Web de données et RDA

Le Web de données ?

- Un Web constitué de données accessibles, structurées, dans un format non-propriétaire, identifiées et liées entre elles sémantiquement

  (Définition de Tim Berners-Lee dès 1999)

- Objectif : Mettre à disposition des données en utilisant des techniques standardisées qui garantissent l'interopérabilité (utilisabilité dans des contextes et avec des systèmes différents sans restriction de conditions d'accès ou de mise en œuvre)

Architecture du Web (1)

- **World Wide Web** : toile d'araignée de serveurs d'informations reliés les uns aux autres par des liens physiques (le réseau matériel) et des liens logiques (les liens hypertextes)
- **Architecture du Web** = les standards définissant l'infrastructure technologique
- Rôle du **W3C (World Wide Web Consortium)** : s'occupe de la standardisation de l'architecture du Web

Les objectifs du W3C

- **Accessibilité** pour les logiciels et machines
  - Interopérabilité et portabilité
  - Production de contenu Web facilitée
  - Réduction du volume des pages
  - Meilleure visibilité et indexation par les moteurs de recherche
  - Compatibilité
  - Pérennité des documents
  - Validation des pages par des services de validation pour garantir la cohérence et la qualité du code
- **Accessibilité** universelle aux contenus

Architecture du Web (2)

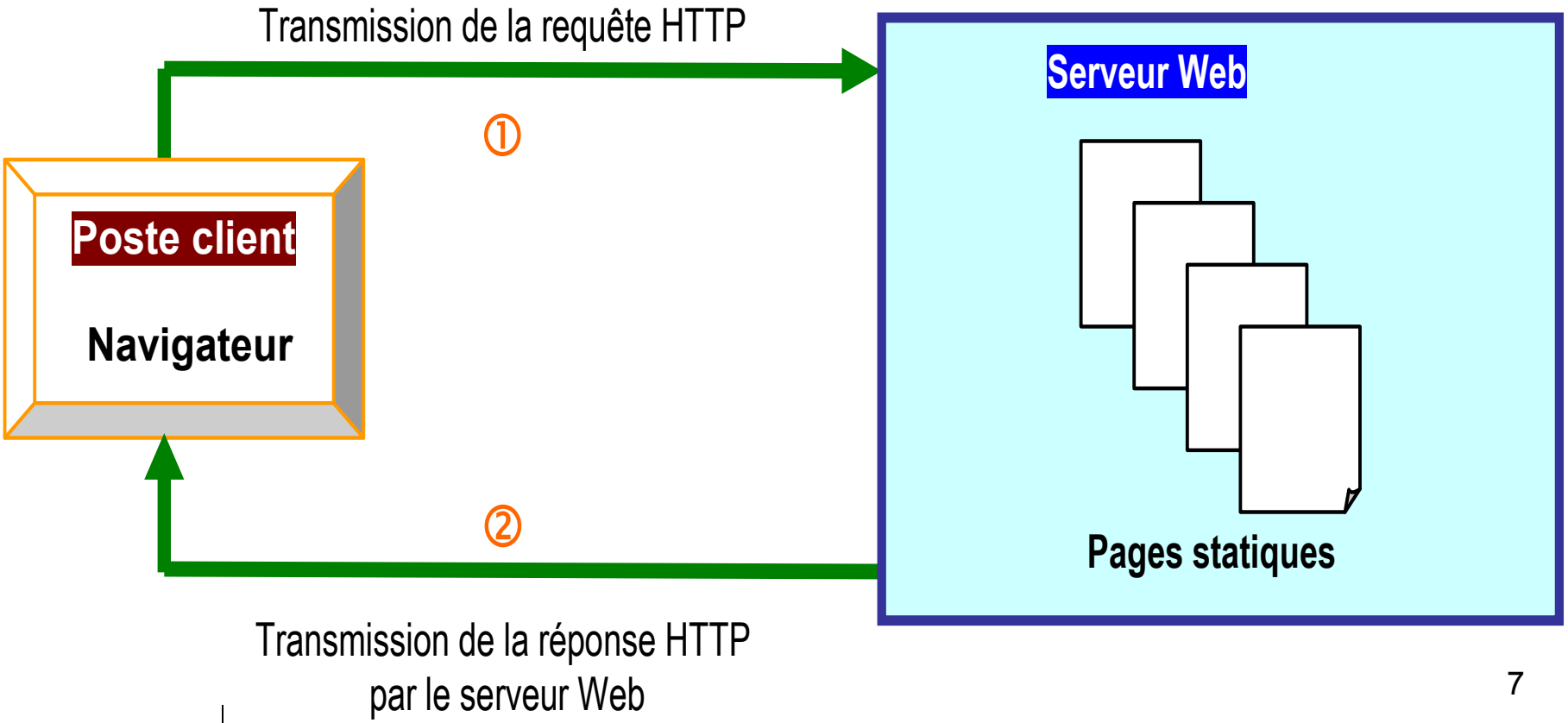
- **Repose sur 3 technologies** :

- **URI** (Uniform Resource Identifier)
  - Chaîne de caractères normalisés permettant d'identifier de manière permanente une ressource abstraite ou physique, accessible ou non sur Internet (personne, organisme, lieu, évènement, concept, …)
  - 3 déclinaisons :
    - **URC** (Uniform Resource Characteristic) : caractéristiques d'une ressource
    - **URN** (Uniform Resource Name) : nom d'une ressource
    - **URL** (Uniform Resource Locator) : spécification de l'adresse physique de localisation d'une ressource sur Internet et de la méthode permettant d'y accéder
- **HTTP** (Hypertext Transfer Protocol)
- **HTML** (Hypertext Markup Language) : standard défini par le W3C pour la diffusion de documents sur le Web pour pouvoir afficher de l'information à l'aide de balises dont le nombre est limité. Il est interprété par le navigateur

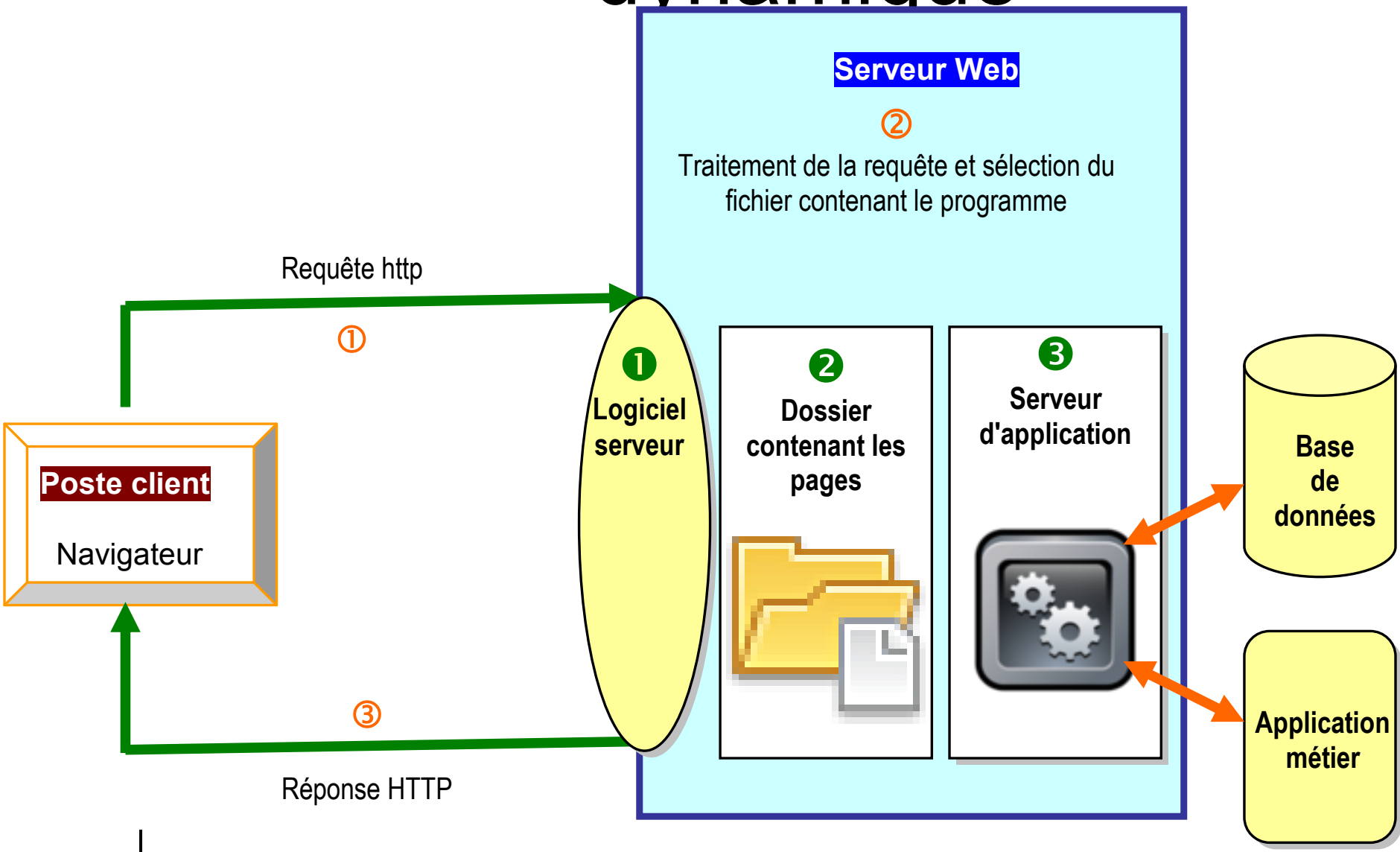
Une histoire : Web 1.0

- Web 1.0 = Web documentaire
  - Web statique
    - 1 page = 1 document
    - Traitement des ressources limité à leur mise en forme
    - Navigation entre les pages avec les liens hypertexte
  - ➢Limites du Web statique
    - Contenu non structuré
    - Pas de possibilité de requête
    - Impossibilité de renvoyer une page personnalisée selon le visiteur
    - Impossibilité d'exploiter les ressources d'une base de données

Web statique



Une évolution : le Web dynamique



Le Web 2.0

- **Pratiquement aujourd'hui tous les sites sont dynamiques**
  - Exploitation de volumes importants d'informations (bases de données, moteurs de recherche)
  - Personnalisation de l'accès à l'information
  - Naissance du **Web 2.0** = Web contributif
    - Les utilisateurs font partie du processus documentaire
    - Ajout de connaissances et de commentaires aux contenus

Le Web 3.0

- Extension du Web permettant de relier non pas des documents (pages HTML) mais les données elles-mêmes, et de les rendre exploitables par des machines
- Repose sur les mêmes technologies de base
  - **URI** : nommage des ressources
  - **HTTP** : transfert des données
- Mais :
  - Evolution du langage de balisage : il ne s'agit plus d'échanger des documents destinés à être immédiatement visualisés mais des données structurées
  - **XML** (eXtensible Markup Language) :
    - Format textuel d'échanges de documents et données structurées lisible par les machines
    - Chacun peut définir la structure et le balisage
    - Garantit l'interopérabilité, la portabilité et l'extensibilité des données et de leur structure
- Construction du Web de données liées grâce au langage **RDF** = 1 modèle et plusieurs syntaxes dont une en XML

RDF = Resource Description Framework

- Nouveau modèle généraliste et standardisé pour encoder, échanger et réutiliser des métadonnées structurées
- Proposé en 1999 par le W3C
- Langage dans lequel on décrit, représente et relie des ressources (= données) à échanger sur le Web
- Permet de décrire des ressources simplement : document, personne, objet, évènement
- **Objectif** : partager les mêmes métadonnées pour des ressources partagées par utilisation d'une syntaxe commune

RDF : un modèle conceptuel (1)

- Principe de base : toute chose peut être décrite avec des phrases minimales composées d'un verbe, d'un sujet et d'un complément = **déclaration RDF**

  Exemple :

  Honoré de Balzac a écrit "La Comédie humaine"

  Sujet : Honoré de Balzac ⇨**Ressource**

  Verbe : a écrit ⇨**Predicat**

  Complément : La Comédie humaine ⇨**Objet**

RDF : un modèle conceptuel (2)

- Modèle de données élémentaires constitué de 3 types d'objets :
  - **Ressource** : toute chose décrite par une expression RDF = entité d'information pouvant être identifiée par un identificateur (URI)
  - **Propriété** : caractéristique, attribut ou relation utilisé pour décrire une ressource
  - **Déclaration** : association d'une propriété à une ressource

RDF : la notion de triplet

- Une déclaration est composée de 3 éléments = **triplet**
- **Triplet {ressource – propriété – valeur}**
  - **sujet** = ressource
  - **predicat** = propriété : nature de la relation
  - **objet** = valeur : caractéristique ou ressource liée

  Exemple :

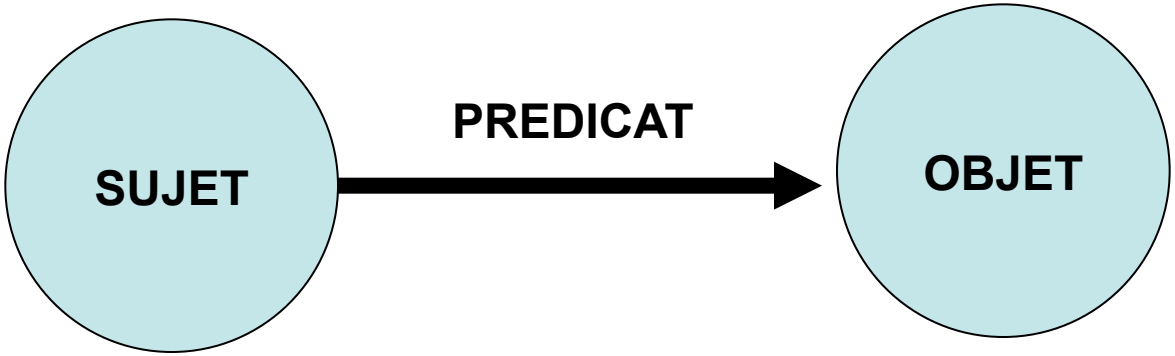
  Sujet (Ressource) : Honoré de Balzac

  Predicat (Propriété) : Creator

  Objet (Valeur) : La Comédie humaine

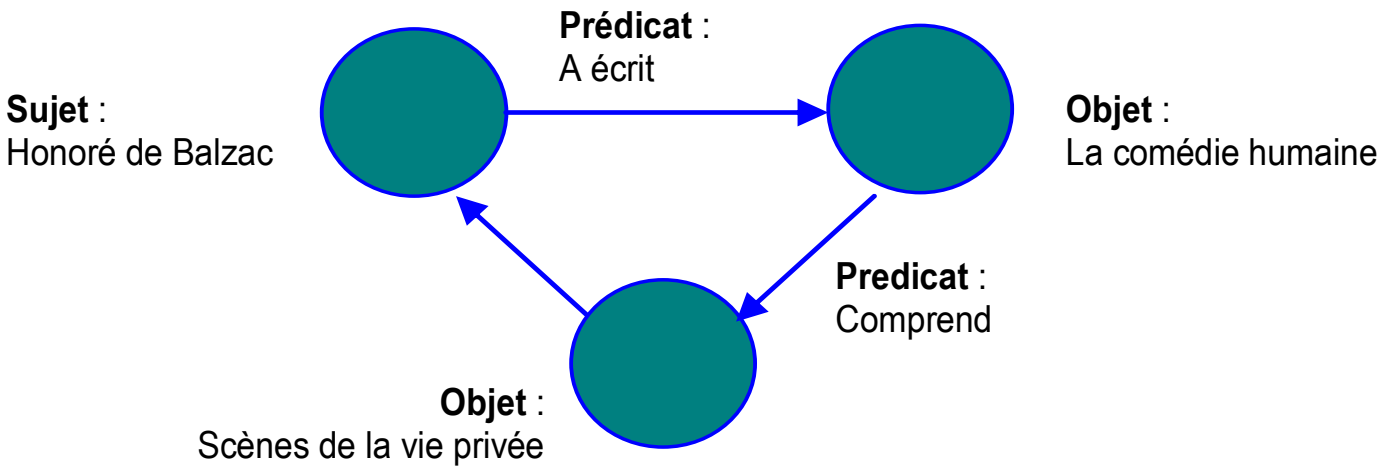
RDF : Graphe

- La déclaration est représentée visuellement par un graphe (système de nœuds reliés par des flèches) qui permet de parcourir l'information de lien en lien



RDF : modèle de graphe

- Chaque membre du triplet est une ressource qui peut être le sujet ou l'objet d'autres déclarations
- On construit ainsi un modèle de graphe



Formalisme RDF (1)

- Modèle permettant de représenter un nombre considérable de ressources désignées par un URI

- Eclatement de l'information
  - Des données et pas des « documents »
  - Plus de souplesse pour manipuler, sélectionner…

Formalisme RDF (2)

- Mais, des inconvénients :
  - Problème pour la création et la maintenance des URI
  - Complexité pour représenter certaines déclarations, les périodes, les provenances, l'absence d'information, globalement, tout ce qui nécessiterait un 4e élément dans le triplet
    - Exemples :
      - Jacques Chirac a été président de la République française de 1995 à 2007
      - Ce livre a été publié en 1981 par l'IGN à Paris et par la Chambre d'Agriculture de l'Indre à Blois
      - La date de publication de ce livre est inconnue mais doit se situer entre 1917 et 1923 selon la source X

RDF : un langage extensible

- Cadre conceptuel de description des ressources applicable à n'importe quel domaine d'application
- Peut être exprimé en utilisant la syntaxe **RDF/XML** (eXtensible Markup Language) : seule syntaxe qui fait l'objet actuellement d'une recommandation du W3C

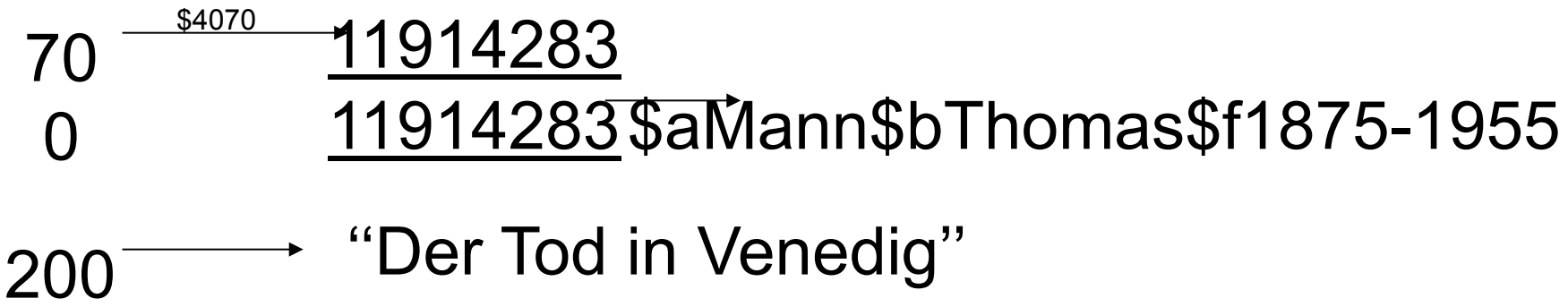
Syntaxe RDF

exemple : métadonnées de la DCMI

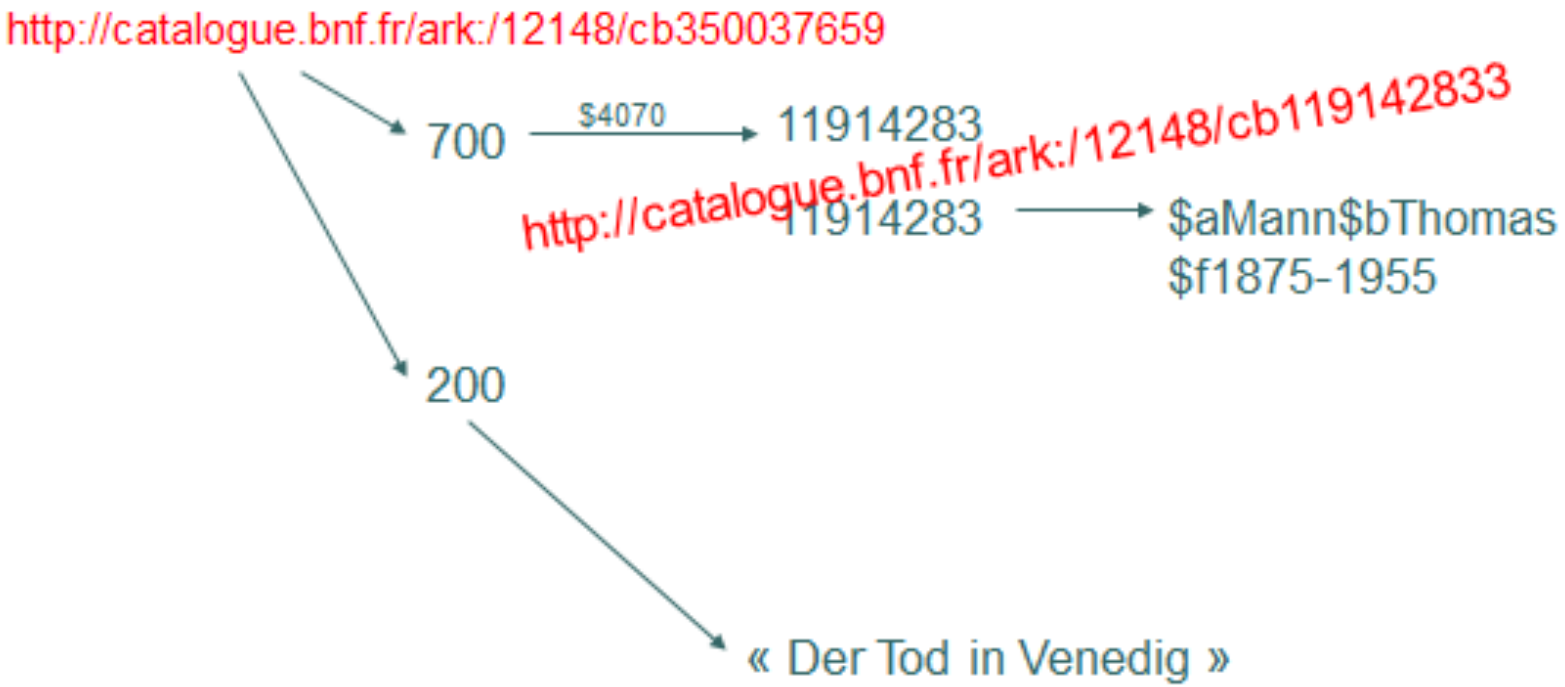
- <?xml version="1.0" encoding="iso-8859-1" ?>
- <!DOCTYPE rdf:RDF PUBLIC "-//DUBLIN CORE//DCMES DTD 2002/07/31//EN" "http://dublincore.org/documents/2002/07/31/dcmes-xml/dcmes-xml-dtd.dtd">
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
- xmlns:dc ="http://purl.org/dc/elements/1.1/">
- <rdf:Description rdf:about="http://www.w3.org">
- <dc:Title>L'avenir des méta-tags</dc:Title>
- <dc:Description> Avenir des métadonnées </dc:Description>
- <dc:Publisher>W3C</dc:Publisher>
- <dc:Date>2004-02-10</dc:Date>
- <dc:subject>Avenir des métadonnées</
- <dc:Type>World Wide Web Home Page</dc:Type>
- <dc:Format>text/html</dc:Format>
- <dc:Language>en</
- </rdf:Description>
- </rdf:RDF>

RDF (1) : Structurer l'information

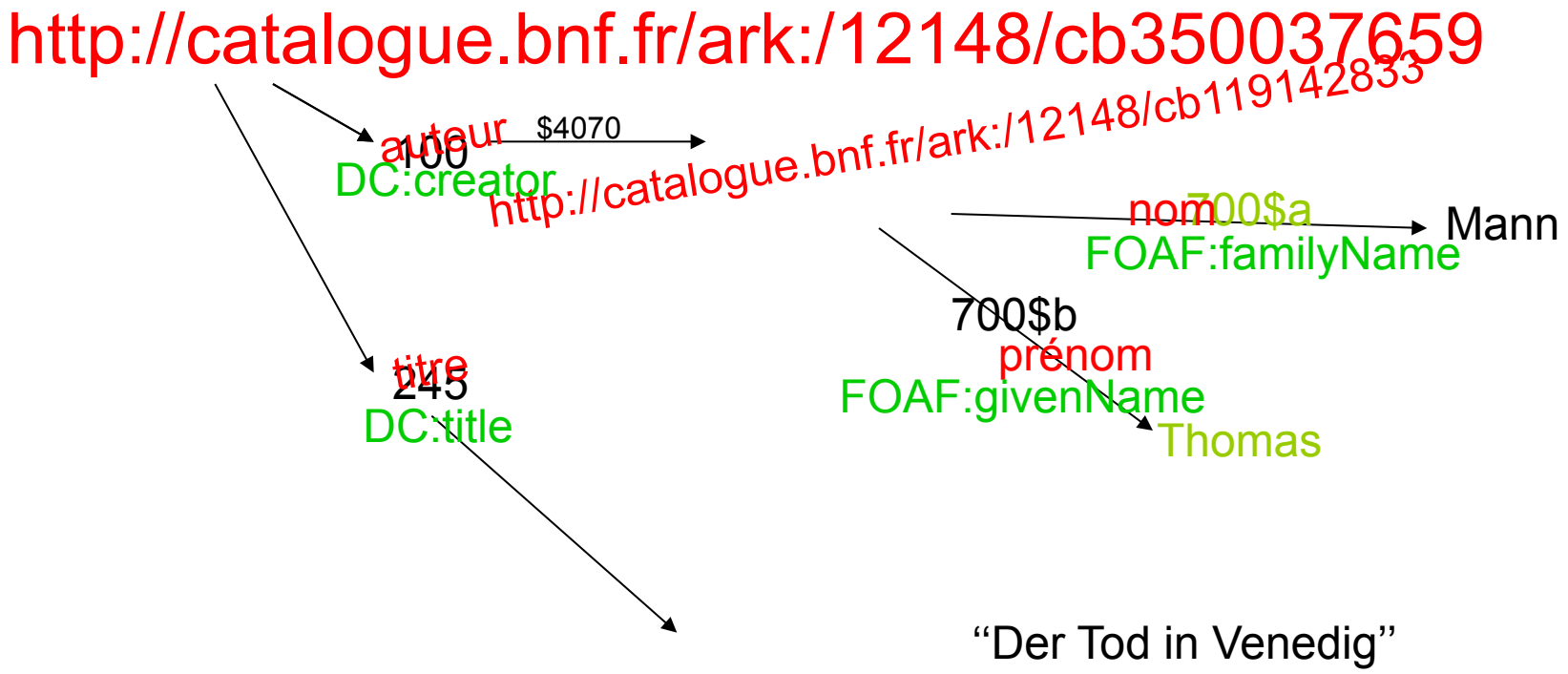
700 $311914283 $aMann$bThomas$f1875-1955$4070
200 1 $aDer Tod in Venedig$bTexte imprimé$fThomas Mann



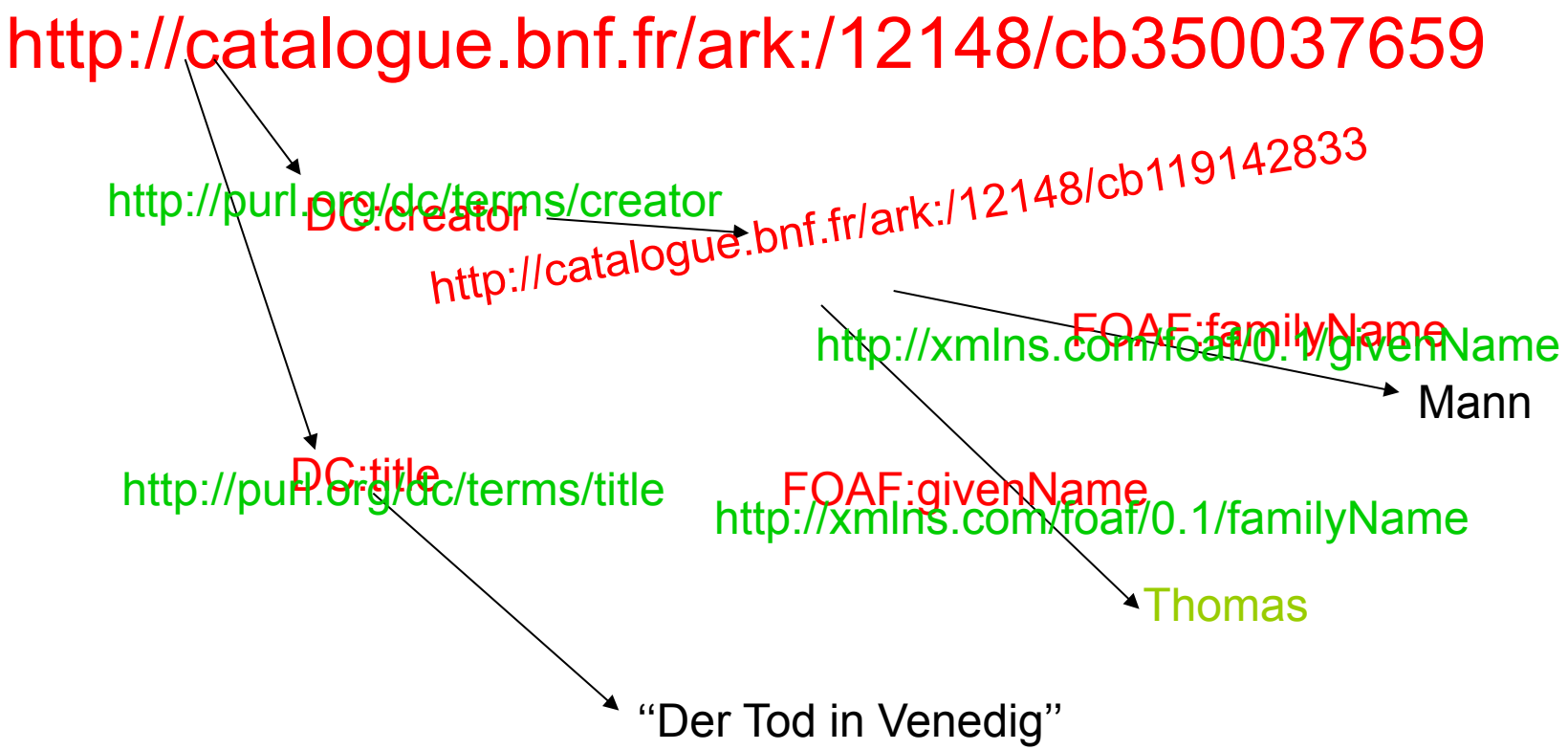
RDF (2) : Nommer les ressources avec des identifiants URI



RDF (3) : définir les relations entre les ressources en utilisant des vocabulaires normalisés



RDF (4) : Nommer les relations avec des URIs



RDF (5) : Exprimer le tout avec une syntaxe normalisée

@prefix dc: < http://purl.org/dc/terms/>

@prefix foaf: < http://xmlns.com/foaf/0.1/>

http://catalogue.bnf.fr/ark:/12148/cb350037659 dc:creator **http://catalogue.bnf.fr/ark:/12148/cb119142833**.

http://catalogue.bnf.fr/ark:/12148/cb119142833foaf:familyName "Mann";

foaf:givenName "Thomas".

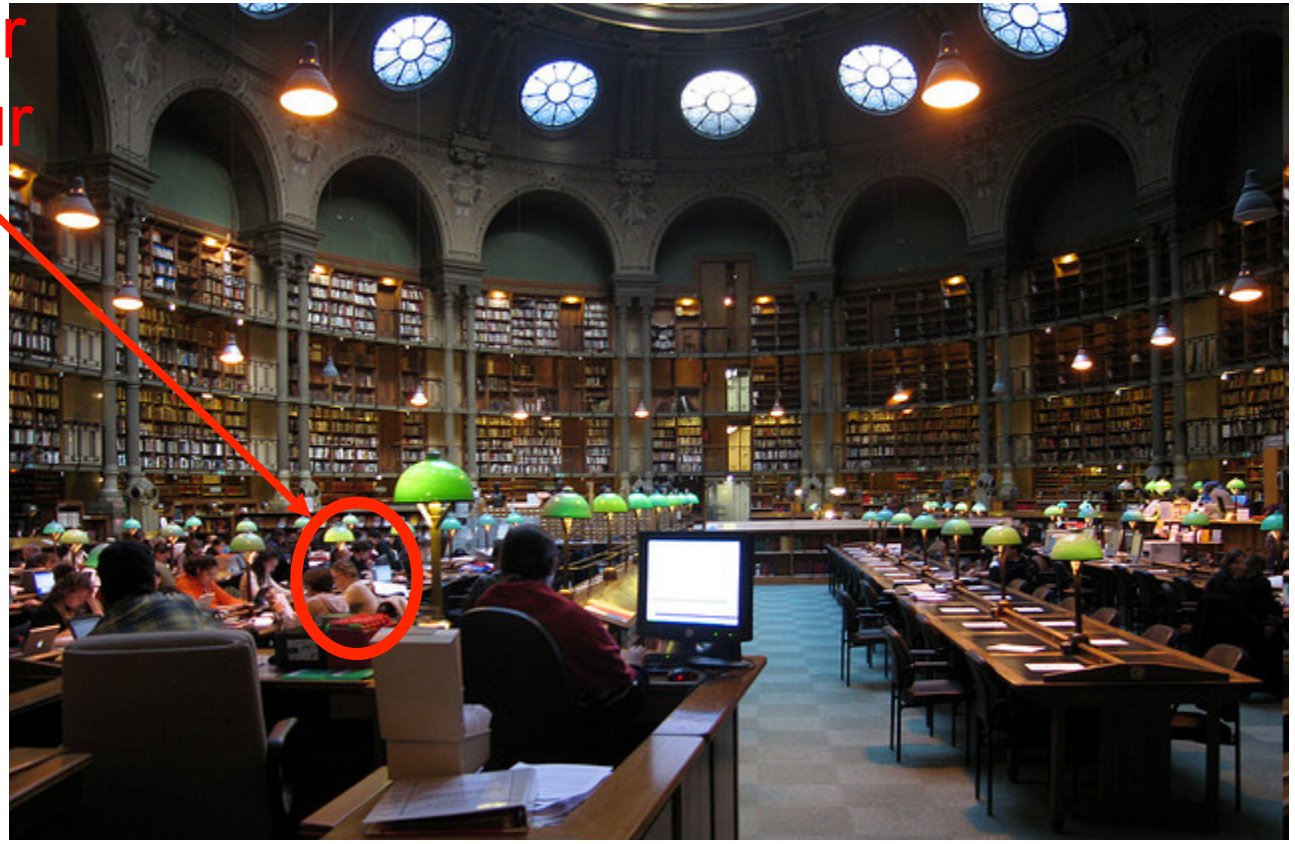
http://catalogue.bnf.fr/ark:/12148/cb350037659dc:title "Der Tod in Venedig".

Web de données et Web sémantique

- **Web de données** : possibilité de relier et d'échanger des données identifiées par des URI
- **Web sémantique** : possibilité d'échanger les schémas des données et la sémantique associée
  - Objectif : permettre aux machines de comprendre la sémantique, la signification de l'information sur le Web

C’ est très bien tout ça mais… Quel rapport avec la bibliothèque, le catalogue, le catalogage ?

Zoom sur
Le lecteur



Les bibliothèques dans le Web de données aujourd'hui

Occupons l'espace !!

Que peut nous apporter le Web de données ?

- Relier les catalogues des bibliothèques avec d'autres données existantes
- Ouverture à d'autres communautés (libraires, éditeurs, …)
- Navigation par les utilisateurs sans avoir à connaître les formats des bases de données et les langages de requête spécifiques
- Plus de visibilité par les moteurs de recherche
- Tirer parti des données structurées des catalogue et des référentiels
- Interopérabilité = Souplesse pour la réutilisation des données