SODUCO ANR-18-CE38-0013

Création d'un graphe de connaissances géohistorique à partir des annuaires du commerce parisien du 19ème siècle: application aux métiers de la photographie

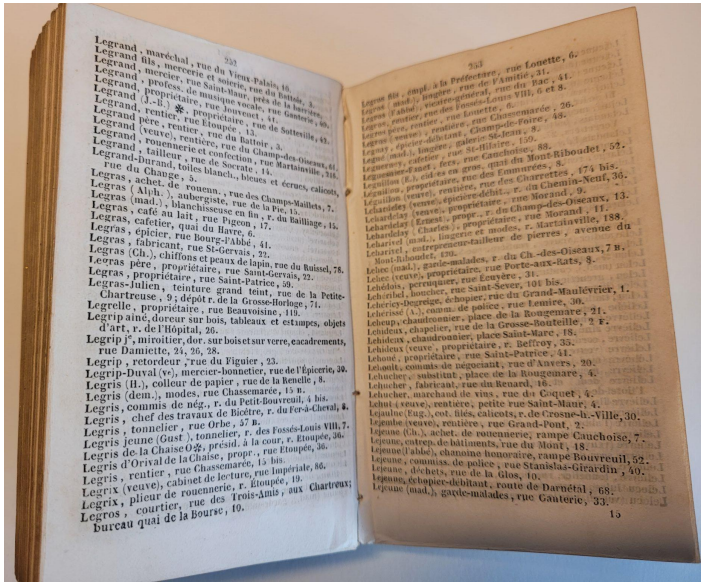
Solenn Tual [1], Nathalie Abadie [1], Bertrand Duménieu [2], Joseph Chazalon [3], Edwin Carlinet [3]

SemWebPro, Paris, 15 novembre 2023

(1) (2) (3)

Annuaires du commerce de Paris

- Édités à partir de la fin du XVIIIème siècle jusqu'au milieu du XXème siècle par différents éditeurs.
- Organisés en différentes listes (alphabétique, par profession, par adresse).
- Chaque entrée contient :
  - Nom/Raison sociale
  - Activité/Profession/Statut
  - Adresse(s)
  - Distinctions militaires/professionnelles

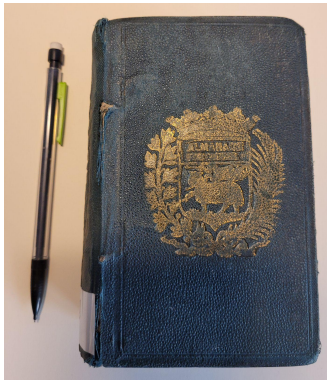


Cherot, *joaillier*. S. Martin, 51.
Chéroux, *orf*. Ste Avoye, 42.
Cherre, *layet*. Caire, 7.
Cherré, *tap*. Moulins, 12.
Cherrier, *tabl*. St Denis, 277.

Chevalier, *opt. du Roi*, Tour de l'Horl. du Palais, 1.
Chevalier (L.), *opt*. q. Horl. Pal. 65.
Chevalier (Vinc.), *opt*. q. Horl. Pal. 69.
Chevalier (Victor), *opt*. q. Horl. Pal. 77 b.

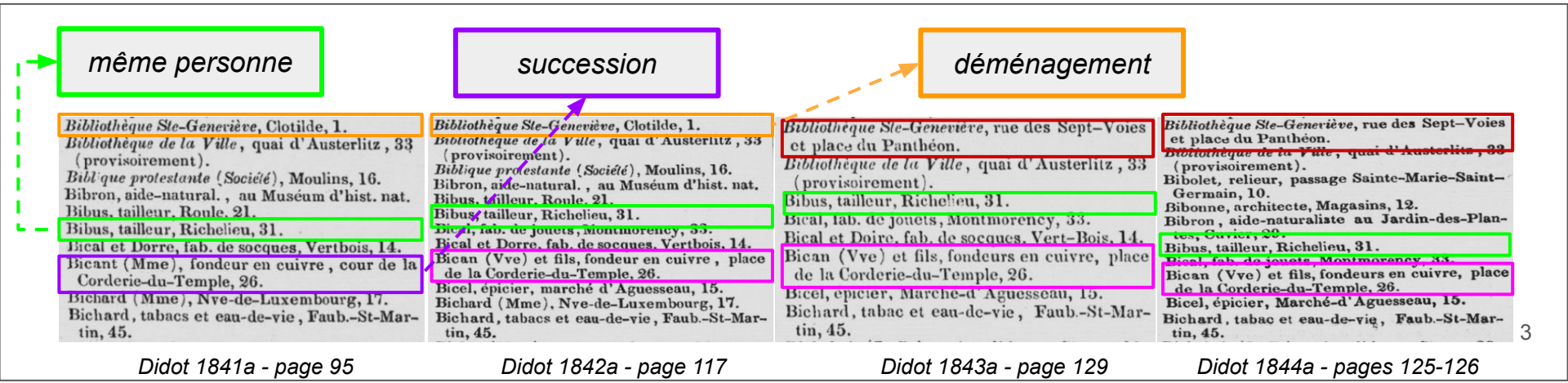
Liste ordonnée par nom - Annuaire Cambon Almgène de 1839.

Source : Gallica BNF



Annuaires du commerce de Paris

- Redondance importante des entrées d'une année ou d'une édition à une autre
- Snapshots temporels



Didot 1841a - page 95 | *Didot 1842a - page 117* | *Didot 1843a - page 129* | *Didot 1844a - pages 125-126*

Source images : Gallica BNF

Des sources géographiques anciennes à l'analyse géohistorique

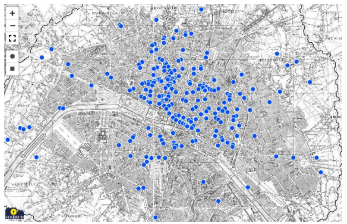
Reconnaissance | Classification | Géo-référencement | Structuration

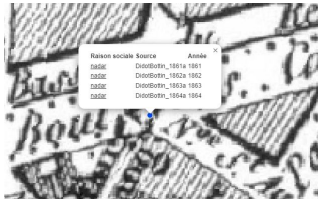
Scan → Texte

Texte →
Entités nommées spatiales

Entités nommées spatiales →
Données géographiques

Données géographiques →
Données géohistoriques





Segmentation des pages + Optical Character Recognition (OCR)

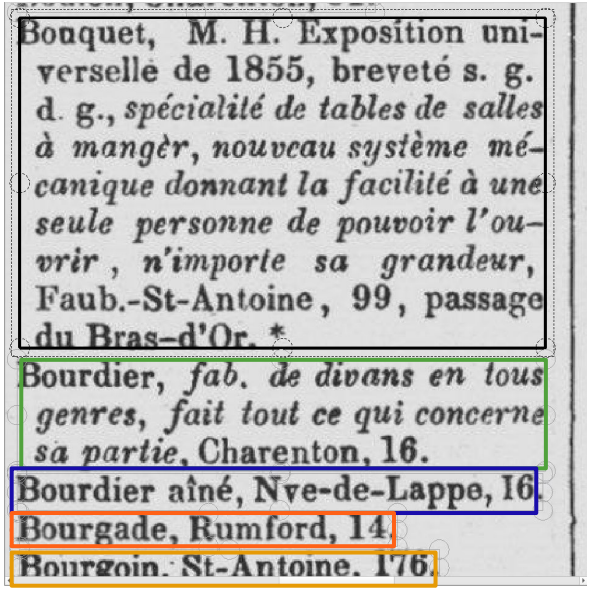
Reconnaissance des Entités Nommées (NER)

Géocodage

Filtrage et liage des entrées

Segmentation des entrées et reconnaissance optique des caractères (OCR)

Segmentation des pages en colonnes, paragraphes (XY cut), lignes (watershed) regroupées en entrées.

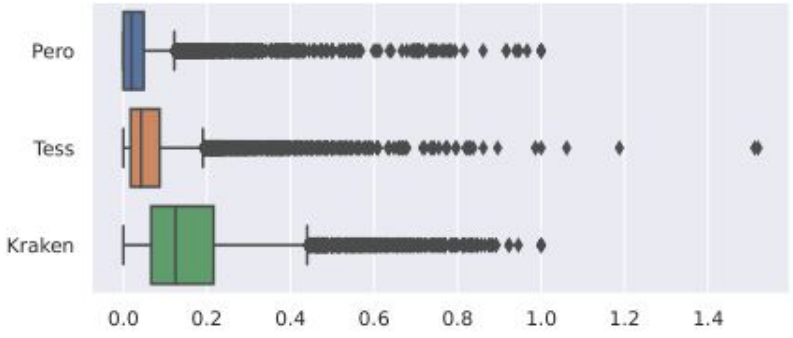


Source : Gallica BNF

Comparaison de 3 modèles d'OCR

	PERO OCR	Tesseract	Kraken
CER	3.78%	6.56%	15.72%

CER : Character Error Rate



Abadie et al. (2022)

Reconnaissance des entités nommées (NER)

Datasets

PER, ACT, LOC, CARD

Dataset Référence : entrées transcrites et annotées manuellement

Dufant (Victor), libraire, r. du Gros-Chenet, 2. 392
TDufant (Victor), libraire, r. du Gros-Chenet. 2.JO

Dataset Bruité : mêmes entrées transcrites avec PERO-OCR et annotations alignées automatiquement

Dutils, limonadier, r. des Colonnes, 6. 395
lukils. limonauier, 1. des Colonnes, 6. J9

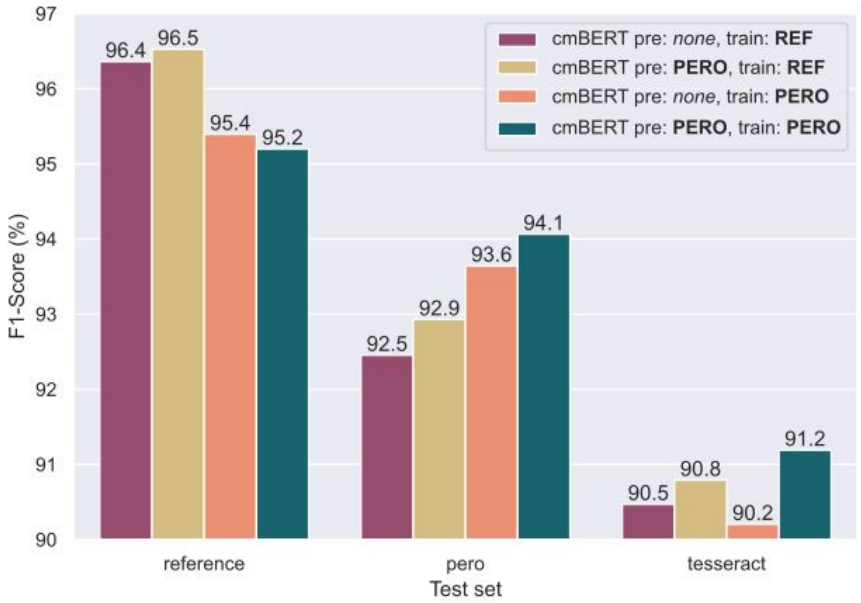
Modèles

- CamemBERT
- CamemBERT pré-entraîné avec des entrées d'annuaires OCrisées

8341 entrées annotées :

- **Train : 6004 entrées**
- **Dev : 668 entrées**
- **Test : 1669 entrées**

Reconnaissance des entités nommées (NER)



Abadie et al. (2022)

Mettereau, prop., quai d'Anjou, 7.\n
Mettemberg, élig., méd., St-Thomas-d'Enf., 5.\n
Metz (de), rentier, St-Guillaume, 30.\n
Metzinger, avocat, Rameau, 6.\n
Metzmacher, peint. sur émaux, St-Martin, 124.\n
Meurgey, épicier-herboriste, Dragon, 33.\n
Meurice, Chaussée-d'Antin, 3.\n
Meurice (Eug.), tapissier, Vivienne, 12.\n
Meurillon, marbrier-sculpteur, butte Mont-\n
Parnasse, 15.\n

NER

Mettereau PER, prop. ACT, quai d'Anjou LOC, 7 CARDINAL.
Mettemberg PER, élig. TITRE, méd. ACT, St-Thomas-d'Enf. LOC, 5 CARDINAL.
Metz (de) PER, rentier ACT, St-Guillaume LOC, 30 CARDINAL.
Metzinger PER, avocat ACT, Rameau LOC, 6 CARDINAL.
Metzmacher PER, peint. sur émaux ACT, St-Martin LOC, 124 CARDINAL.
Meurgey PER, épicier-herboriste ACT, Dragon LOC, 33 CARDINAL.
Meurice PER, Chaussée-d'Antin LOC, 3 CARDINAL.
Meurice (Eug.) PER, tapissier ACT, Vivienne LOC, 12 CARDINAL.
Meurillon PER, marbrier-sculpteur ACT, butte Mont-Parnasse LOC, 15 CARDINAL.

Base de données contenant environ 22 millions d'entrées structurées

=> Environ 10 millions avec la première extraction

Géocodage

Sources géohistoriques des données d'adresses et de rues utilisées pour localiser les annuaires :

Rues de l'Atlas de Verniquet et adresses "Paris artistique, 1800-1820"

Rues et adresses de l'atlas Jacoubet, 1827-1836

Rues et adresses de l'atlas municipal, 1888

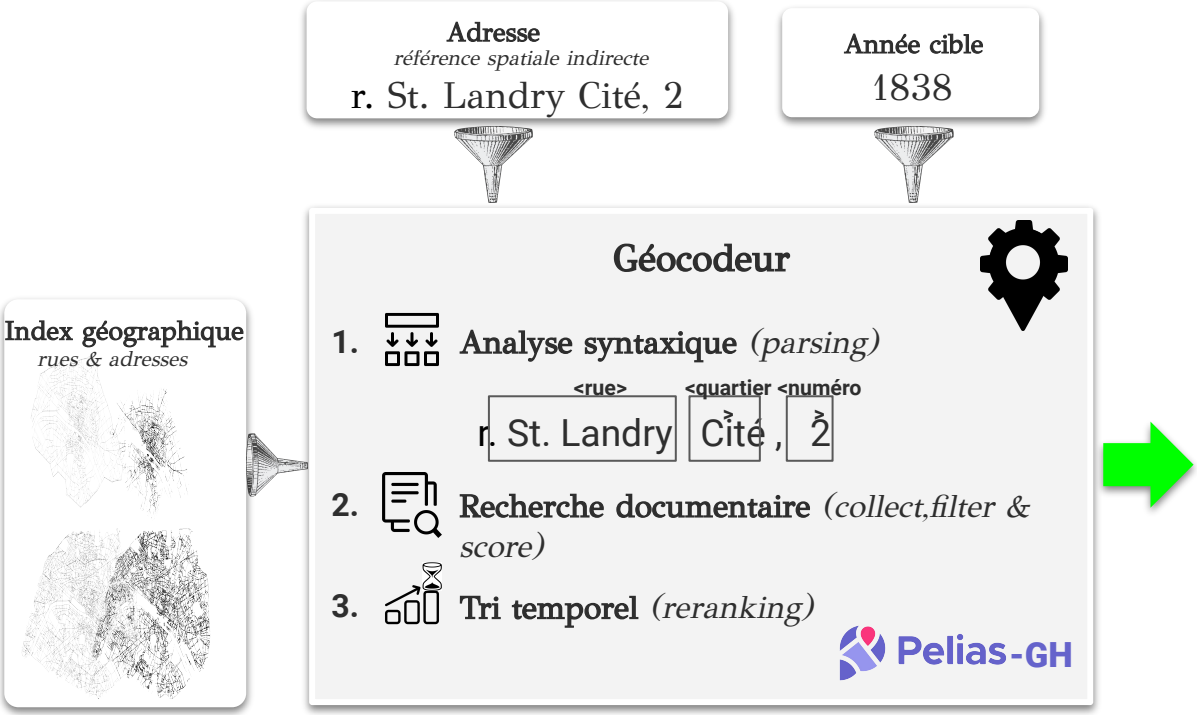
Rues et adresses OpenStreetMap, 2020

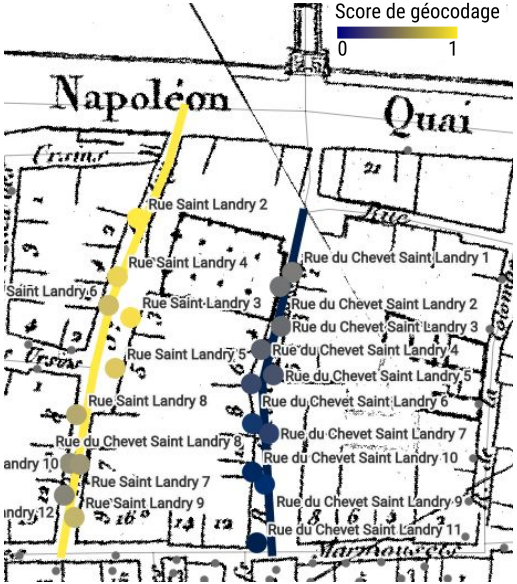
Duménieu (2023)

Géocodage

Géocodeur sensible à la dimension temporelle

96% des entrées sont géocodées.





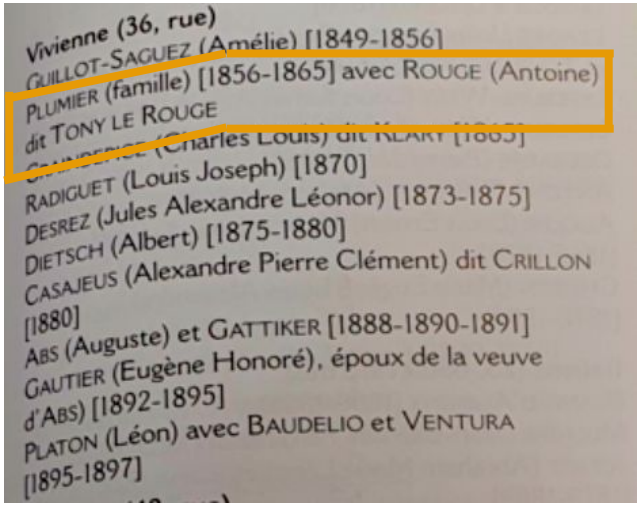
Outil open-source : https://github.com/GeoHistoricalData/historical_geocoding

Duménieu (2023)

- ★ CQ1. Quelle est l'adresse du commerce X en *1862* ?
- ★ CQ2. Combien y a-t-il de commerces de ce type localisé *rue de Rivoli* en *1856* ?
- ★ CQ3. Quels sont les commerces situés dans une zone définie par un polygone ou un rectangle englobant en *1875* ?
- ★ CQ4. Quels commerces ont potentiellement déménagé au cours de leur existence ?
- ★ CQ5. Quels commerces ont été repris par un autre commerçant exerçant la même activité ?

- Constitution d'une liste de mots-clés utilisés pour filtrer les entrées pertinentes pour notre cas d'étude

Recherche des photographes et ateliers (~230) listés par des historiens de l'art dans la base de donnée des extractions


Vivienne (36, rue)
GUILLOT-SAGUEZ (Amélie) [1849-1856]
PLUMIER (famille) [1856-1865] avec ROUGE (Antoine) dit TONY LE ROUGE
RADIGUET (Louis Joseph) [1870]
DESREZ (Jules Alexandre Léonor) [1873-1875]
DIETSCH (Albert) [1875-1880]
CASAJEUS (Alexandre Pierre Clément) dit CRILLON [1880]
ABS (Auguste) et GATTIKER [1888-1890-1891]
GAUTIER (Eugène Honoré), époux de la veuve d'ABS) [1892-1895]
PLATON (Léon) avec BAUDELIO et VENTURA [1895-1897]

Marc Durand (dir.). **« De l'image fixe à l'image animée : 1820-1910. Tome 2 : actes des notaires de Paris pour servir à l'histoire des photographes et de la photographie »**. Archives nationales (2015) . Pierrefitte-sur-Seine.

```
SELECT *
FROM directories.elements AS e
WHERE e.persons ILIKE '%plumi%'
ORDER BY e.published
```

Exemple de requête SQL
Toutes les entrées dont l'entité "Nom de personne/commerce" contient "plumi"

Plumier (Victor), **portraits** sur plaques et sur pap., Vivienne, 36.

Didot_1856a
Exemple de résultat

- Constitution d'une liste de mots-clés utilisés pour filtrer les entrées pertinentes pour notre cas d'étude

Réduction de la liste de mots-clés utilisés aux trois mots les plus couramment associés aux photographes listés dans la référence.

PHOTO	DAGUER	OPTI
Photographe	**Daguerréotype**	**Optique**
Photographie		**Opticien**

Juin 2022 : 34 062 entrées

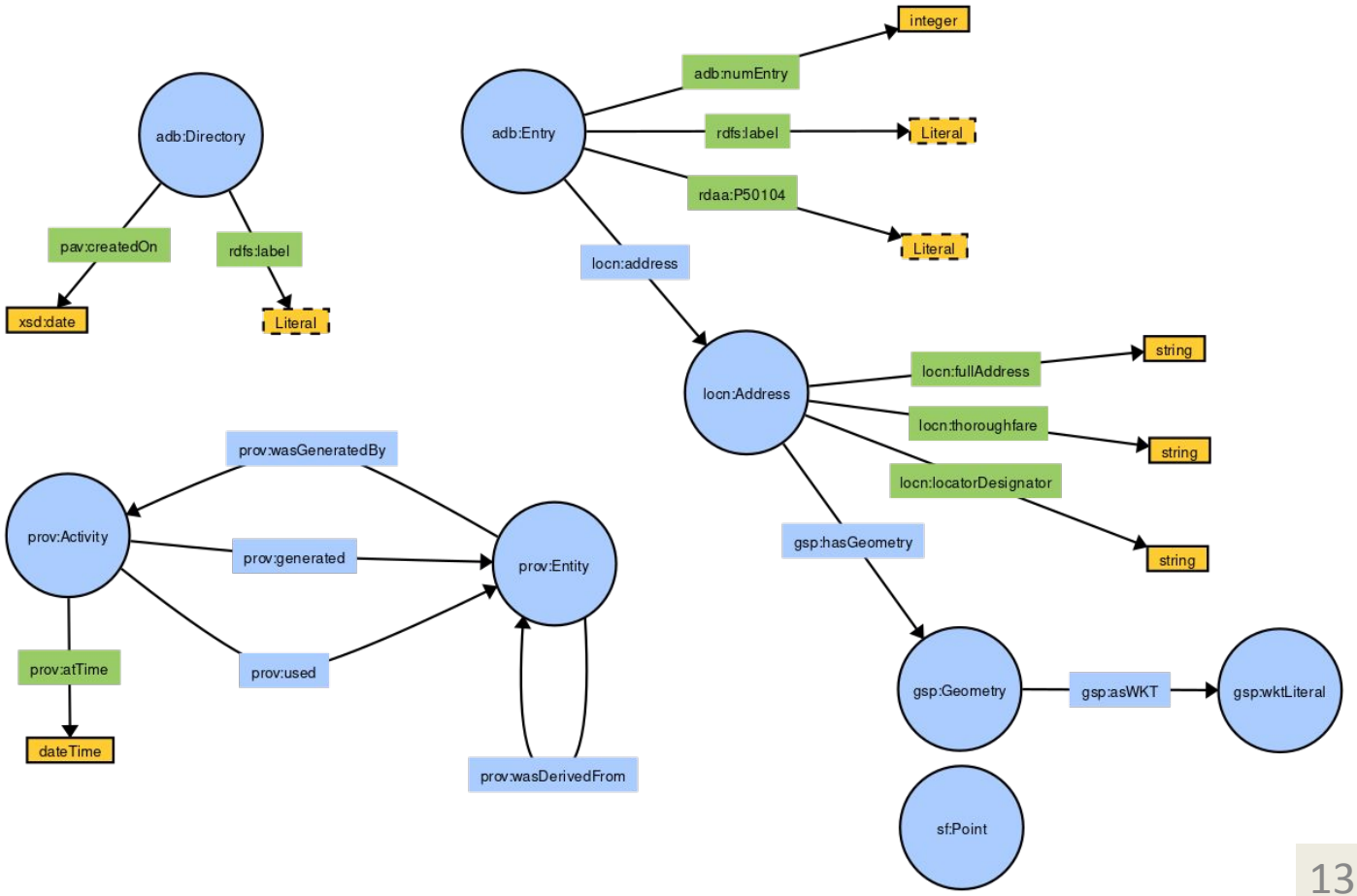
UPDATE Octobre 2023 : 45 807 entrées (listes alphabétiques uniquement)

Création de ressources RDF à partir de la base de donnée relationnelle contenant ces entrées (avec *ontop*)

- Ontologie

Vocabulaires utilisés :

- **RDFS :** label des entrées
- **LOCN** : description des adresses
- **Geosparql** : description des objets géographiques
- **RDA Registry** : description des activités
- **PROV** : description des données



- Méthodes de liage

Méthode fondée sur les connaissances

Appariement des entrées par **comparaison stricte** de clés

→ Clé = combinaison de propriétés, découvertes avec *Sakey*

- Numéro de l'entrée (si plusieurs activités ou adresses dans l'entrée)
- Nom et Activité
- Nom et Adresse
- Adresse et Activité

→ Création des liens par inférence

- Méthodes de liage

Méthode fondée sur les données

→ Appariement par **comparaison numérique** des propriétés

→ Distances d'édition : Levenshtein, TokenWise

Distance de Levenshtein : compte le nombre d'insertions, de remplacements et de suppressions de caractères pour passer d'un mot à un autre

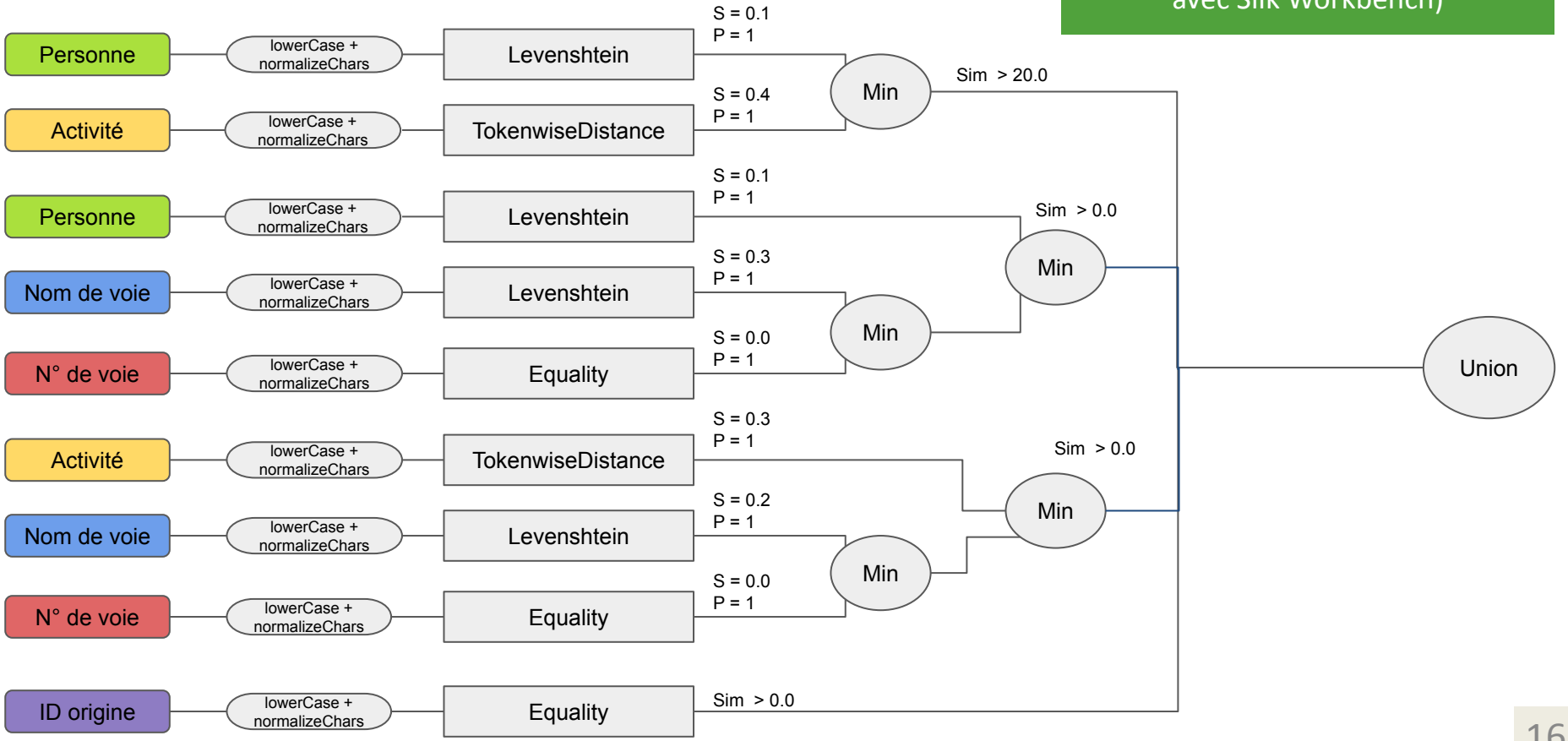
ex : **CH**AT <=> **PL**AT 2 remplacements

Distance TokenWise :

- Calcul de la distance de Levenshtein entre tous les mots du texte
- Score final normalisé (entre 0 et 1)

Liage réalisé avec Silk
Single-machine (tests des critères avec Silk Workbench)

- Critères de liage (méthode numérique)



- Méthodes de liage

Méthode logique

Paramétrage simple = clés

Comparaison très stricte :

- peu adaptée pour traiter des textes bruités
- liens sûrs
- temps et ressources de calcul importantes

1ère extraction

34 062 entrées

⬇️

250 622 liens

Méthode numérique

Paramétrage complexe = identifier les seuils de tolérance pertinents

Comparaison numérique :

- plus tolérante au bruit OCR
- risque plus élevé de produire des liens erronés (importance de choisir des seuils plutôt strictes)

& Propagation des liens sameAs

34 062 entrées

⬇️

357 130 liens

- Méthodes de liage

Méthode numérique

Paramétrage complexe = identifier les seuils de tolérance pertinents

Comparaison numérique :

- plus tolérante au bruit OCR
- risque plus élevé de produire des liens erronés (importance de choisir des seuils plutôt strictes)

→ Graphe géohistorique : https://dir.geohistoricaldata.org/sparql

2ème extraction

45 807 entrées

⬇️

841 609 liens

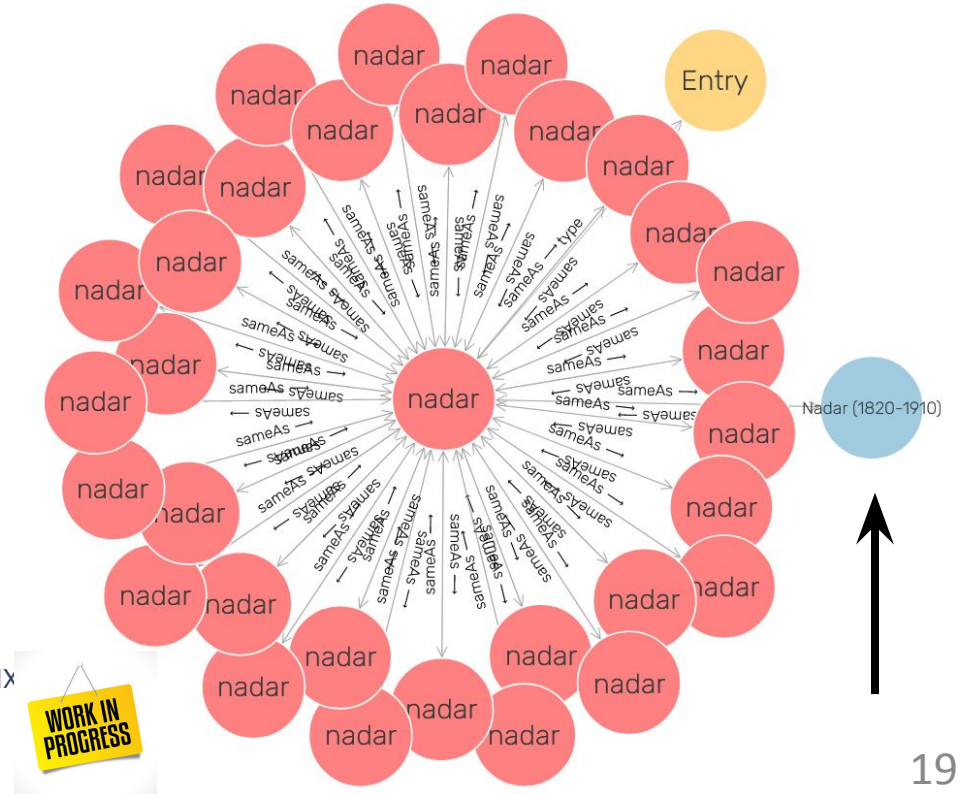
Liage avec d'autres ressources

Test d'appariement par comparaison numérique des propriétés relatives au nom des photographes entre :

Annuaires → Raison sociale

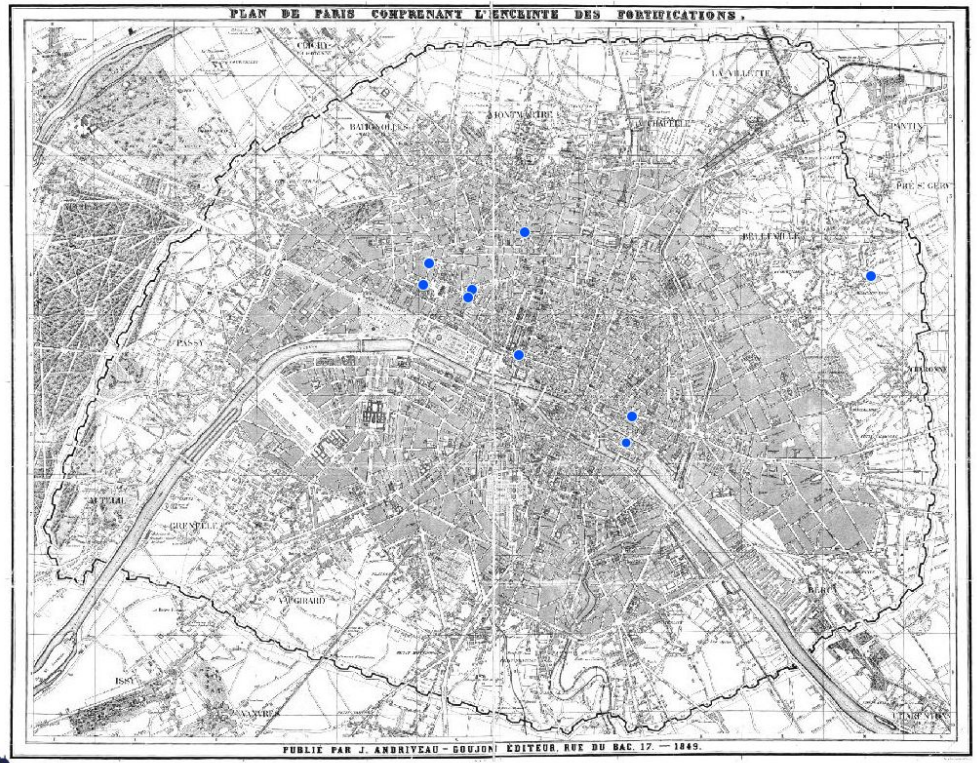
DATA BNF → Libellé principal, Libellés alternatifs

- 84 ressources "Photographes" sur data.bnf.fr
- 192 liens soit 10 ressources *data.bnf.fr* liées aux ressources issues des Annuaires



Visualisation spatiale…

bit.ly/graphes_geohistoriques_soduco



Plan de Paris contenant l'enceinte des fortifications - Andriveau-Goujon - 1849 © GeoHistoricalData

Visualisation des graphes géohistoriques construits à partir des entrées d'annuaires du commerce de Paris (XIXème siècle)

SPARQL Endpoint | Dépôt Git-Hub | Aide

Dataset

Photographes et professions associées

Statistiques du dataset

Filtres

Propriétés

Raison sociale

nadar

Description

Ex : photo

Adresse

Ex : rivoli

Les champs du formulaire acceptent les expressions régulières. Deux regex utiles :

- La propriété contient *mot1* **ET** *mot2*: (?=.*mot1)(?=.*mot2)
- La propriété contient *mot1* **OU** *mot2*: *mot1|mot2*

Tester une regex complexe.

Période

1790 1915

1790 1821 1853 1884 1915

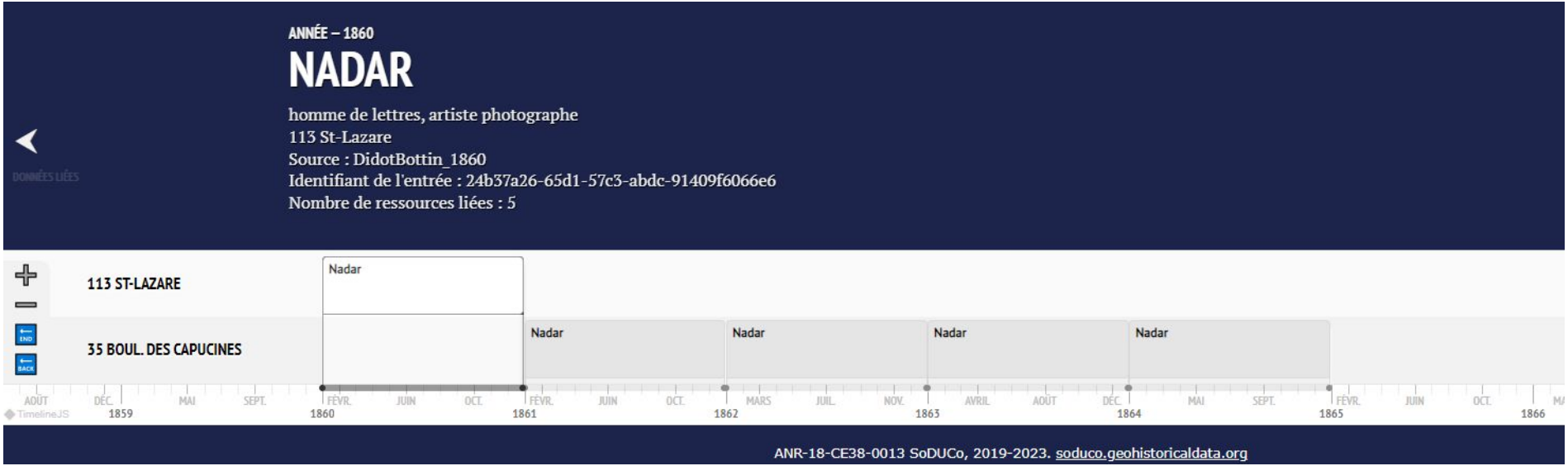
Le filtre temporel permet de faire varier l'affichage des points préalablement chargés sur la carte sans lancer une nouvelle recherche.

Données chargées pour la période **1790-1915**.

Leaflet | Plan de Paris contenant l'enceinte des fortifications - Andriveau-Goujon - 1849 © GeoHistoricalData

… et temporelle

bit.ly/graphes_geohistoriques_soduco

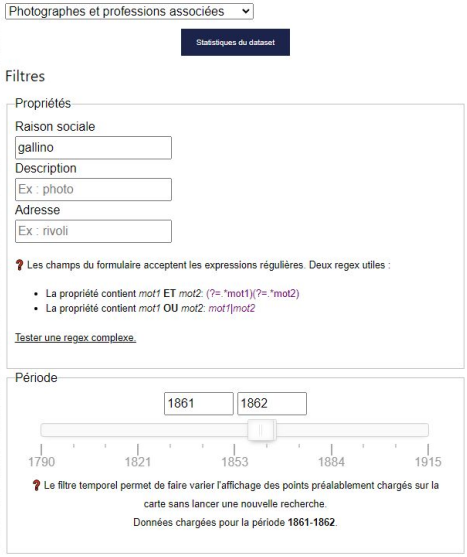


Questions de compétences

Quelle est l'adresse du commerce de Gallino en 1862 ?

11 rue suger


```
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX ont: <http://rdf.geohistoricaldata.org/def/directory#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX pav: <http://purl.org/pav/>
select * where { graph <http://rdf.geohistoricaldata.org/id/directories/photographes> {
    ?e a ont:Entry.
    ?e rdfs:label ?label.
    ?e prov:wasDerivedFrom ?directory.
    ?directory pav:createdOn 1862.
    ?e locn:address ?add.
    ?add locn:fullAddress ?fullAdd.
    ?add prov:wasGeneratedBy <http://rdf.geohistoricaldata.org/id/directories/activity/0001>.
    Filter regex(lcase(?label), "gallino").
}
}
```



Questions de compétences

Combien y a-t-il de commerces de ce type localisés rue de Rivoli en 1856 ?

14


```
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX ont: <http://rdf.geohistoricaldata.org/def/directory#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX pav: <http://purl.org/pav/>
select (count(distinct ?e) as ?nombre) where {graph <http://rdf.geohistoricaldata.org/id/directories/photographes> {
    ?e a ont:Entry.
    ?e prov:wasDerivedFrom ?directory.
    ?directory pav:createdOn 1856.
    ?e locn:address ?add.
    ?add locn:thoroughfare ?voie.
    ?add prov:wasGeneratedBy <http://rdf.geohistoricaldata.org/id/directories/activity/0001>.
  Filter regex(lcase(?voie), "rivoli").
}
}
```

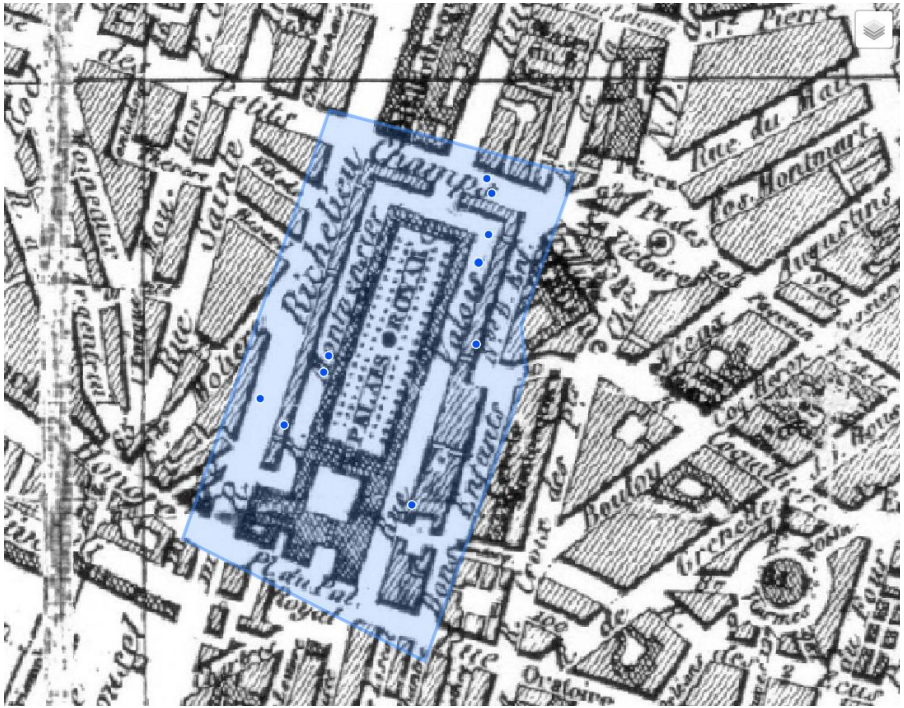
Questions de compétences

Quels sont les commerces situés dans une zone définie par un polygone ou un rectangle englobant en 1875 ?

```
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX ont: <http://rdf.geohistoricaldata.org/def/directory#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX pav: <http://purl.org/pav/>
PREFIX gsp: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
select ?label where { graph <http://rdf.geohistoricaldata.org/id/directories/photographes> {
    ?e a ont:Entry.
    ?e rdfs:label ?label.
    ?e prov:wasDerivedFrom ?directory.
    ?directory pav:createdOn 1875.
    ?e locn:address ?add.
    ?add gsp:hasGeometry ?geom.
    ?geom gsp:asWKT ?wkt.
    FILTER (geof:sfIntersects(?wkt, "<http://www.opengis.net/def/crs/OGC/1.3/CRS84> Polygon((2.3360 48.8625,2.3360 48.8663,2.3385 48.8663,2.3385 48.8625,2.3360 48.8625))"^^gsp:wktLiteral))
}
}
```

Questions de compétences

Quels sont les commerces situés dans une zone définie par un polygone ou un rectangle englobant en 1875 ?



Exemple de la zone située autour du Palais royal :

Allévy

Barrès

Baur Vve

Bureau S

Crillon A

Derepas

Legros

Mustière

Prudent

Vantier

Questions de compétences

Quels commerces ont potentiellement déménagé au cours de la période 1860-1870 ?

```
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX ont: <http://rdf.geohistoricaldata.org/def/directory#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX pav: <http://purl.org/pav/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select distinct ?label ?voie ?labelo ?voieo where { graph <http://rdf.geohistoricaldata.org/id/directories/photographes> {
    ?e a ont:Entry.
    ?e rdfs:label ?label.
        ?e prov:wasDerivedFrom ?directory.
        ?directory pav:createdOn ?date.
        ?e locn:address ?add.
        ?add locn:thoroughfare ?voie.
    ?e owl:sameAs ?o.
    ?o rdfs:label ?labelo.
    ?o locn:address ?addo.
        ?addo locn:thoroughfare ?voieo.
        Filter ((?date > 1860) && (?date < 1870)).
    Filter (!sameTerm(?voie, ?voieo))
  }}
```

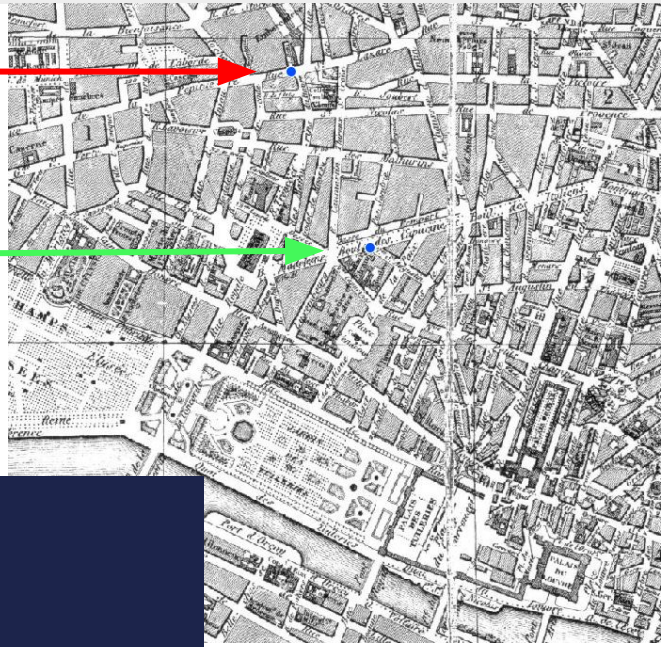
Exemple : Déménagement de Nadar de la Rue St-Lazare au Boulevard des Capucines vers 1860-1861

Questions de compétences

Quels commerces ont potentiellement déménagé au cours de la période 1860-1870 ?

St-Lazare

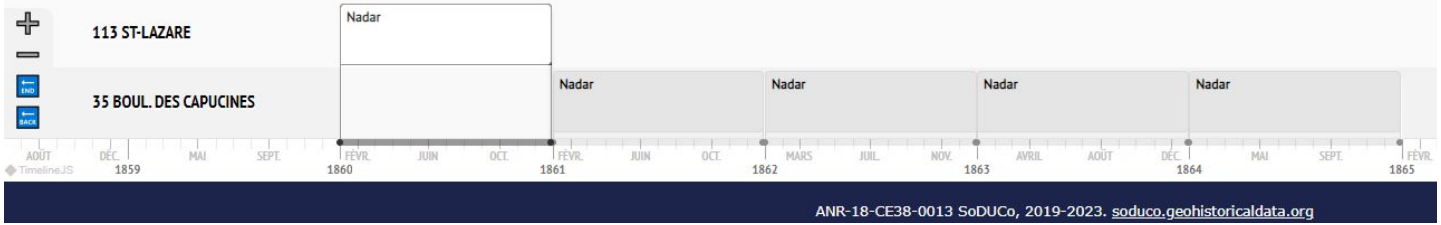
Boulevard des Capucines



ANNÉE – 1860

NADAR

homme de lettres, artiste photographe
113 St-Lazare
Source : DidotBottin_1860
Identifiant de l'entrée : 24b37a26-65d1-57c3-abdc-91409f6066e6
Nombre de ressources liées : 5



ANR-18-CE38-0013 SoDUCo, 2019-2023. soduco.geohistoricaldata.org

Questions de compétences

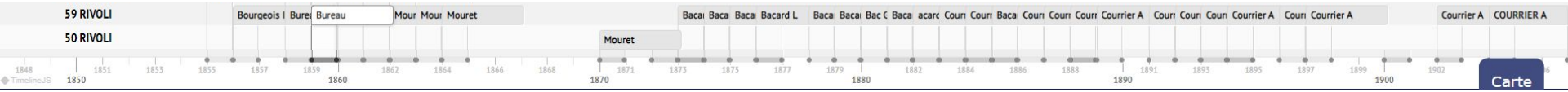
Quels commerces ont été repris par un autre commerçant exerçant la même activité ?

```
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX ont: <http://rdf.geohistoricaldata.org/def/directory#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX pav: <http://purl.org/pav/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select distinct ?label ?fadd ?labelo ?faddo where { graph <http://rdf.geohistoricaldata.org/id/directories/photographes> {
    ?e a ont:Entry.
    ?e rdfs:label ?label.
        ?e locn:address ?add.
        ?add locn:fullAddress ?fadd.
    ?e owl:sameAs ?o.
    ?o rdfs:label ?labelo.
    ?o locn:address ?addo.
        ?addo locn:fullAddress ?faddo.
    Filter (!sameTerm(?label, ?labelo) && sameTerm(?fadd, ?faddo) )
  }}
```

Exemple : Succession des photographes Bourgeois puis Bureau puis Mouret puis Bacard puis Courrier (A.) au 59 rue de Rivoli entre 1856 et 1908

Questions de compétences

Quels commerces ont été repris par un autre commerçant exerçant la même activité ?



Bourgeois

Bureau

Mouret

Bacard

Courrier A

Exemple : Succession des photographes Bourgeois puis Bureau puis Mouret puis Bacard puis Courrier (A.) au 59 rue de Rivoli entre 1856 et 1908

Conclusion

Démarche fonctionnelle d'exploitation des entrées issues des annuaires

- extraction fine et en masse
- entrées structurées, souvent bruitées.

Travaux à venir :

- évaluation quantitative des appariements,
- mise en œuvre de l'approche pour d'autres types de commerces,
- aide pour l'explicitation des relations spatio-temporelles entre les entrées (succession, déménagement),
- liage avec des ressources externes

Méthode, code, application et données accessibles sur le web.

Merci pour votre attention

Tutoriel, ontologie, ressources, code et questions de compétences
https://github.com/soduco/atelier_graphes_geohistoriques_annuaires

Application
bit.ly/graphe_geohistorique_soduco

Endpoint
https://dir.geohistoricaldata.org/sparql

Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Joseph Chazalon et Edwin Carlinet. Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19ème siècle: application aux métiers de la photographie. *IC 2023, 34èmes journées francophones d'Ingénierie des connaissances*, Strasbourg, France, 3-7 Juillet 2023. hal-04121643

Références

Nathalie Abadie, Edwin Carlinet, Joseph Chazalon, et Bertrand Duménieu. **A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories**. Number 13237 in Document Analysis Systems. DAS 2022., La Rochelle, France, May 2022. Springer, Cham.

Bertrand Duménieu. **Localiser les adresses anciennes ? Plans historiques & annuaires parisiens au XIXe siècle.** Journée d'étude scientifique "Pour un traitement scientifique automatisé des cartes anciennes", Festival Printemps des Cartes 2023

Marc Durand (dir.). **De l'image fixe à l'image animée : 1820-1910. Tome 2 : actes des notaires de Paris pour servir à l'histoire des photographes et de la photographie**. Archives nationales (2015). Pierrefitte-sur-Seine.

Danai Symeonidou, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. **SAKey : Scalable Almost Key Discovery in RDF Data.** In proceedings of the 13th International Semantic Web Conference, ISWC 2014, volume Lecture Notes in Computer Science of The Semantic Web – ISWC 2014, pages 33–49, Riva del Garda, Italy, October 2014. Springer.

Silk Linked Data Integration Framework : outil de liage de données, https://github.com/silk-framework/silk

Pero OCR, https://pero-ocr.fit.vutbr.cz/