

Web de données et Web sémantique

D'où vient le Web sémantique ? Du Web de documents au Web de données

Cours Web de données

Maman Sani ABOUBACAR
(Cours adapté des diapositives : HAI824 de K. Todorov)

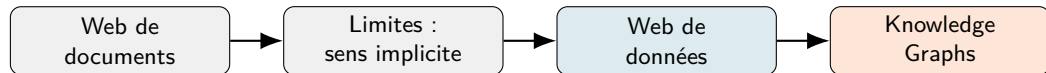
Septembre, 2026

Objectifs de la séance

À la fin de cette séance, vous devez pouvoir :

- expliquer **d'où vient** le Web sémantique ;
- distinguer le **Web de documents** du **Web de données** ;
- comprendre le rôle des **URI**, de **RDF**, des **vocabulaires** et des **ontologies** ;
- expliquer l'intérêt du Web sémantique pour la recherche, l'intégration de données et l'IA ;
- représenter un petit texte sous forme de **triplets RDF** et de graphe.

- ① **The Big Picture** : pourquoi le Web sémantique ?
- ② **The WWW** : le Web de documents, son architecture et ses limites
- ③ **Limites du Web de documents** : volume, analyse, représentation
- ④ **Vision du Web sémantique** : RDF, URI, vocabulaires, ontologies
- ⑤ **Linked Open Data et FAIR** : rendre les données connectables et réutilisables
- ⑥ **Exercice** : décrire un film comme données RDF



Idée centrale

Le Web sémantique ne remplace pas le Web. Il ajoute une couche permettant de décrire les **choses**, leurs **types** et leurs **relations**, afin que les machines puissent mieux les exploiter.

Part I

The Big Picture

From hyperlinks between documents to relationships between entities

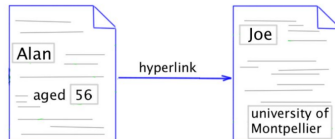


Le Web classique : des liens entre documents

Unité principale

La page Web ou le document HTML.

- un document pointe vers un autre ;
- le lien est navigable ;
- le sens du lien reste souvent implicite.

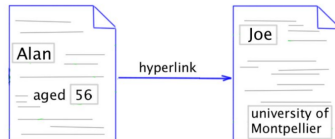


Le Web classique : des liens entre documents

Unité principale

La page Web ou le document HTML.

- un document pointe vers un autre ;
- le lien est navigable ;
- le sens du lien reste souvent implicite.



Question

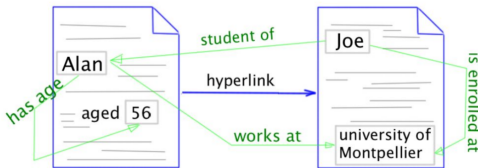
Que signifie réellement ce lien ? Ami ? collègue ? étudiant ? page similaire ?

Web de données : des relations entre entités

Changement d'unité

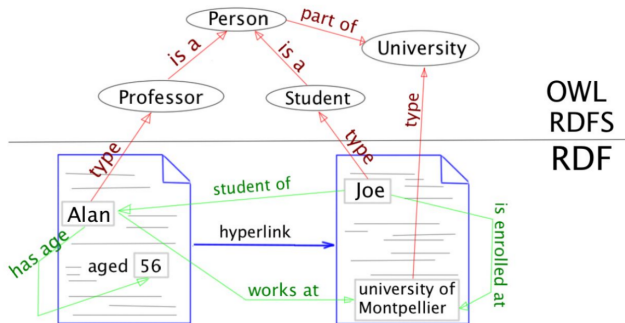
On ne relie plus seulement des documents ; on relie des **entités du monde réel**.

- personnes ;
- organisations ;
- lieux ;
- oeuvres ;
- événements.



RDF

Plus de sémantique : types, classes, ontologies

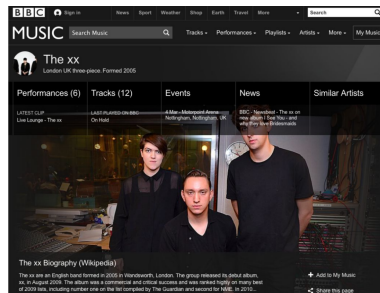


Ce que l'on ajoute

Des relations typées, des ressources typées et des vocabulaires partagés pour dire ce que les choses signifient.

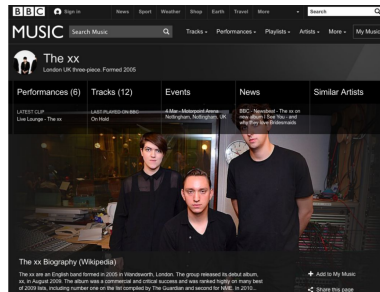
Exemple 1 : BBC Music (2008–2020)

- Artistes, albums, morceaux, événements et biographies proviennent de sources différentes.
- Le Web sémantique aide à relier ces objets au lieu de gérer uniquement des pages séparées.
- L'utilisateur navigue dans un réseau d'entités musicales.



Exemple 1 : BBC Music (2008–2020)

- Artistes, albums, morceaux, événements et biographies proviennent de sources différentes.
- Le Web sémantique aide à relier ces objets au lieu de gérer uniquement des pages séparées.
- L'utilisateur navigue dans un réseau d'entités musicales.



Message

Le site n'est pas seulement une collection de pages : il exploite un réseau de données reliées.

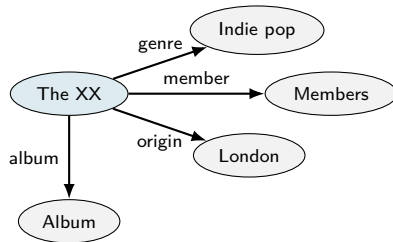
Exemple 2 : Google Knowledge Graph

Recherche par documents

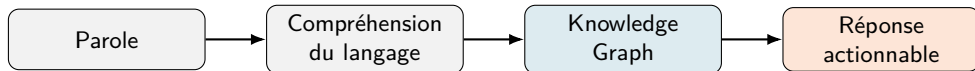
Retourner des pages contenant les mots de la requête.

Recherche par entités

Comprendre qu'une requête concerne une personne, une organisation, un film, un lieu, une date, etc.



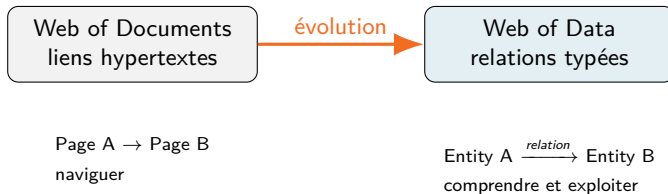
Exemple 3 : Siri, assistants vocaux et IA



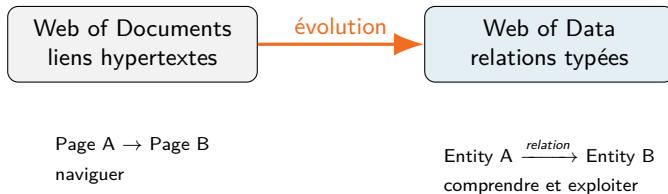
Pourquoi la sémantique est utile ?

Un assistant doit relier des intentions, des personnes, des lieux, des événements et des services. Il a besoin de connaissances structurées, pas seulement de documents.

En résumé : le big picture



En résumé : le big picture



Question suivante

Pour comprendre pourquoi cette évolution est nécessaire, il faut revenir à l'architecture du World Wide Web.

Part II

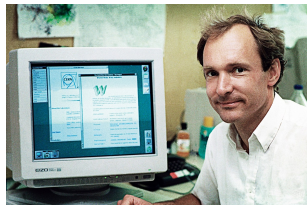
The World Wide Web

A brilliant Web designed primarily around documents



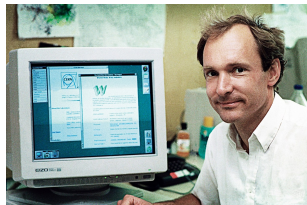
Origine : Tim Berners-Lee et le CERN

- Le Web naît à la fin des années 1980 au CERN.
- L'idée : relier des documents et des informations distribuées.
- Trois briques deviennent fondamentales : **URI**, **HTTP**, **HTML**.



Origine : Tim Berners-Lee et le CERN

- Le Web naît à la fin des années 1980 au CERN.
- L'idée : relier des documents et des informations distribués.
- Trois briques deviennent fondamentales : **URI**, **HTTP**, **HTML**.



Point clé

Le Web a d'abord été conçu pour partager et consulter des documents dans un environnement décentralisé.

Les principes du Web de documents

- **Page Web** comme unité de publication.
- **Hyperlien** comme mécanisme de navigation.
- **Décentralisation** : pas de dépôt central unique.
- **Liberté** : chacun peut publier et créer des liens.

Les principes du Web de documents

- **Page Web** comme unité de publication.
- **Hyperlien** comme mécanisme de navigation.
- **Décentralisation** : pas de dépôt central unique.
- **Liberté** : chacun peut publier et créer des liens.

Conséquence

Le Web est très robuste et scalable, mais le sens des contenus reste généralement interprété par les humains.

Les principes du Web de documents

- **Page Web** comme unité de publication.
- **Hyperlien** comme mécanisme de navigation.
- **Décentralisation** : pas de dépôt central unique.
- **Liberté** : chacun peut publier et créer des liens.

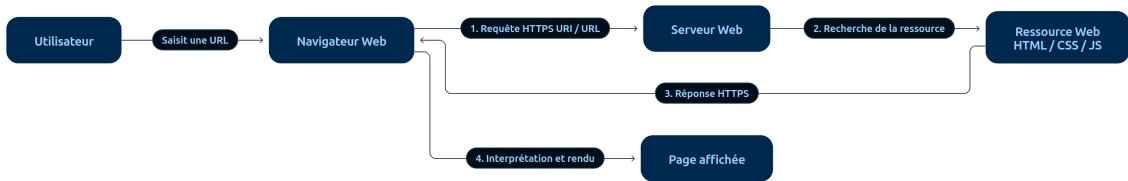
Conséquence

Le Web est très robuste et scalable, mais le sens des contenus reste généralement interprété par les humains.

À retenir

Internet $\neq$ Web : le Web est un service construit au-dessus d'Internet.

Architecture simplifiée du Web



Web content : conçu pour les humains

Constat

Même lorsque le contenu vient d'une base de données, il est souvent transformé en page HTML lisible par un humain.

Dans la base

Entité	Valeur
Centre	Montpellier Physiotherapy Centre
Métier	Physiotherapy
Therapist	Lisa Davenport
Secretary	John Reeds

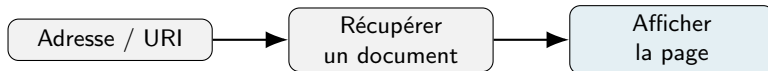
Sur le Web

- une page ;
- des titres ;
- des paragraphes ;
- des liens.

Problème

La structure métier originale disparaît partiellement dans la présentation HTML.

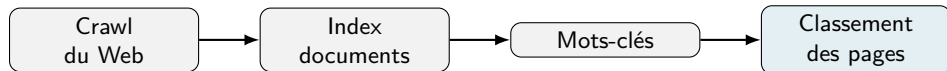
Browser : accéder et afficher



Limite

Un navigateur sait très bien rendre une page. Mais il ne sait pas nécessairement que *Lisa Davenport* est une thérapeute, que *John Reeds* est un secrétaire, ou que *9 :00-18 :00* est une plage horaire d'ouverture.

Search engines : indexer et classer des documents



Exemple de limite

Rechercher des pages contenant `knowledge graph` est différent de demander : “Quels chercheurs ont publié avec Tim Berners-Lee sur le Web sémantique ?”

Part III

Limits of the Web of Documents

Volume, analysis and representation problems



Problème 1 : le volume



Data deluge

Le problème n'est plus seulement d'accéder à l'information : il faut la trouver, la relier, l'analyser, la réutiliser et la contextualiser.

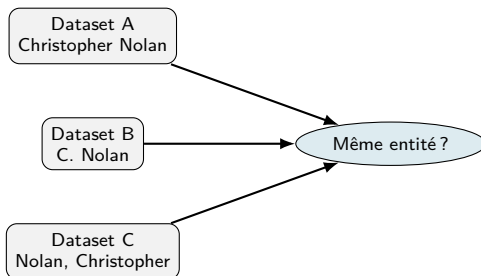
Problème 2 : l'analyse et l'intégration

Documents

Chaque site publie dans son propre format, avec son propre vocabulaire et ses propres pages.

Données

Pour analyser, il faut relier des entités identiques ou proches malgré des formats différents.

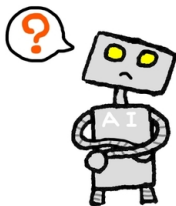


Problème 3 : la représentation

- Les données sur le Web sont créées par et pour les humains. Enfin, c'était le cas autrefois (cf. l'IA générative)
- Elle est donc dominée par des données non structurées ou semi-structurées (textes, images, vidéos, etc.), reliées les unes aux autres, et. . .
- ... compréhensible pour les humains, mais pas pour les machines

Problème 3 : la représentation

- Les données sur le Web sont créées par et pour les humains. Enfin, c'était le cas autrefois (cf. l'IA générative)
- Elle est donc dominée par des données non structurées ou semi-structurées (textes, images, vidéos, etc.), reliées les unes aux autres, et...
- ... compréhensible pour les humains, mais pas pour les machines
- Que veulent dire ces humains ?



Problème 3 : la représentation HTML

Le contenu du Web est largement formaté en utilisant le HTML (Hypertext Markup Language).

Qui est essentiellement structuré pour les humains, et non pour la machine, avec un texte structuré et non les données.

Problème 3 : la représentation HTML

Le contenu du Web est largement formaté en utilisant le HTML (Hypertext Markup Language).

Qui est essentiellement structuré pour les humains, et non pour la machine, avec un texte structuré et non les données.

```
<h1>Montpellier Physiotherapy Centre</h1>
<p>Welcome to homepage of our center. Let our staff Lisa Marc
and Dupond Arnaud (our secretary) take care of you.</p>
<h2>Consultation hours</h2>
<ul>
  <li>Monday to Friday: 9:00-18:00</li>
  <li>Saturday: 10:00-14:00</li>
</ul>
<p>But we do not offer consultation during the weeks of the <a href="...">School Holidays</a></p>
```

La représentation : Humain vs machine

Aucun problème pour les humains.

Plusieurs problèmes pour la machine :

- un thérapeute, une secrétaire... c'est quoi la différence ?
- Les heures de consultation ? Quand est-ce que les horaires de vacances commencent ?
- Le centre de consultation et les horaires de consultations sont ils de mêmes entités ?

=> Impossibilité d'exprimer les relations entre les choses.

La représentation : Humain vs machine

Aucun problème pour les humains.

Plusieurs problèmes pour la machine :

- un thérapeute, une secrétaire... c'est quoi la différence ?
- Les heures de consultation ? Quand est-ce que les horaires de vacances commencent ?
- Le centre de consultation et les horaires de consultations sont ils de mêmes entités ?

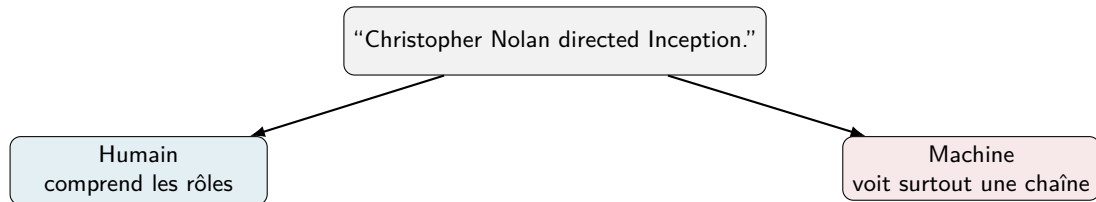
=> Impossibilité d'exprimer les relations entre les choses.

La question clé

À quoi servent les données, si ce n'est à décrire des relations ?

La représentation : Humain vs machine

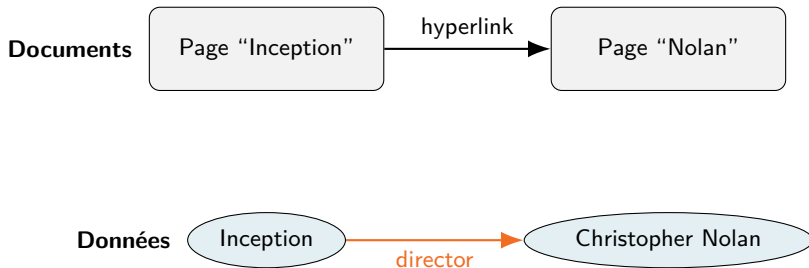
Un autre exemple :



La question clé

Et si les relations entre les choses faisaient elles-mêmes partie des données ?

Solution : Web of Documents -> Web of Data ?



Part IV

The Vision of the Semantic Web

Make meaning explicit and data linkable



Définition intuitive du Web sémantique

Web sémantique

Extension du Web dans laquelle les données sont décrites de manière suffisamment explicite pour permettre leur identification, leur connexion, leur intégration et leur traitement automatique.

Définition intuitive du Web sémantique

Web sémantique

Extension du Web dans laquelle les données sont décrites de manière suffisamment explicite pour permettre leur identification, leur connexion, leur intégration et leur traitement automatique.

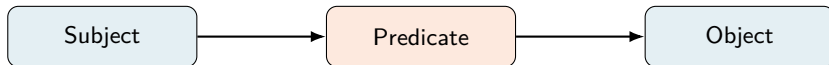
Avant

Documents reliés par des liens hypertextes.

Après

Entités reliées par des relations typées.

RDF : Resource Description Framework



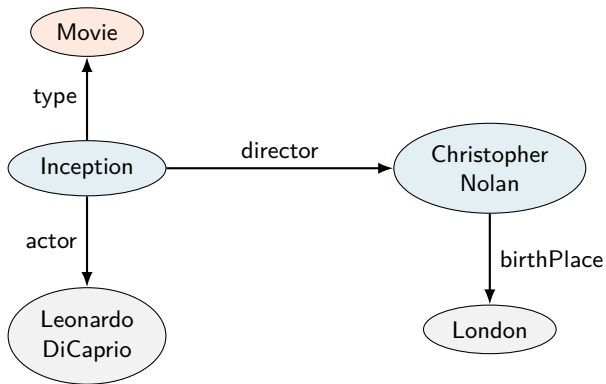
Exemple

Inception - director - Christopher Nolan

Idée

RDF représente les faits comme des triplets. Un ensemble de triplets forme un graphe.

Plusieurs triplets forment un graphe



URI : identifier les choses sans ambiguïté

Ambiguïté

Paris peut désigner une ville, une personne, un film ou une organisation.

URI : identifier les choses sans ambiguïté

Ambiguïté

Paris peut désigner une ville, une personne, un film ou une organisation.

Solution

Utiliser des identifiants globaux pour nommer les ressources.

URI : identifier les choses sans ambiguïté

Ambiguïté

Paris peut désigner une ville, une personne, un film ou une organisation.

Solution

Utiliser des identifiants globaux pour nommer les ressources.

Exemple conceptuel

`https://dbpedia.org/resource/Paris`

URI : identifier les choses sans ambiguïté

Ambiguïté

Paris peut désigner une ville, une personne, un film ou une organisation.

Solution

Utiliser des identifiants globaux pour nommer les ressources.

Exemple conceptuel

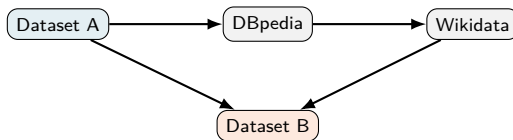
<https://dbpedia.org/resource/Paris>

Message

Le Web sémantique utilise les mécanismes du Web pour identifier et relier les choses, pas seulement les pages qui en parlent.

Les 4 principes du Linked Data

- 1 Utiliser des **URI** pour nommer les choses.
- 2 Utiliser des **URI HTTP** pour que ces noms soient consultables.
- 3 Fournir des informations utiles avec des standards comme **RDF** et **SPARQL**.
- 4 Inclure des liens vers d'autres URI pour découvrir d'autres choses.



Vocabulaires : partager les mêmes termes

Sans vocabulaire commun

ex:director
ex:directedBy
ex:filmmaker

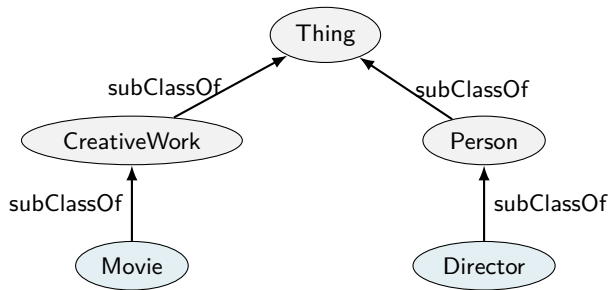
Avec vocabulaire partagé

schema:director
schema:Person
schema:Movie

Exemples de vocabulaires

Schema.org, FOAF, Dublin Core, SKOS.

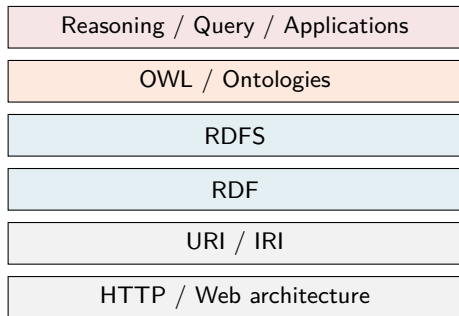
Ontologies : expliciter le sens



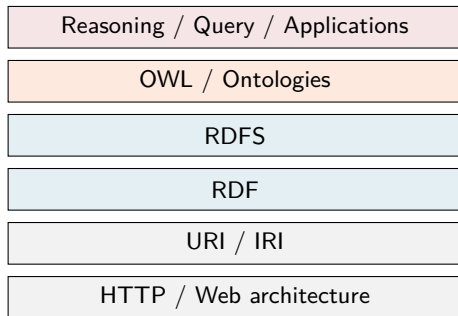
Différence simple

Un vocabulaire donne des termes ; une ontologie ajoute des relations formelles entre ces termes.

La pile du Web sémantique



La pile du Web sémantique



Point clé

Le Web sémantique est construit au-dessus du Web existant : il ajoute une couche de données et de sens.

Part V

Linked Open Data and FAIR

From open datasets to reusable connected knowledge

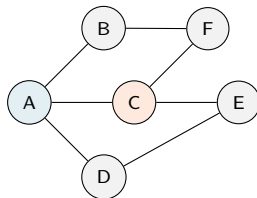


Linked Open Data : publier et relier les jeux de données

Objectif

Publier des données ouvertes de façon à ce qu'elles puissent être reliées, interrogées et réutilisées par d'autres.

- datasets distribués ;
- identifiants communs ;
- liens entre ressources ;
- graphe global de données.



Linked Open Data

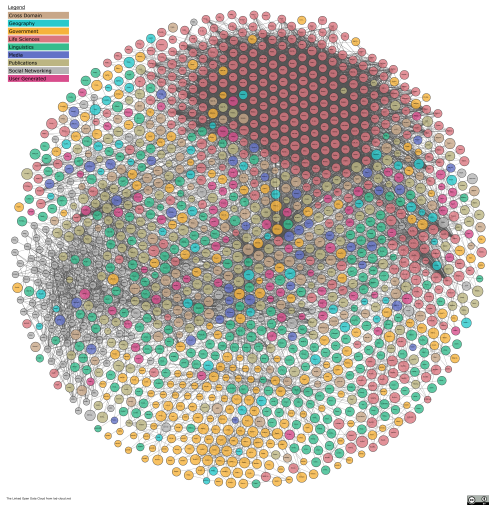


Figure – source : <https://lod-cloud.net/>

Open Data -> Linked Open Data : modèle 5 étoiles

Niveau	Principe	Exemple
★	disponible sur le Web	PDF, image, page
★★	données structurées	tableur
★★★	format ouvert	CSV, JSON
★★★★	URI pour identifier les choses	ressources nommées
★★★★★	données liées à d'autres données	Linked Data

Open Data -> Linked Open Data : modèle 5 étoiles

Niveau	Principe	Exemple
★	disponible sur le Web	PDF, image, page
★★	données structurées	tableur
★★★	format ouvert	CSV, JSON
★★★★	URI pour identifier les choses	ressources nommées
★★★★★	données liées à d'autres données	Linked Data

Message

Plus les données sont structurées, identifiées et liées, plus elles deviennent réutilisables par des humains, des logiciels et des agents IA.

F**Findable****A****Accessible****I****Interoperable****R****Reusable**

Important

FAIR et Open Data ne sont pas synonymes. Une donnée peut être FAIR sans être ouverte. Mais les principes du Web sémantique aident fortement à rendre les données trouvables, interopérables et réutilisables.

Synthèse : pourquoi le Web sémantique est pertinent

- ① Il rend les entités et les relations **explicites**.
- ② Il facilite l'**intégration** de données hétérogènes.
- ③ Il permet de publier des données **liables** et **réutilisables**.
- ④ Il prépare les bases des **Knowledge Graphs**.
- ⑤ Il donne un socle utile pour le **RAG**, le **GraphRAG** et les systèmes agentiques.

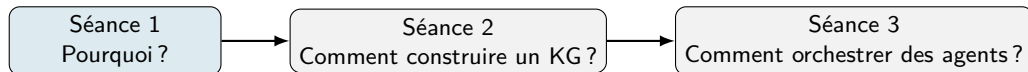
Synthèse : pourquoi le Web sémantique est pertinent

- ① Il rend les entités et les relations **explicites**.
- ② Il facilite l'**intégration** de données hétérogènes.
- ③ Il permet de publier des données **liables** et **réutilisables**.
- ④ Il prépare les bases des **Knowledge Graphs**.
- ⑤ Il donne un socle utile pour le **RAG**, le **GraphRAG** et les systèmes agentiques.

Idée clé

Les documents nous indiquent où aller. Les données sémantiques indiquent aux machines ce que sont les choses et quels liens les unissent.

Transition vers la suite du module



Aujourd'hui

Comprendre l'origine, la motivation et les grands concepts : Web, documents, données, RDF, vocabulaires, ontologies, Linked Data et FAIR.

Part VI

Exercise

Describe a Movie as RDF Data



Exercice : Describe a Movie as RDF Data

Texte

The movie *Inception* was directed by *Christopher Nolan* and released in *2010*. It stars *Leonardo DiCaprio* and won *four Oscars*. *Christopher Nolan* was born in *London* in *1970*.

Objectif

Transformer un court texte en données RDF : entités, relations, triplets, graphe et vocabulaire.

Étape 1 : identifier les entités

Questions

- De quelles choses parle le texte ?
- Quelles choses peuvent devenir des sujets ou objets RDF ?
- Quels types peut-on leur associer ?

Étape 1 : identifier les entités

Questions

- De quelles choses parle le texte ?
- Quelles choses peuvent devenir des sujets ou objets RDF ?
- Quels types peut-on leur associer ?

Entités

- Inception : Movie
- Christopher Nolan : Person
- Leonardo DiCaprio : Person
- London : City
- Oscar : Award

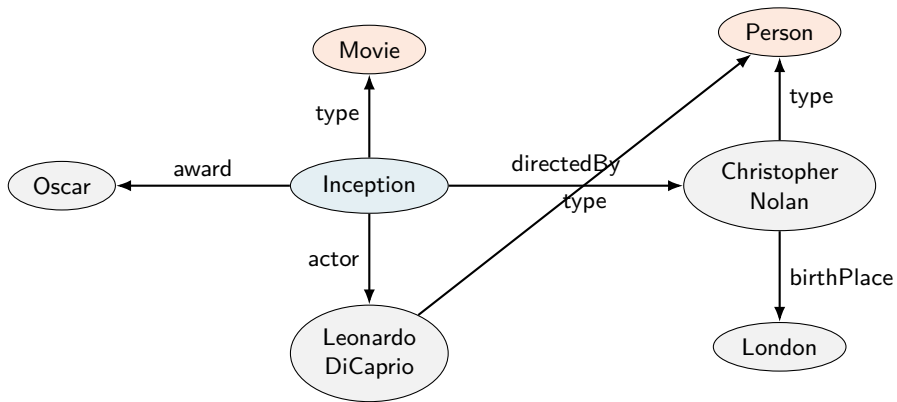
Étape 2 : identifier les relations

Sujet	Relation	Objet
Inception	directedBy	Christopher Nolan
Inception	releaseYear	2010
Inception	actor	Leonardo DiCaprio
Inception	award	Oscar
Christopher Nolan	birthPlace	London
Christopher Nolan	birthYear	1970

Consigne

Compléter jusqu'à obtenir au moins 8 triplets, en ajoutant notamment les types : Movie, Person, City, Award.

Étape 3 : dessiner le graphe



Étape 4 : écrire en Turtle

```
@prefix ex: <http://example.org/movie/> .  
@prefix schema: <https://schema.org/> .
```

```
ex:Inception a schema:Movie ;  
    schema:name "Inception" ;  
    schema:director ex:ChristopherNolan ;  
    schema:actor ex:LeonardoDiCaprio ;  
    schema:datePublished "2010" .
```

```
ex:ChristopherNolan a schema:Person ;  
    schema:birthPlace ex:London ;  
    schema:birthDate "1970" .
```

Étape 5 : mapping vers des vocabulaires connus

Concept local	Vocabulaire conseillé	Pourquoi ?
ex :Movie	schema :Movie	type d'oeuvre audiovisuelle
ex :Person	schema :Person ou foaf :Person	personne
ex :City	schema :City ou schema :Place	lieu
ex :directedBy	schema :director	réalisateur
ex :actor	schema :actor	acteur
ex :birthPlace	schema :birthPlace	lieu de naissance

Consigne

Pour chaque classe ou propriété inventée, chercher si un terme équivalent existe dans Schema.org ou FOAF.

Étape 6 : question Linked Data

Comment relier votre ressource ex:ChristopherNolan
à une ressource externe comme Wikidata ou DBpedia ?

Étape 6 : question Linked Data

Comment relier votre ressource `ex:ChristopherNolan` à une ressource externe comme Wikidata ou DBpedia ?

Solution

Ajouter un lien vers une URI externe pour enrichir le graphe et le connecter au Web de données.

Exemple conceptuel

```
ex:ChristopherNolan owl:sameAs <URI externe> .
```

Étape 6 : question Linked Data

Comment relier votre ressource `ex:ChristopherNolan` à une ressource externe comme Wikidata ou DBpedia ?

Solution

Ajouter un lien vers une URI externe pour enrichir le graphe et le connecter au Web de données.

Exemple conceptuel

```
ex:ChristopherNolan owl:sameAs <URI externe> .
```

Discussion

Qu'est-ce que le graphe rend explicite que le texte ne rendait pas directement exploitable par une machine ?

- Le Web classique relie principalement des **documents**.
- Le Web sémantique cherche à relier des **entités** par des **relations typées**.
- RDF représente les faits sous forme de **triplets**.
- Les vocabulaires et ontologies rendent les données plus **interopérables**.
- Linked Data et FAIR prolongent cette vision vers la publication et la réutilisation des données.

Prochaine séance

Construire et exploiter un Knowledge Graph : requêtes, raisonnement, RAG et IA.