ÉTAT DE L'ART DES RÉFÉRENTIELS GÉO-HISTORIQUES SÉMANTISÉS POUR LES HUMANITÉS

Marion Brunet et Guillaume Guébin

Carmen Brando, Nathalie Abadie, Vincent Jolivet

WEB DE DONNÉES

Les gazetiers

Un gazetier est un tableau à deux colonnes associant un nom à ses coordonnées.
Il a alors deux buts :

- Transformer des noms de lieux en localisations géographiques
- Apporter une description de ces lieux.

Trois autres prérequis sont :

- Apporter les dates de début et de fin pendant lesquelles la ressource existe.
- Si le nom apparait dans une carte, la date de cette carte doit être notée.
- Si le nom apparait dans un texte, la date de ce texte doit être notée

La plupart des gazetiers se concentrent sur la dimension spatiale des données plutôt que sur leur dimension culturelle. Il faut alors les enrichir.

Le web sémantique

Le web de données ou Linked Data vise à favoriser la publication de données structurées sur le web, en les reliant entre elles pour constituer un réseau global d'informations.

Il repose sur quatre grands principes :

- Nommer les ressources avec des URI (uniform ressouce identifier) qui permettent d'identifier sur le web ce qui existe par ailleurs.
- Utiliser des URI http pour pouvoir accéder à des informations sur les ressources.
- Lorsque l'on regarde une URI, renvoyer des in formations utiles grâce à RDF et SPARQL.
- Se relier avec d'autres URI pour créer un réseau de liens.

Le langage RDF

Pour structurer ces données, on utilise le langage RDF (ressource description framework). C'est un modèle de triplet (sujet, prédicat, objet). On découpe chaque énoncé dans un langage élémentaire.
Chaque triplet peut être vu comme un arc d'un graphe. L'utilisation des URI permet de lier entre eux les graphes et de construire un web mondial de données liées.

En RDF, les ressources appartiennent généralement à des classes qui les regroupent par types (documents, concepts, personnes, etc.). Un vocabulaire (ou ontologie) est défini comme une description formelle explicite des concepts dans ces classes, Il contient des définitions lisibles en machine des concepts de base de ce domaine et de leurs relations.

ÉTAT DE L'ART

Notre état de l'art s'est concentré sur quatre gazetiers : Pleiades, Data bnf, Geonames et Getty thesaurus of geographic names. Nous avons étudié à la fois leurs ontologies, mais aussi les données décrites dans les gazetiers.

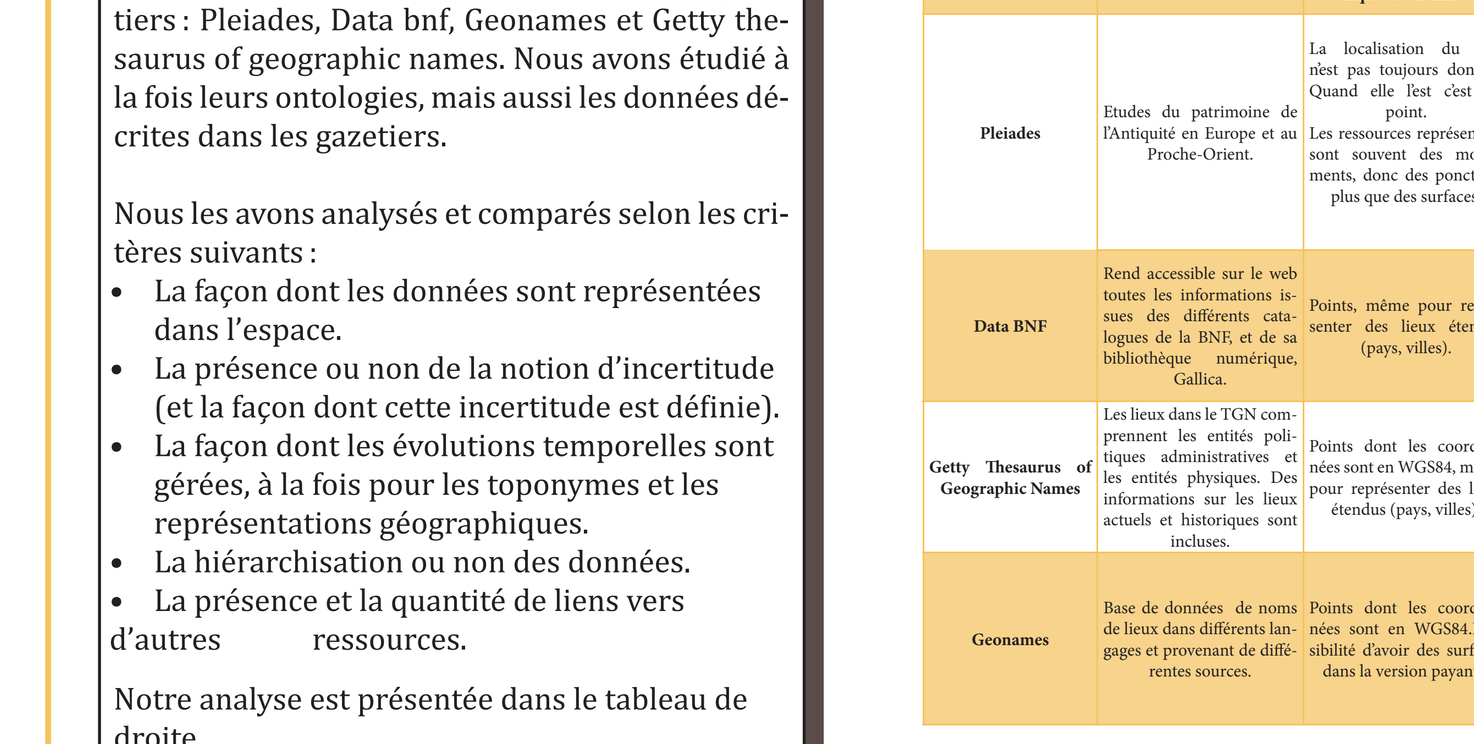
Nous les avons analysés et comparés selon les critères suivants :

- La façon dont les données sont représentées dans l'espace.
- La présence ou non de la notion d'incertitude (et la façon dont cette incertitude est définie).
- La façon dont les évolutions temporelles sont gérées, à la fois pour les toponymes et les représentations géographiques.
- La hiérarchisation ou non des données.
- La présence et la quantité de liens vers d'autres ressources.

Notre analyse est présentée dans le tableau de droite.

Nom du gazetier	Courte description	Représentation géométrique des données	Gestion de la donnée temporelle	Incertitude spatiale et temporelle	Hiérarchisation des ressources	Lien vers d'autres ressources
Pleiades	Etudes du patrimoine de l'Antiquité en Europe et au Proche-Orient.	La localisation du lieu n'est pas toujours donnée. Quand elle l'est c'est un point. Les ressources représentées sont souvent des monuments, donc des ponctuels plus que des surfaces.	Répertorie les différents noms qui ont été attribués au lieu au cours du temps, ainsi que les périodes temporelles pendant lesquelles ces noms ont été utilisés. Pas d'informations sur l'évolution temporelle dans l'espace.	Les noms ont des degrés de certitude (certain, confident), mais pas d'informations sur la façon dont ils sont attribués. Pour les localisations géographiques, lorsqu'elles sont incertaines, elles sont symbolisées par un rectangle bleu plutôt qu'un point.	Hiérarchisation en fonction du type de lieu.	Très bien relié à l'intérieur de Pléiades, moins de références à d'autres gazetiers. Références à des textes attestant de l'existence du lieu à une date ancienne, ainsi qu'à des sites ayant pour sujet la ressource.
Data BNF	Rend accessible sur le web toutes les informations issues des différents catalogues de la BNF, et de sa bibliothèque numérique, Gallica.	Points, même pour représenter des lieux étendus (pays, villes).	Les différents noms portés par un lieu sont listés, mais aucune date ne leur est associée.	Aucun attribut	Pas de hiérarchie entre les données.	Beaucoup de références de documents qui citent la ressource à différentes dates (cartes, textes, images). Lien vers d'autres sections du gazetier BNF (auteurs, thèmes) qui ont un rapport avec la ressource.
Getty Thesaurus of Geographic Names	Les lieux dans le TGN comprennent les entités politiques administratives et les entités physiques. Des informations sur les lieux actuels et historiques sont incluses.	Points dont les coordonnées sont en WGS84, même pour représenter des lieux étendus (pays, villes).	Présence de noms anciens avec des attributs prévus pour préciser les dates estimées de début est de fin d'utilisation. Cependant, ces attributs sont peu utilisés.	Aucun attribut	Hiérarchisation administrative, politique et physique Il peut y en avoir plusieurs grâce à l'existence de parents préférés.	Lien vers les données spatialement proches du TGN grâce à l'attribut narrower, Pas d'alignement vers d'autres jeux de données.
Geonames	Base de données de noms de lieux dans différents langages et provenant de différentes sources.	Points dont les coordonnées sont en WGS84.Possibilité d'avoir des surfaces dans la version payante.	Présence de noms anciens mais aucune date associée.	Aucun attribut	Hiérarchie administrative, à plusieurs niveaux seulement.	Liens vers les ressources proches géographiquement avec l'attribut nearby. Lien vers l'article Wikipédia avec l'attribut gn:wikipediaArticle, Lien vers la donnée de dpbedia dans un seeAlso.

PROPOSITION D'UN MODÈLE



A gauche se trouve le digramme UML de l'ontologie geotir que nous avons créée. Nous nous sommes inspirés du modèle Pleiades qui nous semblait le plus complet, et nous l'avons enrichi.
Les dimensions qui nous semblaient manquer aux modèles précedemment étudiés et que nous avons ajoutées sont :

- L'ajout de surfaces avec la classe *geo:Geometry*,
- La gestion des incertitudes temporelles avec la classe *geotir:Interval*
- L'utilisation de la propriété *pleiades:during* à la fois sur les classes *pleiades:Name* et *pleiades:Location* pour pouvoir gérer les évolutions temporelles à la fois des noms et des formes physiques
- L'ajout de relations topologiques et de relations «floues» pour situer les lieux entre eux.

Conclusion

Après avoir testé notre modèle sur des jeux de données de GeoHistoricalData et du DicoTopo, la plupart des données est bien décrite. Il faut maintenant tester le modèle à plus grande échelle.

BIBLIOGRAPHIE

- "Place, Period, and Setting for linked Data Gazetteers" Karl Grossner, Krzysztof Janowicz, and Carsten Keßler.
- *Placing Names : Enriching and Integrating Gazetteers* Ruth Mostern