Algorithmes pour l'optimisation et la gestion des réseaux – Introduction aux graphes de connaissances

I&I @ M2 MVA – ENS Paris Saclay

Lionel Tailhardat
lionel.tailhardat@orange.com
genears.github.io

Février/Mars 2025

MATHÉMATIQUES VISION APPRENTISSAGE

orange

Syllabus & planning

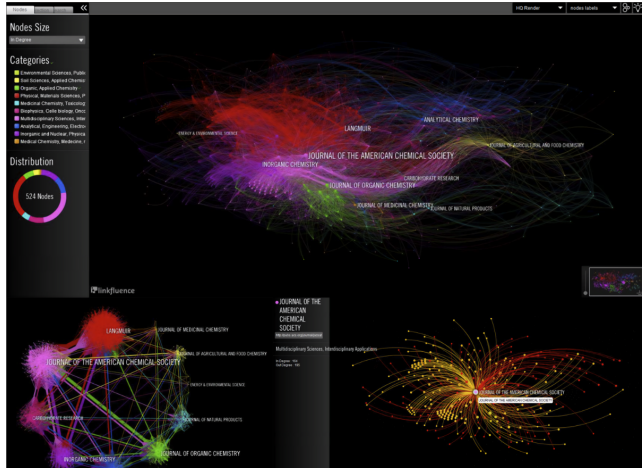
Le concept de **graphe de connaissances** a émergé autour de 2010 mais trouve ses racines dans des travaux bien plus anciens en IA en rapport à la **représentation formelle des connaissances** et le **raisonnement logique** sur celles-ci (p.ex. les graphes sémantiques, les logiques de description).

En quoi ces graphes diffèrent de ceux évoqués classiquement par la théorie des graphes ? et en quoi ces graphes peuvent-ils nous être utiles pour la gestion des réseaux ?

Nous verrons dans cette leçon d'introduction que les graphes de connaissances et les techniques algorithmiques associées permettent de répondre aux problématique de nombreux domaines d'application dès lors qu'il s'agît de travailler avec des **données hétérogènes**, de **raisonner sur la sémantique et la structure des données**, de réaliser des **inférences avec des garanties intrinsèques d'explicabilité**, et ce sans pour autant s'empêcher d'utiliser des techniques et approches d'autres registres tels que l'optimisation combinatoire, l'apprentissage automatique, ou les réseaux de neurones.

7 février	Présentation de l'utilisation des graphes de connaissances pour la détection d'anomalies sur des infrastructures réseaux.
14 février	Introduction aux techniques de construction et d'exploitation des graphes de connaissances + présentation du projet étudiant #2.
28 mars	Restitution du projet #2.

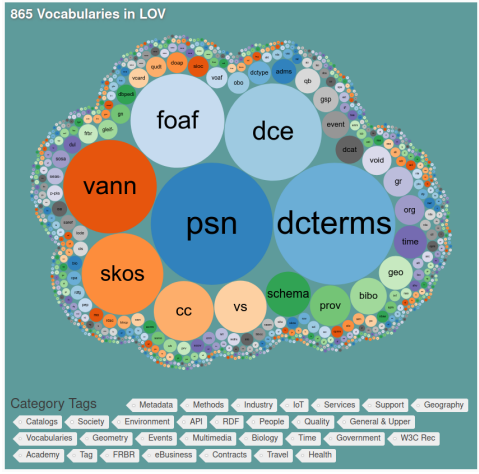
Un avant goût – Décrire et réfléchir le monde (1)



Cartographie numérique de la recherche en chimie en France en 2009 à partir des données du WoS (Linkfluence, WebAtlas, 2010). Extrait de “Qu’est-ce que la cartographie du web ?” (Franck Ghitalla, 2021). https://doi.org/10.4000/books.oep.15358

↪ Mettre ensemble des données parlant de choses différentes pour disposer d’une vue d’ensemble et comprendre les interrelations.

Un avant goût – Décrire et réfléchir le monde (2)



Les vocabulaires référencés dans le catalogue Linked Open Vocabularies (LOV) au 17 janvier 2025. https://lov.linkeddata.es/dataset/lov

↪ Des façons standardisées et interopérables de représenter et interpréter divers domaines de discours.

Bibliographie & références

Quelles compétences et quelles ressources pour approfondir ?

Quelles compétences ? (1)

Dans “Teaching Knowledge Graph for Knowledge Graphs Education” (Eleni Ilkou, et al., 2024, pré-publication SWJ), les auteurs s’intéressent au socle de compétences pour travailler avec les graphes de connaissances et les technologies du Web Sémantique.

Ils proposent d’organiser ces compétences à l’aide d’un graphe de connaissances :-)

Main Skills	**Knowledge Topics**
RDF	RDF (Resource Description Framework) introduction. RDF syntax (Turtle, RDF/XML, JSON-LD).
Knowledge Graph Standards and Tools	Overview of standards like RDF, RDFS, OWL, and SHACL, FAIR principles. Popular knowledge graph tools and frameworks (e.g., Apache Jena, Neo4j).
SPARQL	SPARQL query language for querying RDF data.
Secondary Skills	**Knowledge Topics**
Linked Data	Linked Data principles and best practices. Ontologies and their role in the Semantic Web.
Graph Theory	Introduction to graph theory concepts (nodes, edges, etc.) Types of graphs (property, directed, etc.) Graph algorithms for KGs (e.g., community detection, centrality measures) Graph traversal techniques.
Validation of KGs	Data accuracy, consistency, and adherence to defined schemas or constraints like SHACL and SHEx. Schema validation, data quality dimensions (e.g., correctness, completeness), and resolving inconsistencies through logical inference and debugging.
Ontology Engineering	Ontology design patterns and best practices. Ontology alignment and merging.
Knowledge Graph Construction	Data extraction and integration techniques. Entity recognition and disambiguation. Knowledge graph population and curation.
Knowledge Graph Reasoning.	Inference and reasoning in KGs. Rule-based and ontology-based reasoning.
Scalability and Performance	Knowledge graph storage and processing. Scalability challenges and solutions. Benchmarking and performance optimization.
NLP, ML and Knowledge Graphs	Named entity recognition, extraction and linking. Relation extraction from text. Graph neural networks for KG analytics. Knowledge graph completion, link prediction and entity embeddings
Case Studies	Industry-specific use cases (e.g., healthcare, e-commerce). Practical examples of KG implementations (e.g. fact checking, QnA)
Knowledge Graph Visualization	Visualization techniques for KGs. Tools and libraries for creating interactive visualizations.
AI Ethics, Bias and Legal Considerations	Data privacy and security in KGs.Ethical concerns related to KG construction and usage. Addressing bias and fairness issues in KGs Compliance with regulations (e.g., GDPR).
Blockchain and Decentralized Knowledge Graphs	Decentralized KG architectures using blockchain technology. Data ownership and security in decentralized KGs.
Domain Knowledge Graphs	In-depth exploration of KGs in specialized fields (e.g., healthcare, geospatial). Domain-specific standards and ontologies (e.g., HL7 FHIR in healthcare).
Knowledge Graph Governance and Quality	Techniques for assessing and improving data quality in KGs. Governance models and policies for maintaining KGs.

Quelles compétences ? (2)

Dans le cours magistral “Les technologies du Web Sémantique et de l’extraction d’information” (Raphaël Troncy, EURECOM. 2024, WebSem-2024fall), l’enseignement s’effectue par un module de 21h au sein d’un tronc commun d’ingénieur en sciences des données.

Le cours est présenté comme suit:

“Le Web sémantique désigne un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles, utilisant notamment une famille de langages développés par le W3C. Ce cours est une visite guidée d’un certain nombre de recommandations du W3C permettant de représenter (RDF/S, SKOS, OWL) et d’interroger (SPARQL) des connaissances sur le web ainsi que les formalismes logiques sous-jacents à ces langages, leur syntaxe et leur sémantique formelle. Nous aborderons également les problèmes posés par la construction de systèmes à base de connaissances et de leur mise sur le réseau (alignement). Nous montrerons finalement comment extraire de la connaissance à partir de textes en utilisant des techniques de traitement de la langue naturelle.”

REQUIREMENTS

Connaissances basiques des technologies du web (html, css, javascript) ou des bases de données est un plus

DESCRIPTION

Ce cours est une visite guidée d'un certain nombre de recommandations du W3C permettant de représenter (RDF/S, SKOS, OWL) et d'interroger (SPARQL) des connaissances sur le web ainsi que les formalismes logiques sous-jacents à ces langages, leur syntaxe et leur sémantique formelle. Nous aborderons également les problèmes posés par la construction de systèmes à base de connaissances et de leur mise sur le réseau (alignement). Nous montrerons finalement comment extraire de la connaissance à partir de textes en utilisant des techniques de traitement de la langue naturelle.

Objectifs d'apprentissage:

- Maîtriser la pile des technologies du web sémantique
  - RDF: représenter de la connaissance sur le web
  - RDFS, SKOS, OWL: développer ses propres vocabulaires
  - SPARQL: interroger le web de données
- Extraction d'Information 101
  - Reconnaissance et désambigüisation d'entités nommées
  - Analyse de sentiment
- Développer des applications sur le web sémantique
  - Les principes des données liées: des données brutes maintenant !
  - Réconcilier des données sur le web en utilisant des techniques d'apprentissage
  - Interagir avec le web de données : RDFa, microdata, JSON-LD

Nb d'heures : 21.00

Evaluation :

- Rapports de travaux pratiques 1+2+3 (40% de la note finale),
- Examen Final (60% de la note finale)

Ressources (1)

Cours en ligne

- Web sémantique et Web de données. INRIA, MOOC gratuit de 21h. https://www.fun-mooc.fr/fr/cours/web-semantique-et-web-de-donnees/
- RDF and JSON-LD for Metadata. DublinCore Academy. https://academy.dublincore.org/

Livres

- G. Antoniou, et al. A Semantic Web Primer. 2nd Edition, MIT Press, 2009, 288p. http://www.semanticwebprimer.org/
- D. Allemang, et al. Semantic Web for the Working Ontologist. 1st Edition, Morgan Kaufmann, 2008, 384p. http://workingontologist.org/
- T. J. Pollock. Semantic Web for Dummies. Wiley, 2009, 432p. http://www.semanticwebfordummies.com/
- J.G. Breslin, et al. The Social Semantic Web. Springer-Verlag, 2009, 300p. http://socialsemanticweb.net/

Ressources (2)

Articles

- A. Hogan, et al. Knowledge Graphs. 2020. https://doi.org/10.1145/3447772

Journaux

- Semantic Web journal (SWJ, par IOS Press). https://www.semantic-web-journal.net/
- Transactions on Graph Data and Knowledge (TGDK). https://drops.dagstuhl.de/entities/journal/TGDK

Organismes et groupes de normalisation

- W3C standards and drafts. https://www.w3.org/TR/
- W3C KG Construction Community Group. https://www.w3.org/community/kg-construct/

Ressources (3)

Conférences – France

- Conférences EGC (Association Internationale Francophone d'Extraction et de Gestion des Connaissances). https://www.egc.asso.fr/manifestations/conferences
- Les journées francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA), organisées par l'Association Française pour l'Intelligence Artificielle (AFIA). https://afia.asso.fr/

Conférences – International

- European Semantic Web Conference (ESWC). https://eswc-conferences.org/
- International Semantic Web Conference (ISWC). https://iswc.umbc.edu/
- Conference on Information and Knowledge Management (CIKM). http://www.cikmconference.org/
- International Conferences on Knowledge Capture (K-CAP). https://www.k-cap.org/
- International Conference on Knowledge Engineering and Knowledge Management (EKAW). https://ekaw.org/

Ressources (4)

Catalogues de modèles sémantiques

- Linked Open Vocabularies (LOV). https://lov.linkeddata.es/
- IndustryPortal (Ontologies for industry) INP Toulouse / Ecole Nationale d'Ingéieurs de Tarbes (ENIT). https://industryportal.enit.fr/ontologies/
- EU Vocabularies (European Union). https://op.europa.eu/fr/web/eu-vocabularies/data-catalogue

Catalogues d'outils

- Semantic Web Tool Hub. https://semantic-tool-hub.github.io/
- Awesome KGC Tools. https://github.com/kg-construct/awesome-kgc-tools

Applications

Utilisation des graphes de connaissances pour la détection d'anomalies sur des infrastructures réseaux.

Les graphes de connaissances pour la détection d'anomalies

La gestion des réseaux doit satisfaire des **contraintes réglementaires** (disponibilité des services critiques, traçabilité des actions) et répondre à des **objectifs commerciaux** (disponibilité et performance des services, garantie de temps de rétablissement), qui amènent naturellement à chercher une **efficacité opérationnelle** maximale, notamment à travers des questions telles que:

Gestion des incidents. Comment pouvons-nous fournir une approche unifiée à la phase de diagnostic pour des systèmes complexes ?

Modélisation des anomalies. Quelles techniques algorithmiques incluent la notion de temps et disposent intrinsèquement de propriétés d'explicabilité ?

Aide à la décision. Comment apprendre/utiliser une représentation manipulable des anomalies ?

Pour entrer dans le vif du sujet:

↪ "Anomaly Detection using Knowledge Graphs and Synergistic Reasoning – Application to Network Management and Cyber Security" (Lionel Tailhardat, 2024). NNT:2024SORUS293 / manuscrit / pres-pdf.

Prochaine séance : 14 février

Programme :

- Nous nous donnerons les moyens d'"appeler un chat un chat".
- Nous discuterons du projet étudiant #2 "quelles stratégies pour maîtriser l'ordre (le nombre de sommets) et la taille (le nombre d'arrêtes) d'un graphe de connaissances".

Quelques questions à méditer entre-temps :

- Peut-on prendre une décision sur une décision ?
- Quelles techniques du cours "algorithmes pour l'optimisation et la gestion des réseaux" (Nancy Perrot) utiliseriez-vous pour la détection d'anomalies et/ou aider à la résolution d'incidents ?
- Où positionneriez-vous ces techniques dans l'équation (1), qui propose une représentation abstraite d'une chaîne de traitement de données dans un système d'aide à la décision :

$$\mathcal{E} \xrightarrow{\text{e.t.l.}} \mathcal{K}^{\circlearrowleft_{\mathsf{r}}} \xrightarrow{\text{infer.}} \mathcal{P} \tag{1}$$

$\mathcal{E}$ the Environment (i.e. primary and secondary data describing the network), $\mathcal{K}$ the Knowledge (i.e. information about the network behavior and business rules guiding network administration tasks), $\mathcal{P}$ the Propositions (i.e. explained inferences about the network states and behavior), e.t.l. an Extract-Transform-Load process or alike, r a reasoning process (i.e. an "internal" inference process aimed at producing facts and knowledge from basic facts present in $\mathcal{K}$) and infer. an inference process.

Mise en œuvre

Introduction aux techniques de construction et d'exploitation des graphes de connaissances.

Préambule philosophique

“appeler un chat un chat” (fr.wiktionary.org) :

- Probablement du grec ancien “appeler une figue une figue et une barque une barque”.
- Appeler les choses par leur nom ; dire les choses franchement, sans circonlocution.

... d’accord, mais :

- Qui a dit que ce qu’on appelle un chat doit s’appeler un chat ? ⟷ notions de **sémantique** (signifiant versus signifié) et de **provenance** de l’information.
- Qu’est-ce qui caractérise “un chat” et permet d’affirmer que ce qu’on observe / ce dont on discute est “un chat” ? ⟷ notions d’**univocité** (de “chose” en soi), de **domaine de discours**, et de **modèle**.
- Comment dire qu’un chat n’est pas “qu’un chat” ? (il s’agît de Léon, le chat de ma fille, pris en photo lorsqu’il venait d’être diplômé du M2 MVA) ⟷ notions d’**identité** et de **domaine d’interpétation**.

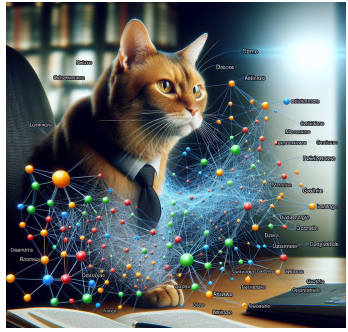
Préambule philosophique

“appeler un chat un chat” (fr.wiktionary.org) :

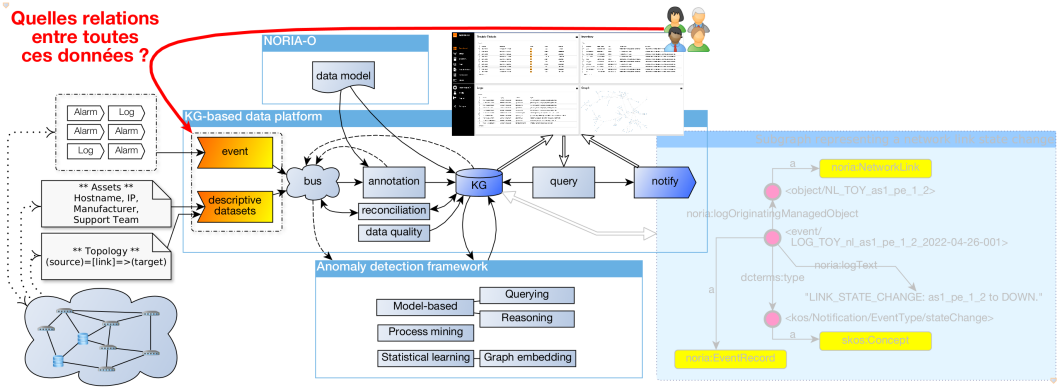
- Probablement du grec ancien “appeler une figue une figue et une barque une barque”.
- Appeler les choses par leur nom ; dire les choses franchement, sans circonlocution.

... d’accord, mais :

- Qui a dit que ce qu’on appelle un chat doit s’appeler un chat ? $\longleftrightarrow$ notions de **sémantique** (signifiant versus signifié) et de **provenance** de l’information.
- Qu’est-ce qui caractérise “un chat” et permet d’affirmer que ce qu’on observe / ce dont on discute est “un chat” ? $\longleftrightarrow$ notions d’**univocité** (de “chose” en soi), de **domaine de discours**, et de **modèle**.
- Comment dire qu’un chat n’est pas “qu’un chat” ? (il s’agît de Léon, le chat de ma fille, pris en photo lorsqu’il venait d’être diplômé du M2 MVA) $\longleftrightarrow$ notions d’**identité** et de **domaine d’interpétation**.



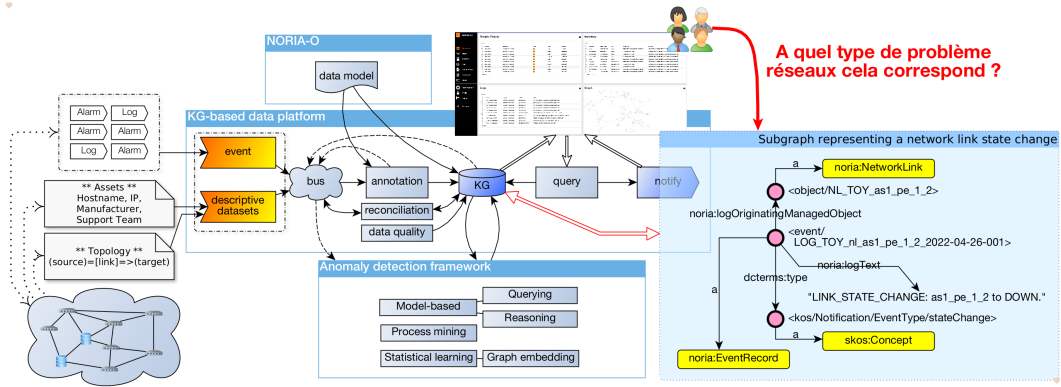
Principes d'architecture pour la construction et l'exploitation des graphes (1)



"appeler un chat un chat" ... il est donc globalement question de :

↪ Etre en capacité de **donner du sens** (analyse, catégorisation, et discrimination) à un **ensemble de données hétérogènes**.

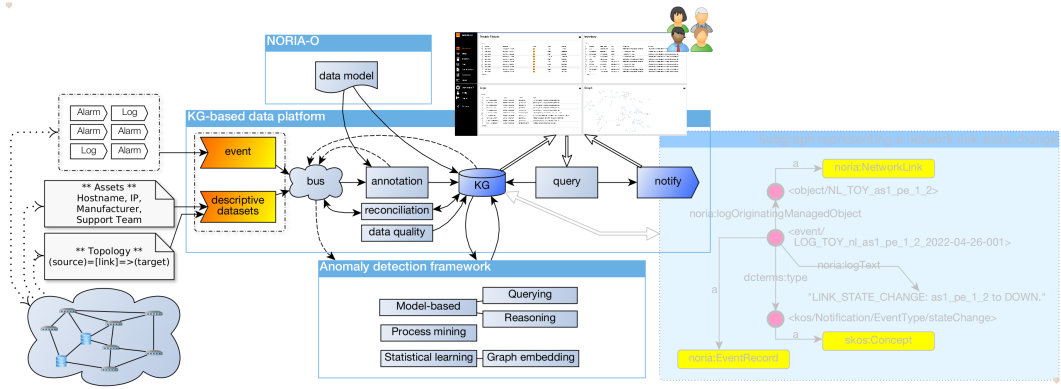
Principes d'architecture pour la construction et l'exploitation des graphes (2)



Nous faisons implicitement le pari que :

↪ La **nature** des entités (tangibles ou intangibles) du domaine étudié et la **structure relationnelle** relative à ces entités permettent de donner du sens à une entité donnée, voire à un groupe d'entités.

Principes d'architecture pour la construction et l'exploitation des graphes (3)



Pourriez-vous associer les divers composantes de l'équation:

$$\mathcal{E} \xrightarrow{\text{e.t.l.}} \mathcal{K}^{\circlearrowleft^{\mathsf{r}}} \xrightarrow{\text{infer.}} \mathcal{P}$$

... aux éléments de la figure ci-dessus ?

Travaux pratiques avec le NORIA-O dataset (1)

Objectif Combiner des données de topologie de réseaux (routeurs, serveurs, liens), d'événements (logs techniques, alarmes, tickets d'incidents) et d'organisation (intervenants, équipes) afin d'identifier les personnes à contacter pour résoudre un incident.

Données Ressources en ligne `noria-0.2` sur https://w3id.org/noria/dataset/

Outils Ce dont nous allons nous servir pour ce cours ...

- RMLmapper – construction de graphe de connaissances à l'aide de règles de transformation déclaratives : code source (image Docker)
- Jena ARC – interrogation de graphe en ligne de commande via des requêtes SPARQL : SPARQL CLI (aide mémoire)
- Editeurs graphique en ligne – visualiser, éditer et interroger un graphe de connaissances : Semantechs turtle-editor-viewer (Zazuko RDF Skecth)

Etapes Typiquement ...

1. Collecter les données à transformer → `dataset/noria-0.2/datasource/`
2. Implémenter des règles de transformation → `dataset/noria-0.2/engine/`
3. Exécuter les règles (produire le graphe) → `dataset/noria-0.2/output/`
4. Interroger le graphe, nous verrons ça un peu plus tard dans ce cours ...

Travaux pratiques avec le NORIA-O dataset (1)

Objectif Combiner des données de topologie de réseaux (routeurs, serveurs, liens), d'événements (logs techniques, alarmes, tickets d'incidents) et d'organisation (intervenants, équipes) afin d'identifier les personnes à contacter pour résoudre un incident.

Données Ressources en ligne `noria-0.2` sur https://w3id.org/noria/dataset/

Outils Ce dont nous allons nous servir pour ce cours ...

- RMLmapper – construction de graphe de connaissances à l'aide de règles de transformation déclaratives : code source (image Docker)
- Jena ARC – interrogation de graphe en ligne de commande via des requêtes SPARQL : SPARQL CLI (aide mémoire)
- Editeurs graphique en ligne – visualiser, éditer et interroger un graphe de connaissances : Semantechs turtle-editor-viewer (Zazuko RDF Skecth)

Etapes Typiquement ...

1. Collecter les données à transformer → `dataset/noria-0.2/datasource/`
2. Implémenter des règles de transformation → `dataset/noria-0.2/engine/`
3. Exécuter les règles (produire le graphe) → `dataset/noria-0.2/output/`
4. Interroger le graphe, nous verrons ça un peu plus tard dans ce cours ...

Travaux pratiques avec le NORIA-O dataset (1)

Objectif Combiner des données de topologie de réseaux (routeurs, serveurs, liens), d'événements (logs techniques, alarmes, tickets d'incidents) et d'organisation (intervenants, équipes) afin d'identifier les personnes à contacter pour résoudre un incident.

Données Ressources en ligne `noria-0.2` sur https://w3id.org/noria/dataset/

Outils Ce dont nous allons nous servir pour ce cours ...

- RMLmapper – construction de graphe de connaissances à l'aide de règles de transformation déclaratives : code source (image Docker)
- Jena ARC – interrogation de graphe en ligne de commande via des requêtes SPARQL : SPARQL CLI (aide mémoire)
- Editeurs graphique en ligne – visualiser, éditer et interroger un graphe de connaissances : Semantechs turtle-editor-viewer (Zazuko RDF Skecth)

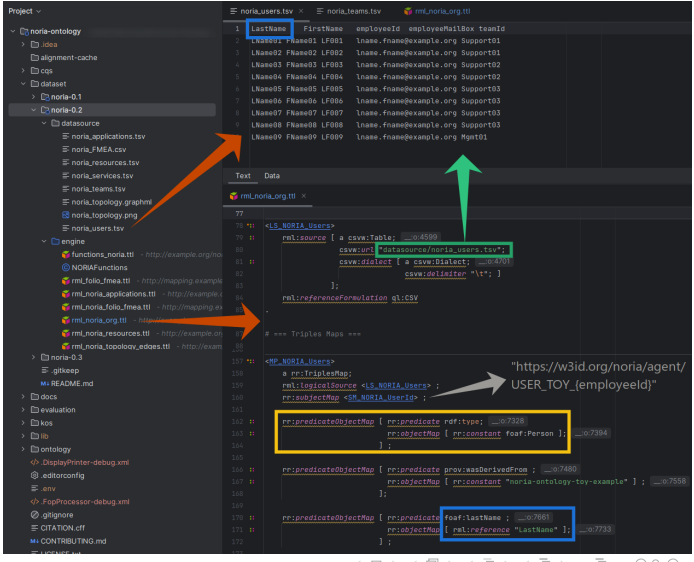
Etapes Typiquement ...

1. Collecter les données à transformer → `dataset/noria-0.2/datasource/`
2. Implémenter des règles de transformation → `dataset/noria-0.2/engine/`
3. Exécuter les règles (produire le graphe) → `dataset/noria-0.2/output/`
4. Interroger le graphe, nous verrons ça un peu plus tard dans ce cours ...

Travaux pratiques avec le NORIA-O dataset (2)

#1 – Collecter les données à transformer

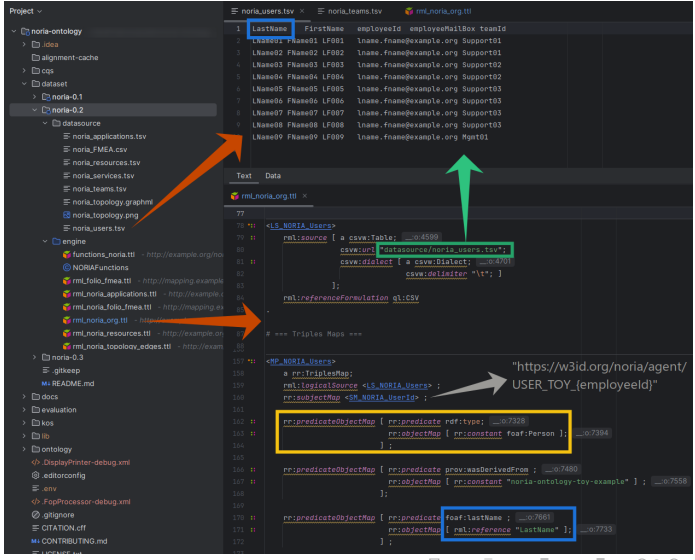
- Exemple : fichier `noria_users.tsv`
- Format : Tabulation-Separated Values (TSV)



Travaux pratiques avec le NORIA-O dataset (3)

#2 – Implémenter des règles de transformation

- Exemple : fichier `rml_noria_org.ttl`
- Syntaxe : Terse RDF Triple Language (Turtle)
- Vocabulaire : RDF Mapping Language (RML)
- Logical Source (vert) : référence au fichier contenant les données
- Triples Map (gris) : pour chaque enregistrement (ligne) de la Logical Source, on génère une entité (un nœud) dont l'identifiant sera de la forme `https://w3id.org/noria/-agent/USER_TOY_{employeeId}`
- Triples Map (jaune) : chaque entité disposera d'une propriété (un arc) pour lui associer la classe `foaf:Person`
- Triples Map (bleu) : chaque entité disposera d'une propriété de sémantique `foaf:lastName` dont la valeur sera issue de la colonne `LastName` de la Logical Source



Travaux pratiques avec le NORIA-O dataset (4)

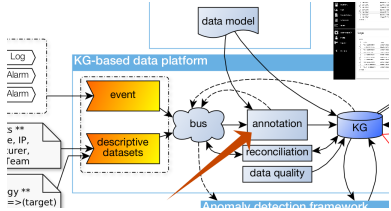
#3.1 – Exécuter les règles (produire le graphe)

- Exemple : fichier `rml_noria_org.ttl`
- Télécharger l'image Docker de RMLMapper :

```
# Get the rmlmapper docker image
docker pull rmlio/rmlmapper-java
```

- Exécuter RMLMapper en se focalisant sur le Triples Map `MP_NORIA_Users` :

```
# Run rmlmapper
docker run --rm -v $(pwd):/data rmlio/rmlmapper-java rmlmapper \
  -v -s turtle \
  -m engine/rml_noria_org.ttl \
  -o output/noria_org_test.ttl \
  -t http://example.org/noria/engine/MP_NORIA_Users
```



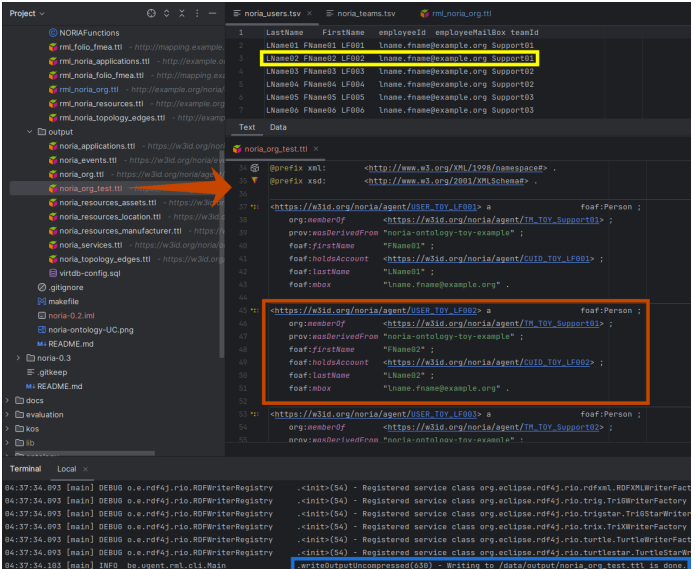
Nous analysons le graphe résultant (fichier `output/noria_org_test.ttl`) dans la diapo suivante ...

Travaux pratiques avec le NORIA-O dataset (5)

#3.2 – Exécuter les règles

- RMLMapper (bleu) : le fichier `output/noria_org_test.ttl` a été produit
- Un enregistrement du fichier source (jaune) est traduit en un ensemble de triplets (rouge)
- Triplet $\equiv \langle$sujet, prédicat, objet$\rangle$
  - Sujet $\simeq$ nœud source
  - Prédicat $\simeq$ arc
  - Objet $\simeq$ nœud cible
- Dans l'exemple encadré (rouge)
  - Sujet = `https://w3id.org/noria/-agent/USER_TOY_LF002`
  - ObjectProperty : un arc qui pointe vers un nœud, p.ex. `org:memberOf`
  - DatatypeProperty : un arc qui pointe vers une valeur, p.ex. `foaf:firstName`

→ Est-ce que cela forme un graphe ?

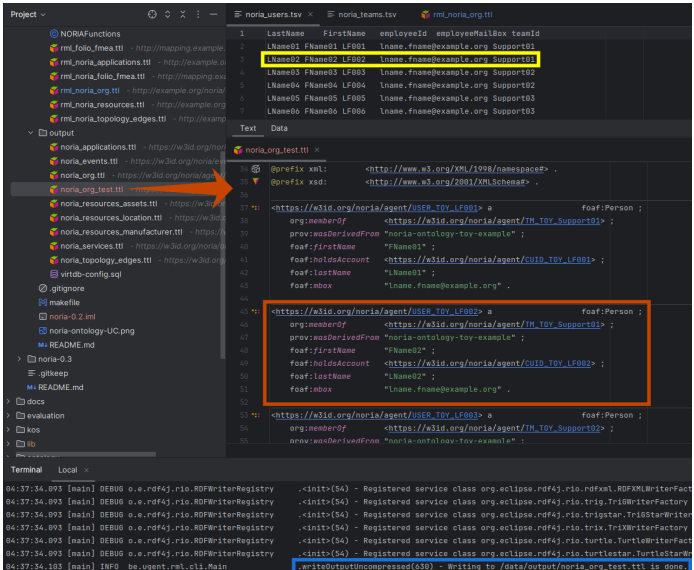


Travaux pratiques avec le NORIA-O dataset (5)

#3.2 – Exécuter les règles

- RMLMapper (bleu) : le fichier `output/noria_org_test.ttl` a été produit
- Un enregistrement du fichier source (jaune) est traduit en un ensemble de triplets (rouge)
- Triplet $\equiv$ ⟨sujet, prédicat, objet⟩
  - Sujet $\simeq$ nœud source
  - Prédicat $\simeq$ arc
  - Objet $\simeq$ nœud cible
- Dans l'exemple encadré (rouge)
  - Sujet = `https://w3id.org/noria/-agent/USER_TOY_LF002`
  - ObjectProperty : un arc qui pointe vers un nœud, p.ex. `org:memberOf`
  - DatatypeProperty : un arc qui pointe vers une valeur, p.ex. `foaf:firstName`

→ Est-ce que cela forme un graphe ?



Travaux pratiques avec le NORIA-O dataset (6)

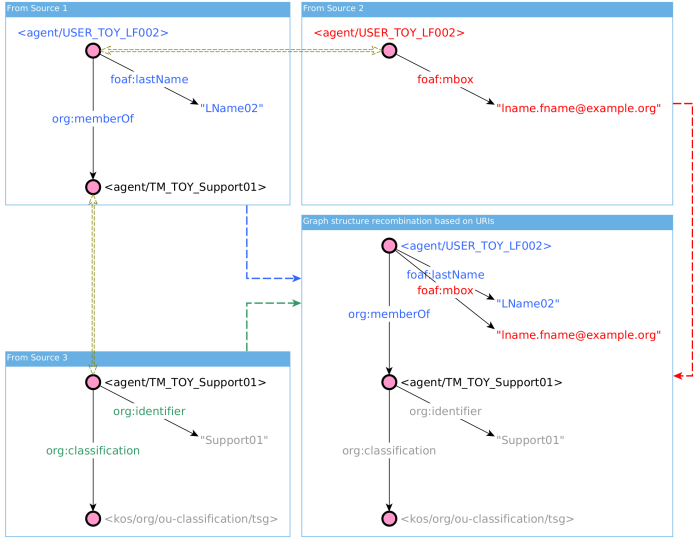
#3.3 – Exécuter les règles

- Exemple complet : exécuter RMLMapper sur l'ensemble des règles à disposition (`rml_noria_applications.ttl`, `rml_noria_org.ttl`, `rml_noria_resources.ttl`, etc.), puis concaténer les fichiers graphes résultants en un seul et même fichier:

```
# Call all RML rules
make build-kg

# Concatenate graph files
cat $(ls output/noria_*.ttl) \
  > output/noria-full.ttl
```

→ L'utilisation d'Uniform Resource Identifiers (URIs) cohérents à travers les règles/processus de construction garantit une déduplication des informations lors d'une agrégation de graphes (analogue à un système multi-silos). Il s'agît de la mise en application de la Unique Name Assumption (UNA).



Travaux pratiques avec le NORIA-O dataset (6)

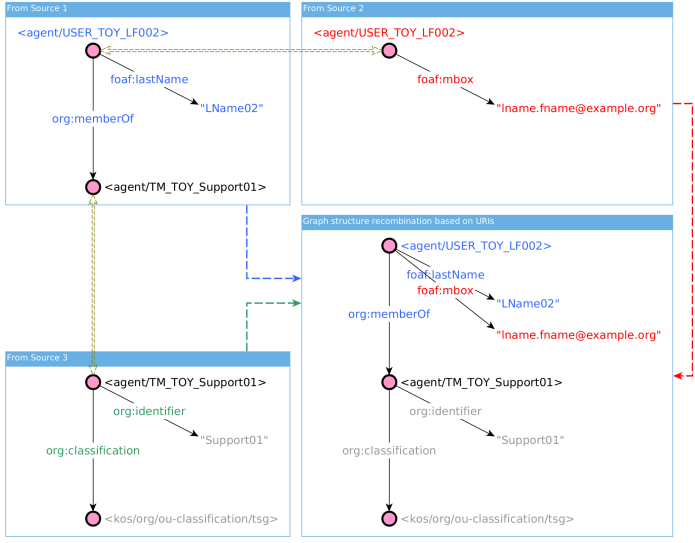
#3.3 – Exécuter les règles

- Exemple complet : exécuter RMLMapper sur l'ensemble des règles à disposition (`rml_noria_applications.ttl`, `rml_noria_org.ttl`, `rml_noria_resources.ttl`, etc.), puis concaténer les fichiers graphes résultants en un seul et même fichier:

```
# Call all RML rules
make build-kg

# Concatenate graph files
cat $(ls output/noria_*.ttl) \
  > output/noria-full.ttl
```

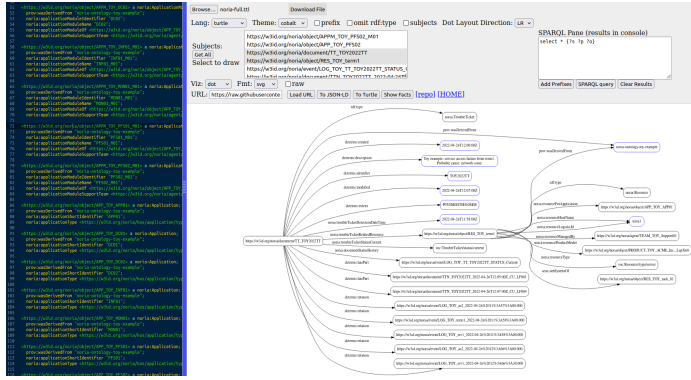
→ L'utilisation d'Uniform Resource Identifiers (URIs) cohérents à travers les règles/processus de construction garantit une déduplication des informations lors d'une agrégation de graphes (analogue à un système multi-silos). Il s'agît de la mise en application de la Unique Name Assumption (UNA).



Travaux pratiques avec le NORIA-O dataset (7)

#4.1 – Interroger le graphe

- Exemple : fichier `noria-full.ttl` produit en #3.3
- Approche : utiliser un a priori visuel en observant/explorant le graphe, p.ex. ici en chargeant le ficher de l'étape précédente dans l'outil en ligne Semantechs turtle-editor-viewer.

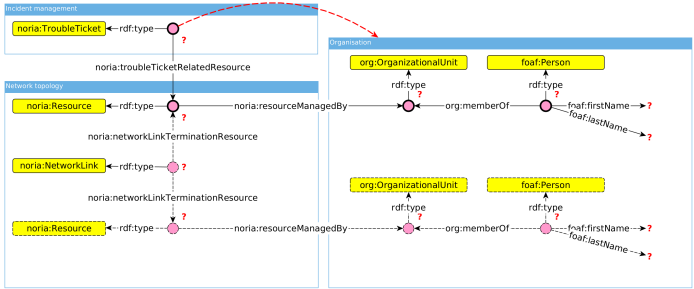


Travaux pratiques avec le NORIA-O dataset (8)

#4.2 – Interroger le graphe

→ Approche : utiliser une requête SPARQL relative à un cas d'application (un besoin métier), la requête correspondant à un motif de graphe à identifier.

- Exemple "identifier les personnes à contacter pour résoudre un incident" :




```
# SPARQL
SELECT ?tt ?res ?tech ?res2 ?tech2
WHERE {
    ?tt rdf:type noria:TroubleTicket ;
        noria:troubleTicketRelatedResource ?res .
    ?res rdf:type noria:Resource ;
        noria:resourceManagedBy ?org .
    ?tech org:memberOf ?org ;
        foaf:firstName ?tech_fname ;
        foaf:lastName ?tech_lname .

  OPTIONAL {
    ?nl rdf:type noria:NetworkLink ;
        noria:networkLinkTerminationResource ?res ;
        noria:networkLinkTerminationResource ?res2 .
    FILTER (?res != ?res2)
    ?res2 rdf:type noria:Resource ;
        noria:resourceManagedBy ?org2 .
    ?tech2 org:memberOf ?org2 ;
        foaf:firstName ?tech_fname ;
        foaf:lastName ?tech_lname . }
}
```

Travaux pratiques avec le NORIA-O dataset (9)

#4.3 – Interroger le graphe

- Exemple : fichier `noria-full.ttl` produit en #3.3

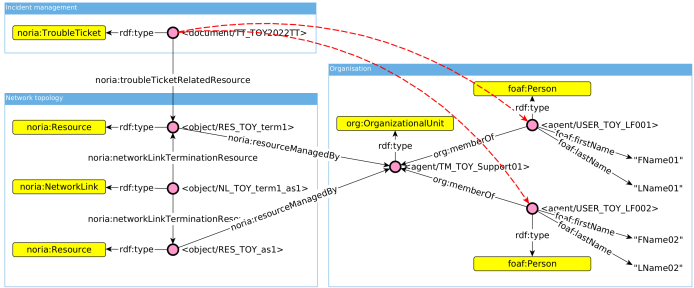
→ Exécuter la requête SPARQL de #4.2 à l'aide de Jena SPARQL CLI ou Semantechs turtle-editor-viewer.

```
# JENA CLI
~/.local/jena/bin/sparql \
  --query ./rq_tt_org.sparql \
  --data ./output/noria-full.ttl
```

- Les entités du graphe satisfaisant le motif de graphe (la forme décrite dans la requête SPARQL en #4.2) sont retournées.

→ Sur ces bases, quelle autres approches & techniques imaginez-vous ?

→ ... quelques pistes que vous pourriez explorer : raisonnement automatique (FOLIO, CFGOwl), caractéristiques du graphe (kglab, statistical relational learning), plongement de graphes (pyRDF2Vec, INK), etc.



Travaux pratiques avec le NORIA-O dataset (9)

#4.3 – Interroger le graphe

- Exemple : fichier `noria-full.ttl` produit en #3.3

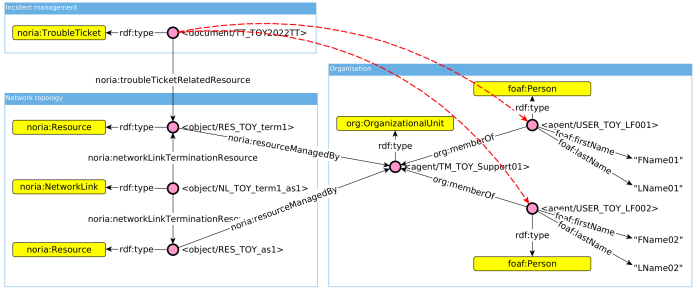
→ Exécuter la requête SPARQL de #4.2 à l'aide de Jena SPARQL CLI ou Semantechs turtle-editor-viewer.

```
# JENA CLI
~/.local/jena/bin/sparql \
  --query ./rq_tt_org.sparql \
  --data ./output/noria-full.ttl
```

- Les entités du graphe satisfaisant le motif de graphe (la forme décrite dans la requête SPARQL en #4.2) sont retournées.

→ Sur ces bases, quelle autres approches & techniques imaginez-vous ?

→ ... quelques pistes que vous pourriez explorer : raisonnement automatique (FOLIO, CFGOwl), caractéristiques du graphe (kglab, statistical relational learning), plongement de graphes (pyRDF2Vec, TNK), etc.



Travaux pratiques avec le NORIA-O dataset (9)

#4.3 – Interroger le graphe

- Exemple : fichier `noria-full.ttl` produit en #3.3

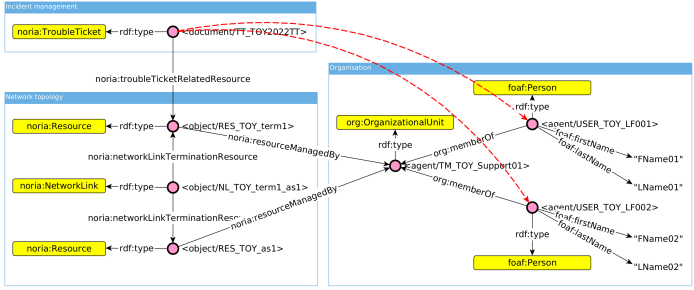
→ Exécuter la requête SPARQL de #4.2 à l'aide de Jena SPARQL CLI ou Semantechs turtle-editor-viewer.

```
# JENA CLI
~/.local/jena/bin/sparql \
  --query ./rq_tt_org.sparql \
  --data ./output/noria-full.ttl
```

- Les entités du graphe satisfaisant le motif de graphe (la forme décrite dans la requête SPARQL en #4.2) sont retournées.

→ Sur ces bases, quelle autres approches & techniques imaginez-vous ?

→ ... quelques pistes que vous pourriez explorer : raisonnement automatique (FOLIO, CFGOwl), caractéristiques du graphe (kglab, statistical relational learning), plongement de graphes (pyRDF2Vec, INK), etc.



Projet

Présentation du projet étudiant #2, une étude sur la maîtrise de la taille des graphes de connaissances.

Présentation du projet (1) – Problématique et objectif

Problématique L'état de l'art sur les techniques de construction de graphe de connaissances montre globalement, à date, la maturité des approches par transformation de données source en triplets et l'**insertion des triplets** dans une base de données orientée graphe. Or travailler uniquement en insertion induit que **le graphe de connaissance ne cesse de croître** dès que les données varient à la source. Pour exemples : un bagôt d'interface réseau ou un décommissionnement d'équipement réseau correspondront à l'ajout et la liaison d'entités-événements représentatifs des changements d'états.

Objectif Proposition et critique de stratégies pour maîtriser l'ordre (le nombre de sommets) et la taille (le nombre d'arrêtes) d'un graphe de connaissances.

Intuition "l'élagage du graphe" (pruning) sera nécessaire à un moment donné en fonction,

1. Des politiques de conservation des données,
2. De la précision des outils de détection-classification qui s'entraîneront sur / consommeront le graphe de connaissances.

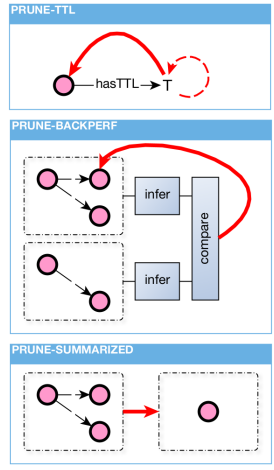
Présentation du projet (2) – Stratégies d'élagage de graphe

Les stratégies suivantes ont émergées lors de discussions autour des présentations "Anomaly detection for telco companies" (KGCW 2024) et "I want a KG-based Digital Twin" (KG4DI 2024) :

PRUNE-TTL Supprimer des triplets en fonction d'une propriété "temps de vie" (TTL) ou d'un horodatage (p.ex. en raison de réglementations sur la conservation des données).

PRUNE-BACKPERF Supprimer des triplets lorsque nous constatons que les oublier lors de l'entraînement n'affecte pas la performance d'un classificateur (p.ex. un modèle de détection d'anomalies reposant sur les graph embeddings) ou n'affecte pas l'exactitude d'un raisonneur automatique.

PRUNE-SUMMARIZED Résumer les données historiques comme moyen de compression (p.ex. en substituant des sous-graphes par des embeddings représentatifs ou une synthèse produite par un Large Language Model).



Présentation du projet (3) – Deux activités complémentaires à réaliser

Activité 1 **Lecture & résumé** de papiers de recherche.

- L. Tailhardat, et al. 2023. “Designing NORIA: a Knowledge Graph-based Platform for Anomaly Detection and Incident Management in ICT Systems.” ceur-ws.org/Vol-3471/paper3.pdf ⇔ une architecture de construction de graphe de connaissances à partir de 15 sources de données dans le domaine des telcos.
- D. V. Assche, et al. 2024. “IncRML: Incremental Knowledge Graph Construction from Heterogeneous Data Sources.” SWJ 3790-5004 ⇔ une méthode pour cibler les modifications d’un graphe de connaissances.
- L. Tailhardat, et al. 2023. “Leveraging Knowledge Graphs For Classifying Incident Situations in ICT Systems.” 10.1145/3600160.3604991 (open access) ⇔ deux approches pour détecter des anomalies dans le domaine des telcos à partir de données dans un graphe de connaissances.

Activité 2 **Analyse & discussion** des stratégies d’élagage : quel algorithme permettrait de décrire le fonctionnement de chaque stratégie ? quels outils utiliseriez-vous pour leur implémentation ? comment évalueriez-vous l’efficacité de chaque implémentation ? quels sont les gains théoriques en nombre de triplets de chaque stratégie ? quels sont les avantages, inconvénients, et limites a priori de chaque stratégie ?

Evaluation **Synthèse écrite + présentation orale**. Production individuelle ou en groupe. Critères : 1) clarté du propos ; 2) argumentation motivée et justifiée, notamment avec des exemples (réels ou inventés), ainsi que des références lorsqu’elles existent ; 3) un relecteur doit être mis en capacité de prendre une décision, en fonction du contexte métier qui est le sien et à partir des informations présentées, sur la stratégie à adopter.

Merci !

lionel.tailhardat@orange.com