



Description des sources de données et modélisation de la wikibase

AlpesTransport

Mohammed ROUABAH
Ilyes ZEGHDALLOU
William MAILLARD

08/02/2023

1 Suivre de projet

1.1 Planning

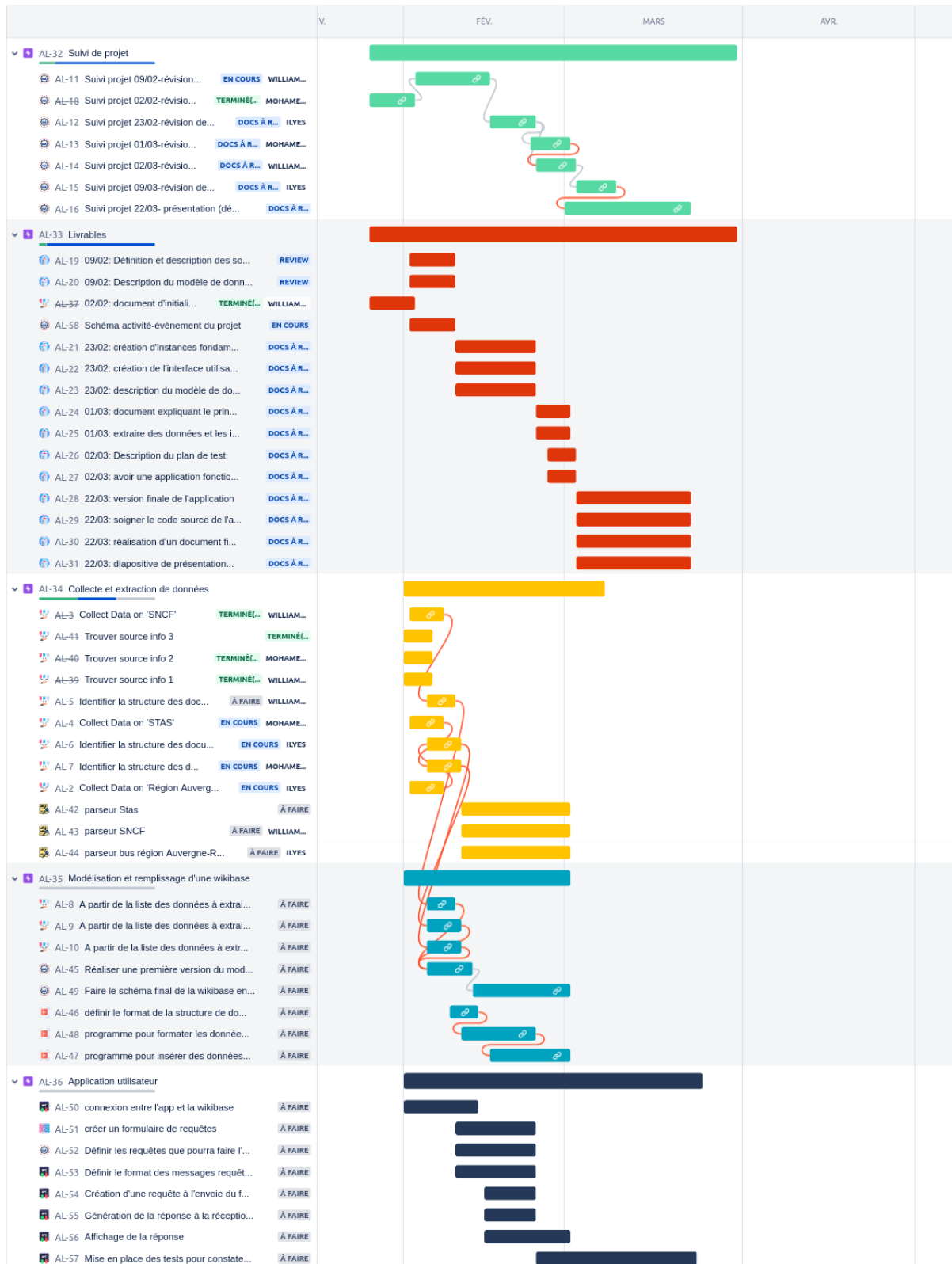


FIGURE 1 – Planning

1.2 Objectifs de la semaine

En accords avec les retours de la semaine dernière nous avons réalisé un nouveau planning en subdivisant au maximum les tâches à réaliser. Pour cela nous avons changer d’outil de gestion de projet, ainsi nous utilisons maintenant **jira** pour maintenir un *taskboard* et créer un **diagramme de gant** du planning.

Le planning divise les tâches à effectuer en 3 parties (collecte, modélisation, application) et affiche les dates des livrables du projet.

Cette semaine nous avons collecté les données de nos sources et analysé leurs structures afin de pouvoir réaliser des *parseurs* au cours de la semaine, pour extraire les données.

Nous avons aussi réalisé une première version du modèle de notre wikibase afin de visualiser la structure à créer lors de l’insertion des données extraites, qui sera réalisé dans la semaine.

2 Définition et description des sources de données

2.1 STAS

La collecte des données concernant les réseaux de transports en communs de Saint-Etienne métropole nous nous sommes orientés sur la plate-forme de stockage des données gouvernementales nommée transport.data.gouv.fr.

Ainsi nous avons pu récupérer des fichiers de données au format *.xml* contenant les informations dont nous avons besoin sur la STAS.

Les documents *.xml* sont regroupés au sein d’une archive ZIP composé de 4 fichiers et d’une archive supplémentaire contenant les fichier *.xml* d’itinéraire de chaque lignes.

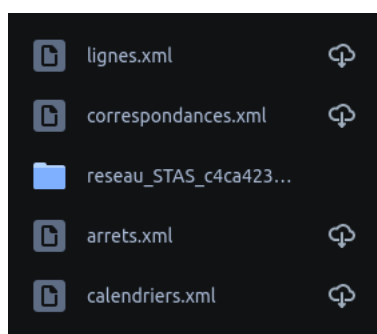


FIGURE 2 – Décomposition de l’archive

Chaque fichier *.xml* est composé d’une balise `<dataObject>` et d’une `<frame>`, les données étudiées sont structurées autour de ce balisage.

Prenons l'exemple du fichier "lignes.xml" :

```
<Line id="FR:Line:10:" version="any">
  <Name>PLACE JEAN JAURES &lt;&gt; BOURG</Name>
  <TransportMode>bus</TransportMode>
  <PublicCode>10</PublicCode>
</Line>
```

FIGURE 3 – Représentation d'une ligne de bus STAS

- les balises **<lines>** et **<line>** sont utilisées pour délimiter les caractéristiques d'une ligne, ici une ligne de bus.
- les sous-balises de **<line>** définissent ces caractéristiques, tel que :
 - **<Name>** qui représente le nom de la ligne en question
 - **<TransportMode>** qui représente le mode de transport
 - **<PublicCode>** qui indique le code de ligne public

Nous pouvons donc nous demander quel est le lien entre ces documents .xml ?

Ce sont des identifiant unique contenus dans chaque fichiers, associé à une ligne ou un arrêt par exemple, qui permet de faire le lien entre chaque contenu de chaque fichier.

Pour comprendre ces liens, prenons l'exemple de l'itinéraire d'un bus qui comprend :

- un identifiant **<fr :Route :48>**
- des caractéristiques tel que :
 - un nom : **<Name>**
 - une distance : **<Dist>**
 - une ligne **<LineRef>** (qui fait référence a l'identifiant unique du fichier lignes.xml)
- une séquence de points **<pointsInSequence>**, du fichier arrêts.xml , qui contient des référence aux points d'arrêt **<RoutePointRef>** entre le point de départ et le point d'arrivée.

```
<Route id="FR:Route:48:" version="any">
  <Name>MAIRIE CHAGNON - CH. DE GAULLE</Name>
  <Distance>0</Distance>
  <LineRef ref="FR:Line:48:">
  </LineRef>
  <DirectionType>inbound</DirectionType>
  <pointsInSequence>
    <PointOnRoute id="FR:PointOnRoute:48_1:" order="1" version="any">
      <RoutePointRef ref="FR:RoutePoint:48_1:">
      </RoutePointRef>
    </PointOnRoute>
    <PointOnRoute id="FR:PointOnRoute:48_2:" order="2" version="any">
      <RoutePointRef ref="FR:RoutePoint:48_2:">
      </RoutePointRef>
    </PointOnRoute>
  </pointsInSequence>
</Route>
```

FIGURE 4 – Exemple de route

2.2 SNCF

Les données relatives à la SNCF seront collectées sur leur page répertoriant leur "opendata" que l'on peut trouver [ici](#).

Parmi ces documents, nous utiliserons les suivants :

1. un référentiel des **gares d'Auvergne** et un des **gares Rhône-Alpes** au format `.csv` , qui nous permettra de récolter les méta-données sur les gares de la région tel que leurs code d'identification (UIC), leurs plate-formes et leurs localisation.
2. un document des **horaires des lignes inter cité** et un document des **horaires des ter** du réseau SNCF, tout deux au format `GTFS` . Pour ne collecter que ceux de la région, on filtrera ces données à l'aide des gares collecté avec les documents du point 1.
3. un document des **tarifs des ter** et un document des **tarifs des tgv inoui-ouigo**, tout deux au format `csv`. Nous filtrerons les données de ces documents comme au point 2 pour ne conserver que les tarifs concernant la région.

Les fichiers au format `.csv` contient une ligne d'en-tête qui indique se que contient chaque colonne, qui sera utilisée afin d'extraire les colonnes contenant des informations qui nous intéressent. Chaque colonne contient soit un chiffre soit une chaine de caractères.

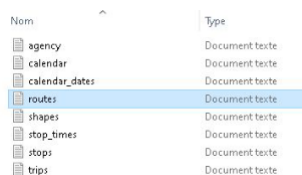
Les fichiers au format `GTFS` sont composés de plusieurs fichiers `.txt` structuré de la même manière que les fichiers `.csv` , mais il contient des références (sorte de clés étrangères) qui relient les informations des différents fichiers entre elles.

2.3 Région Rhône-Alpes

Le site de la région (disponible [ici](#)) regroupe des informations sur les bus permettant de se déplacer au sein de la région Auvergne Rhône-Alpes.

Les jeux de données disponibles nous renseignent notamment sur la position des arrêts, les parcours et les horaires de passage des véhicules.

Deux standards de données peuvent être utilisés : le format `GTFS` ou le format `NeTEx`. Nous avons choisi de prendre le format `GTFS` qui peut se composent de plusieurs fichiers `.txt` listés ci-dessous.



Nom	Type
agency	Document texte
calendar	Document texte
calendar_dates	Document texte
routes	Document texte
shapes	Document texte
stop_times	Document texte
stops	Document texte
trips	Document texte

FIGURE 5 – Fichiers `.txt` contenus dans le format `GTFS`

Voici les informations qui seront considérées lors de l'extraction de données sur les bus de la région AUvergne Rhône-Alpes :

- les lignes : nom, numéros, départ, destination,
- les horaires : départ et arrivé à u arrêt,
- les arrêts : nom, position, ligne qui passe par là,
- l'accessibilité pour les personnes à mobilité réduite,
- les bagages autorisés, notamment les vélos

3 Wikibase : modélisation triple store

La modélisation d'un schéma de données pour notre wikibase a été réalisé en suivant le modèle **Sujet<Prédicat>Objet**, en utilisant 3 éléments de base : **Entités**, **Propriétés**, **Littéraux**. Ces éléments sont identifié de différentes couleurs sur le schéma.

Afin de réaliser ce schéma nous avons donc dans un premier temps étudié les différents documents dont nous disposons et identifié les entités et littéraux apparaissant dans ces documents.

Ensuite, nous avons identifié les relations entre ces différents éléments, ce qui nous a permis de créer un graphe pour représenter la structure des données qui seront stockées dans notre wikibase.

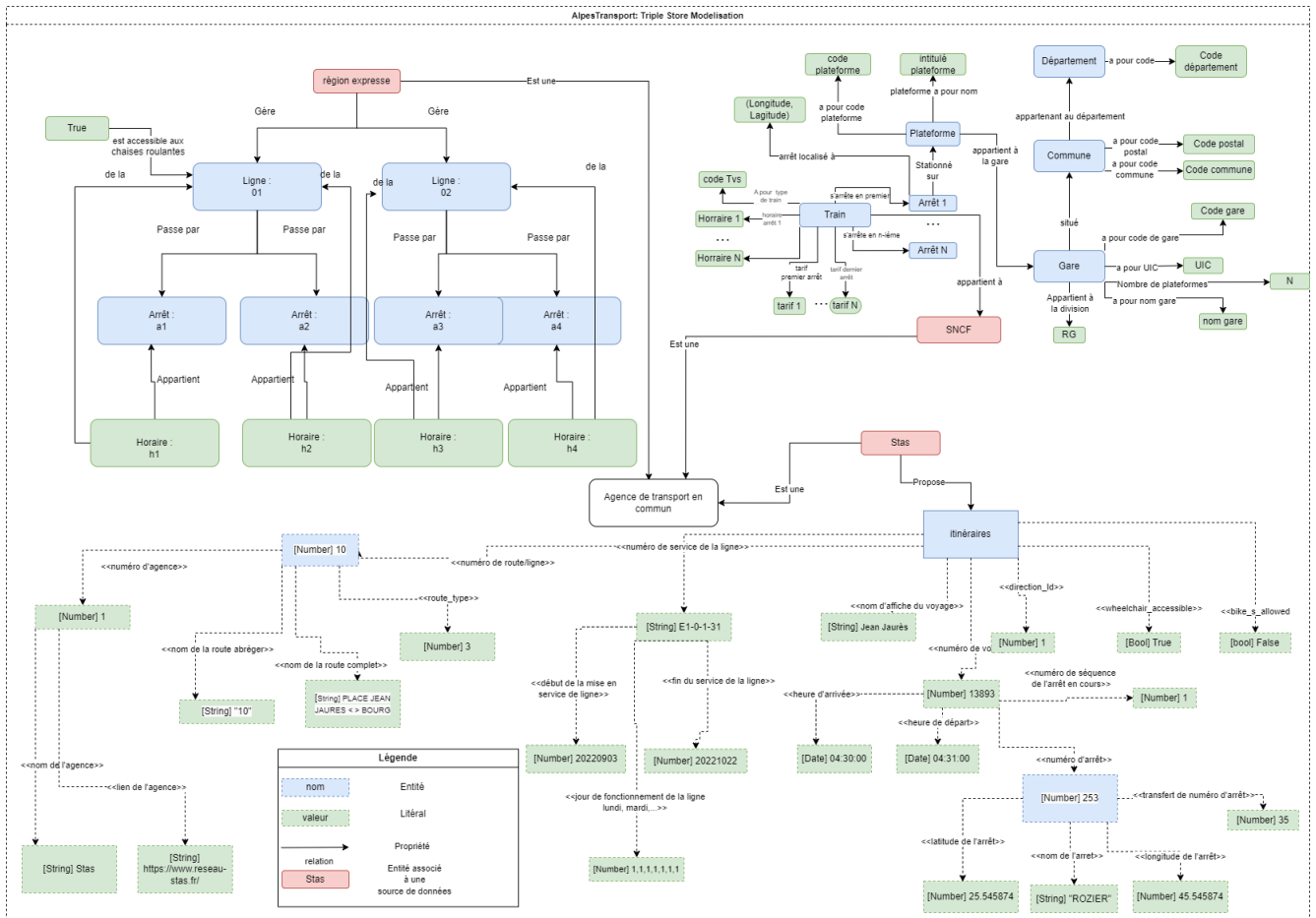


FIGURE 6 – Représentation de la modélisation des données dans la wikibase