/instances/

Une seule instance est configurée : "whisperx".



Dans le cas où la VM présentera un problème de blocage persistant, elle peut être redémarrée via l'option disponible en bout de ligne : "Redémarrer matériellement l'instance".

Le redémarrage prend 2 à 3 minutes.

Une fois relancée, les containers réapparaissent fonctionnels sur Portainer / Infomaniak.

Les accès à docling et xinference doivent être vérifiés (voir points 5 et 6 de https://formagent.vincennes.fr/books/intelligence-artificielle/page/verification-du-bon-fonctionnement).