Découverte des Avantages et des Limites du Web Sémantique pour les Archives Numériques: Regards Croisés avec Memobase

Roberta Padlina

CaféInteractif@Memoriav, 30 avril 2024

Préserver le patrimoine audiovisuel
www.memoriav.ch

Table des matières

Introduction : contexte et perspectives

Contexte : Archives numériques

Archives numériques en tant que fournisseurs de (re)sources pour les chercheurs en sciences sociales et humaines (Humanités numériques).

La mission principale des archives numériques est la **transmission** (préservation + accès) dans l'espace et le temps de :

1. **données** (matériel/objets/contenu numérisé ou né numérique) → matériel audiovisuel
2. **information** (métadonnées/catalogage/descriptions/...)

→ **connaissance** (la mise en relation de tout cela)

Archives numériques

Comment la transmission doit-elle s'effectuer ?

$\rightarrow$ perspective **scientifique** et **informatique** : selon les principes de la publication scientifique ainsi que de les principes de l'information et de la communication (ICT)

Perspective scientifique : principes de recherche

La recherche doit être de plus en plus **axé sur les données** (*data-driven research*) et leur publication scientifique doit être :

- FAIR (*findable-accessible-interoperable-reusable*)
- basée sur le paradigme des Linked Open Data (données ouvertes liées)
- basée sur les standards (normes et conventions)

Le Web comme lieu incontournable pour la publication des données

https ://www.go-fair.org/fair-principles/

Pour implémenter des données FAIR on a besoin de :

- institutions et infrastructures (serveurs, protocoles du Web)
- identifiants uniques, globales et persistantes
- information (métadonnées) riche et explicite pour les êtres humains ET les machines (*machine readable data*)
  $\rightarrow$ un langage formel pour la représentation de la connaissance (*knowledge representation*)
- liens et références qualifiées (non génériques)
- **modèles de données standard et formats internationaux**
- fichiers d'autorité et **vocabulaires** contrôlées
- **licences** ouvertes et transparentes (e.g. CC BY-SA)

Le paradigme des *Linked Open Data* (LOD)

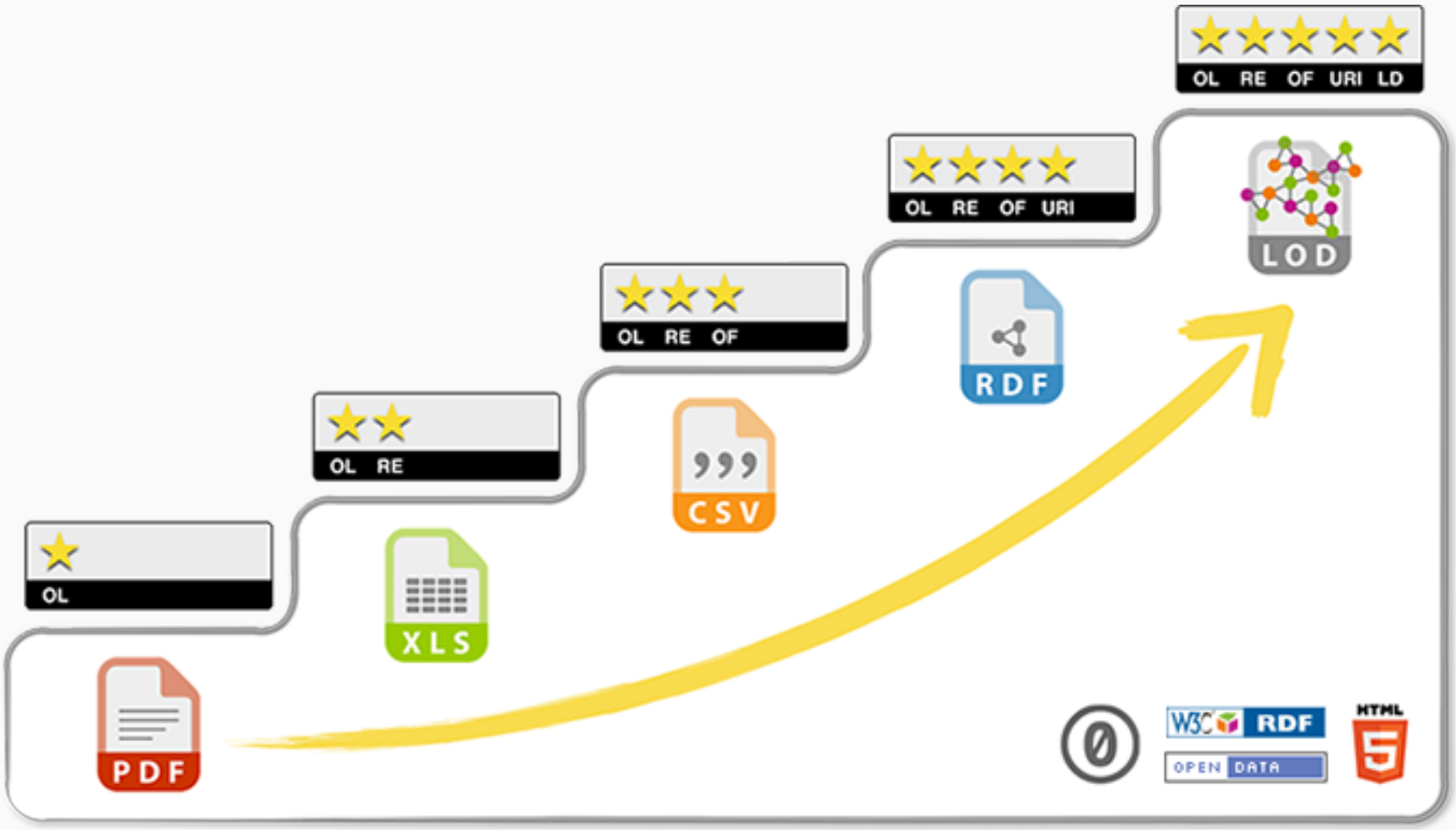
- **Évolution du Web** : Web of Documents → Web of Data
- **Linked Open Data** utilise les technologies du Web Sémantique pour
  1. pour publier des **données structurées** sur le Web, en utilisant le modèle de données RDF
  2. utiliser des liens RDF pour **interconnecter des données** provenant de différentes sources.

→ création d'un Web de données que les **machines** peuvent lire et comprendre → les produits du Web Sémantique sont destinés à être utilisés en premier lieu par des machines, et non pas par des humains !

LOD

« D'un point de vue technique, les données liées sont des données publiées sur le Web de manière à ce qu'elles soient **lisibles par une machine**, que leur **signification soit explicitement définie**, qu'elles soient **liées à d'autres ensembles de données externes** et qu'elles puissent à leur tour être liées à des ensembles de données externes» (Yu L., *Developer's Guide to the Semantic Web, Springer*, Berlin Heidelberg 2011, p. 409)

Différent niveaux de LOD (Five Stars)



Différent niveaux de LOD (Five Stars)

`https://5stardata.info/fr/`

online publiez vos données sur le Web (peu importe leur format) avec une licence ouverte → **PDF**

readable publiez les en tant que données structurées lisibles par une machine → **Excel**

open publiez les dans un format ouvert et non-propriétaire → **CSV** (comma-separated values)

identified utilisez des URI pour désigner des choses dans vos données, afin que les gens puissent faire des références à celles-ci → **RDF**

linked liez vos données à d'autres données pour y ajouter du contexte → **LOD**

Perspective informatique : Comment communiquer par et avec la machine ?

Le langage de l'ordinateur est le code binaire composé de 0 et 1 (**langage ou code machine**).

Or, le code doit être partagé pour pouvoir communiquer, mais les ordinateurs ne peuvent pas traiter directement les codes humains, et vice versa.

« L'ordinateur et l'homme n'utilisent pas un **code de communication** commun. Le problème de la mémorisation informatique d'un texte est toujours un **problème d'encodage**, puisqu'il s'agit de **traduire** ce texte quel qu'il soit pour qu'il soit lisible par une machine, de transposer l'information textuelle, comme on dit, en *Machine Readable Form* (MRF) » (Ferrarini, *La trascrizione dei testimoni manoscritti : metodi di filologia computazionale*, p. 104)

Code et langages de programmation

Encoder de l'information par et pour une machine en permet un **traitement computationnel et informatique**.

Les données encodées sont ensuite traités avec un **langage de programmation**, un ensemble d'instructions que la CPU assemble pour accomplir une certaine tâche.

Il existe de nombreux langages de programmation de différents types et différents niveaux

Comment parler à la machine ?

Plus que le choix d'un langage particulier, c'est la manière dont nous communiquons avec la machine qui importe le plus.

Nous devons nous adresser à la machine de la manière **la plus explicite et la moins ambiguë possible**, car la machine ne tolère pas les erreurs et est rapidement bloquée.

Encoder (exprimer, représenter) les données

- avec **précision**, **rigueur**, **exhaustivité**, **univocité** et **cohérence**
- en utilisant des **représentations standard**, des symboles et des règles d'un système conventionnel convenu (code **partagé**)

Paradigme scientifique et numérique

Raoul Mordenti, *Informatica e critica dei testi*, 2001 :

« **L'informatique ne peut résoudre de manière satisfaisante que des problèmes informatiques, c'est-à-dire posés de manière informatisée** » (p. 24)

« [La transcription] configure le moment crucial de l'encodage, c'est-à-dire l'introduction dans la machine de l'information **dont dépendront tous les traitements et manipulations ultérieurs** » (p. 29)

« les **normes** appelées à présider, de manière analytique et conventionnelle, à la transcription joueraient le rôle d'**intercode** (ou **métacode**) d'équivalence entre les deux systèmes, garantissant le **caractère scientifique** de l'opération de re-codification » (p. 76).

...et ChatGPT alors ?

Succès des applications d'apprentissage automatique (***Machine Learning***) : ChatGPT et autres ***Large Language Models***.

→ **langues naturelles ordinaires**

Nombreuses initiatives et investissements

Les risques de ChatGPT

Nombreux et graves problèmes :

- ***black box*** inintelligibles pour nous (et aussi pour leurs développeurs)
- **erreurs** et **hallucinations**
- entraînés avec des données pleines de **préjugés erronés** (*biases*)

Les deux premiers aspects sont dus à la nature **statistique et probabiliste** de ces instruments, qui fonctionnent par essais et erreurs (*trial-and-error*) et par approximations successives.

D'un point de vue scientifique, on ne peut pas faire **confiance** à ces instruments !

Machine Learning vs. Machine Reasoning

apprentissage automatique (ML) vs **raisonnement automatique** (*Machine Reasoning*, MR)

$\rightarrow$ **ML** nous donne la possibilité de nous exprimer directement dans nos **langues naturelles**, mais il donne aucune garantie de résultats.

$\rightarrow$ **MR** exige qu'on utilise une **langage formel**, mais nous donne la certitude que les connaissances dérivées sont correctes (à condition que les données initiales soient correctes).

Intelligence Artificielle : ML + MR

C'est urgent de produire et fournir à la machine :

1. de **données scientifiques de haute qualité et fiables** (pour contrer les données biaisées et la désinformation)
2. des **règles d'inférence solides** (pour éviter les hallucinations de l'IA)

→ combiner les deux approches pour améliorer l'IA : ML + MR, *Large Language Models* + *knowledge graphs*

→ Web Sémantique !

Structures de données et algorithmes

Deux grandes catégories de nombres, de choses représentées dans la machine :
les nombres qui **signifient** des choses (*structures dans l'espace*) et les nombres qui **font** des choses (*séquences dans le temps*)

→ **données** et instructions (programmes ou algorithmes)

«The title "On Computable Numbers" (rather than "On Computable Functions") signaled a fundamental shift. Before Turing, things were done to numbers. After Turing, numbers began doing things»

(George Dyson, *Turing's Cathedral. The origins of the Digital Universe*, 2012, pp. 305-306)

extract-transform-load (ETL)

Remarque initiale : nombreuses ressources (humaines, temps) consacrées aux processus

ETL

extract-transform-load

1. **Extraire** les données de différents types de sources et systèmes (format d'origine usuellement choisi pour sa facilité d'utilisation)
2. **Transformer** les données extraites en des ressources utilisables et fiables, dans un format (syntaxe) et une structure (sémantique) prévus par le système cible
3. **Charger** ces données dans le système cible (format cible choisit selon l'utilisation envisagée)

Règle générale

En règle générale : plus l'information est **granulaire et précise** (structurée et non-ambigüe), plus c'est facile de programmer un processus ETL.

Associer un **identifiant** (un numéro) à quelque chose donne à la machine un accès direct à cette chose, qu'elle peut ensuite manipuler selon les règles d'un système

$\rightarrow$ **encodage granulaire** et **identifiants uniques**

Standards

« Les informations sont souvent encodées dans les fichiers par le biais de normes qui sont spécifiées et qui existent en dehors des fichiers eux-mêmes. [...] normes qui, tout comme le format, transmettent des informations et fournissent un **contexte** »

« il est essentiel de conserver des informations sur le système d'encodage utilisé afin que les futurs utilisateurs puissent **interpréter correctement** les objets dans le contexte de leur plateforme. »

(Owens, *The Theory and Craft of Digital Preservation* p. 50)

Exemples de formats et standards

- **Formats** ($\sim$ syntaxe) : Excel, XML, RDF, JSON
- **Standards** ($\sim$ sémantique) : ISAD(G), PREMIS, METS, Dublin Core, PBCore, EBUCore, Records in Contexts (cfr. https ://memoriav.ch/fr/recommandations/all/11-metadonnees-pour-la-description-le-catalogage-linventarisation-des-documents-audiovisuels/)

Les données utilisant des standards facilitent l'échange et l'interopérabilité et, donc aussi, la durabilité des données.

Chaque format et standard a ses possibilités et ses limites.

Excel : pro et contra

Pro :

- simplicité d'utilisation (pour des données simples et non-ambiguës)
- universalité (outil connu et utilisé par tout le monde)

Contra :

- ne convient pas pour des données plus complexes et imbriquées $\rightarrow$ problèmes sémantiques :
  - pas de standard $\rightarrow$ la bonne interprétation des donnés dépend de la compréhension des noms des colonnes
  - mélange des catégories de niveaux différentes
  - différentes types de remarques ensemble dans la même cellule
  - erreurs difficilement repérables

XML : pro

- plus de granularité
- beaucoup de flexibilité
- meilleure gestion des dates et d'autres types de données
- plusieurs objets pour la même relation
- possibilité d'ajouter des identifiants aux micro-unités
- référence à des schémas standard

XML : contra

- unicité des identifiants seulement locale (dans le document même)
- XML exige une seule vue hiérarchique de la structure de l'objet
- trop de flexibilité : différents encodages pour le même type d'entité et même type d'encodage pour des entités différentes $\rightarrow$ utilisation ambiguë des balises
- pas se sémantique formelle (pour les machines) mais seulement description textuelle/narrative (pour les humains)

Situation

Grande hétérogénéité des données due aux :

- différentes langues naturelles
- différentes terminologies
- données peu structurées
- données structurées différemment
- portée locale des schémas/modèles des donnés
- sémantique seulement pour les humains et non pour les machines

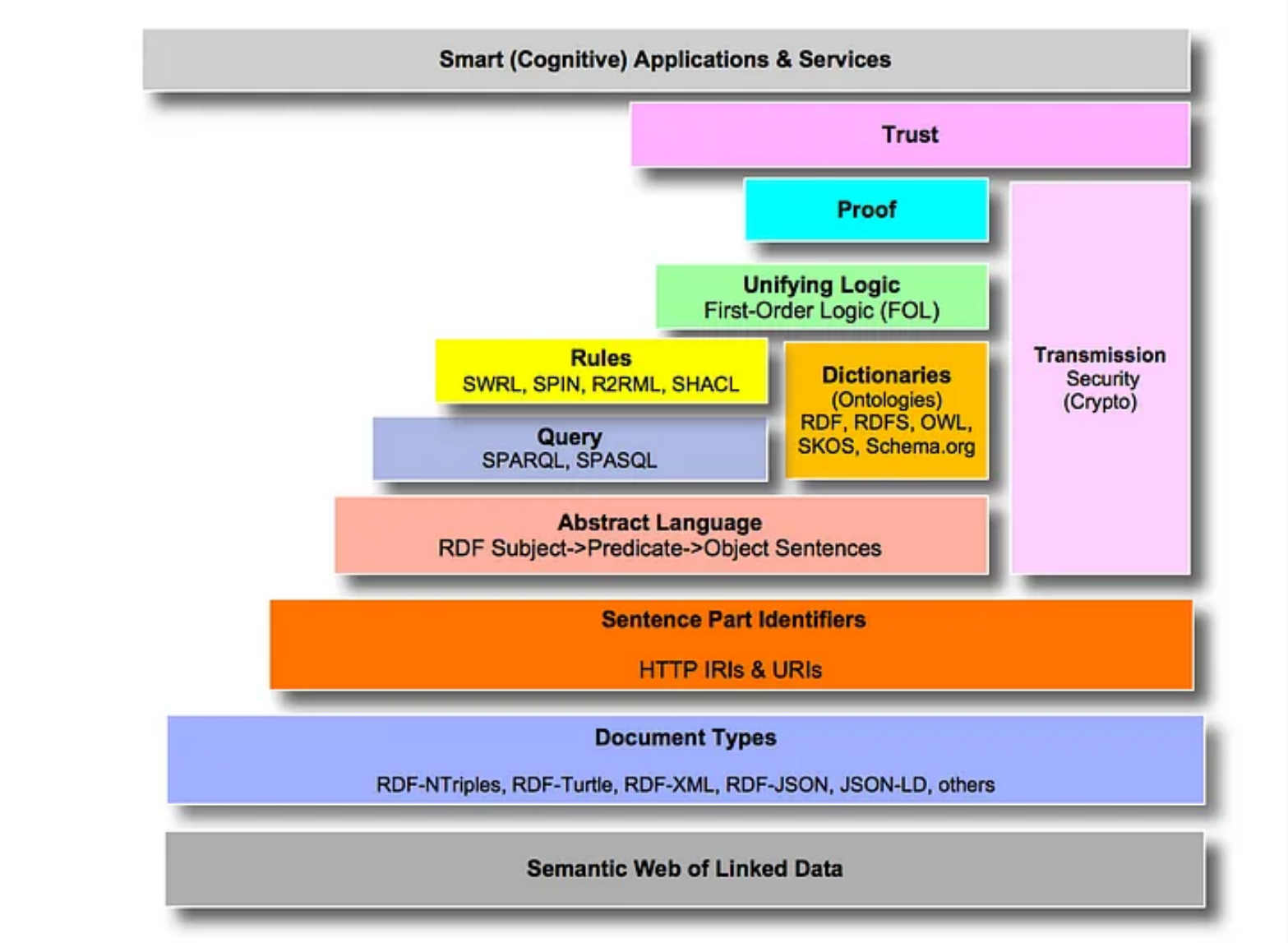
Le Web Sémantique et ses composantes

Le Web Sémantique

Le Web Sémantique est un ensemble de **langages formels open source et standard** recommandé par le **World Wide Web Consortium** :

- `https://www.w3.org/2001/sw/wiki/Main_Page`
- Tim Berners-Lee, James Hendler, Ora Lassila, « The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities », *Scientific America*, May 2001, p. 29-37.

The Semantic Web Stack



(https ://medium.com/openlink-software-blog/semantic-web-layer-cake-tweak-explained-6ba5c6ac3fab)

RDF : représentation de la connaissance et échange de données

Resource Description Framework : modèle de données et logique de base pour la modélisation de l'information et la description conceptuelle, ainsi que pour l'échange de données.

Le SW est formalisé dans RDF, lequel est utilisé pour représenter les descriptions de ressources et exprimer toute déclaration sur des faits ou des concepts sous forme de **triples** :

SUJET - PRÉDICAT - OBJET

RDF

L'ensemble des **termes RDF** est divisé en trois sous-ensembles **disjoints** :

- **URIs/IRIs** : les identifiants globaux pour les ressources web
- **Literals** : des chaînes de caractères simples associées à des types de données (comme par exemple une chaîne de caractères, un nombre ou une date) ; il est possible d'indiquer la langue de ces *literals*
- **Blank nodes** : représentent la quantification existentielle implicite ("il existe", "il y en a au moins un") et sont utilisées pour indiquer l'existence d'une ressource pour laquelle un URI/IRI n'est pas donné

RDF

Les deux formes les plus communes de triples sont :

Subject	*Predicate*	*Object*
IRI	**IRI**	**literal** .
dbr :Aristotle	rdfs :label	"Aristote"@fr .
IRI	**IRI**	**IRI** .
dbr :Aristotle	rdf :type	foaf :Person .

`https://dbpedia.org/page/Aristotle`

RDF

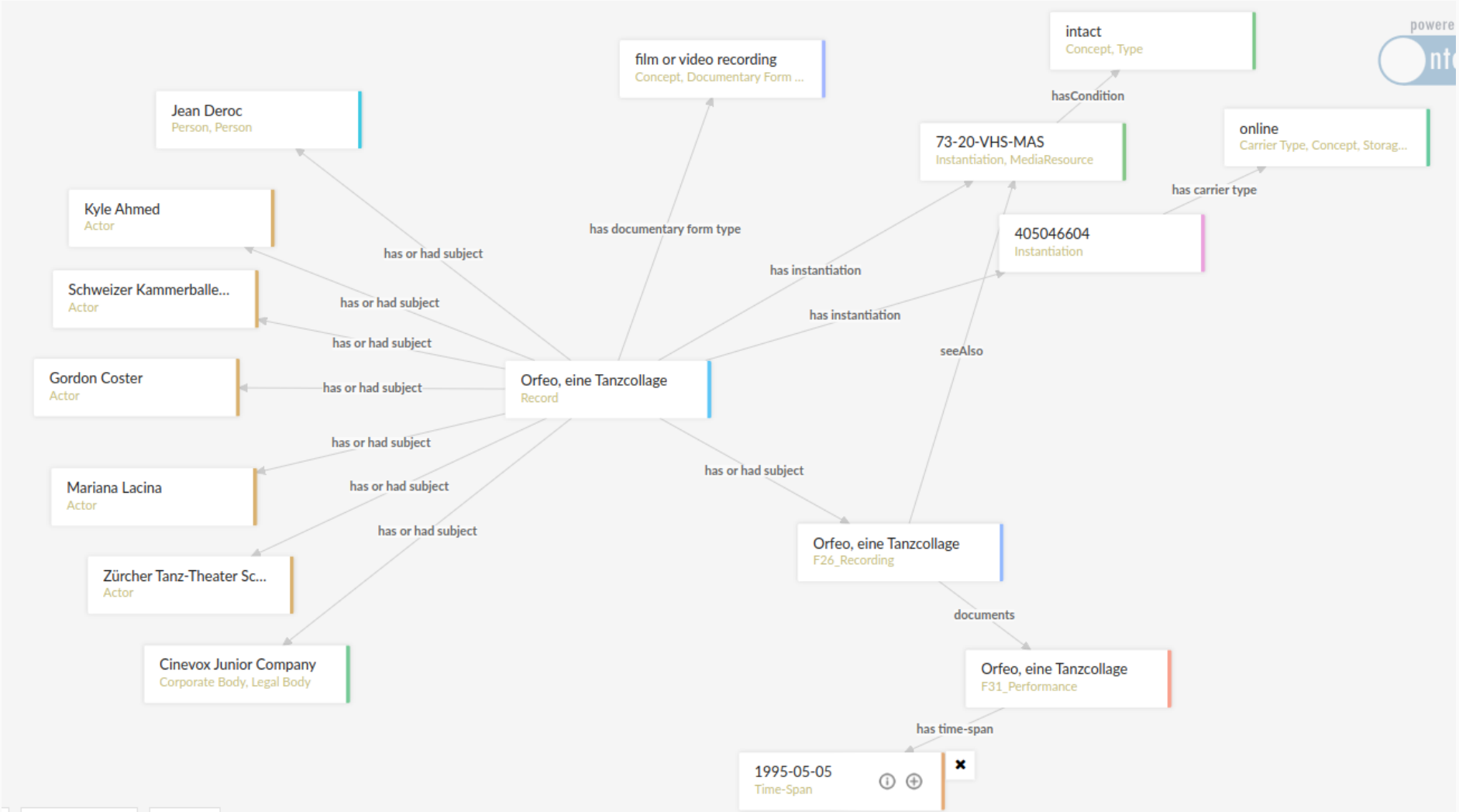
Prefixes for namespaces : `@prefix dbr:`

```
<http://dbpedia.org/resource/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns\#>
@prefix owl: <http://www.w3.org/2002/07/owl\#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
```

Subject	*Predicate*	*Object*
dbr:Aristotle	rdf:type	foaf:Person ;
	owl:sameAs	<www.wikidata.org/... .

Un ensemble de triplets RDF constitue un **graphe** RDF.

Web Sémantique : modèle de graphe (SAPA)



Sémantique formelle : RDFS et OWL

Sémantique → deux niveaux :

1. RDF Schema (**RDFS**) : taxonomies → concepts : ressource, classe, literal, type de données, sous-classe, sous-propriété,...
2. **OWL** 2 ontologies : modèles de données formels qui spécifient les concepts (**classes**) et les relations (**propriétés**) entre les concepts appartenant à un domaine donné → concepts : axiome, restrictions, symétrie, transitivité, cardinalité, équivalence, réflexivité ...

Remarques sur la modélisation des données

- on n'est pas limité à une seule ontologie, mais on peut (devrait) utiliser plusieurs ontologies
- il y a différent types de ontologies, notamment (mais il y a plus) :
  - ***upper/foundation ontologies*** définissant des relations et des objets communément partagés qui sont généralement applicables à un large éventail de domaines
  - ***domain ontologies*** qui représentent des concepts qui appartiennent à un domaine du monde, tel que la biologie ou la politique ou l'audiovisuel
- il y a plusieurs façons de construire/utiliser des ontologies (de manière directe, indirecte, hybride)

Autres composantes du Web Sémantique

- **langage de requête** SPARQL
- **raisonnement machine** (Machine Reasoning) basé sur des **règles** (*Notation3 rules*)
- *Unifying logic*, *Proof*, *Trust*

Memobase

Dans Memobase, on utilise :

- **plusieurs ontologies** (voir `https://api.memobase.ch/context.json`)
- **de manière directe**

Différent niveaux ontologiques dans Memobase

Niveau fondamental (W3C) :

- RDFS : <http ://www.w3.org/2000/01/rdf-schema>
- OWL : <http ://www.w3.org/2002/07/owl>

Ontologies génériques/basiques :

- SKOS : <http ://www.w3.org/2004/02/skos/core>
  (pour définir la proximité/distance entre concepts)
- DublinCore : <http ://purl.org/dc/elements/1.1/>
- Schema : <http ://schema.org/>

Ontologies pour le partage :

- Wikidata : <http ://www.wikidata.org/entity/>
- Europeana :
  <http ://www.europeana.eu/schemas/edm/>

EBUCore

`http://www.ebu.ch/metadata/ontologies/ebucore/`

The EBUCore is the Dublin Core for **media**, i.e. based on Dublin Core, but extended for media.

The EBUCore RDF schema (so called **EBUCore ontology**) is a semantic alternative to the **EBUCore XML schema**.

EBUCore dans Memobase

Classes et propriétés utilisées :

- ebucore :hasGenre
- ebucore :duration
- ebucore :locator
- ebucore :isDistributedOn
- ebucore :hasMimeType
- ebucore :hasFormat
- ebucore :height
- ebucore :width
- ebucore :orientation
- ebucore :mediaResourceDescription

Records In Contexts (Ontology)

ICA (Conseil International des Archives) RiC-O (Records in Contexts-Ontology) est une ontologie OWL pour décrire les ressources d'archives et leurs entités contextuelles

- **Site web** : `https://www.ica.org/en/records-in-contexts-ontology`
- **Spécification** : `https://www.ica.org/standards/RiC/RiC-O_v0-2.html`
- **GitHub** : `https://github.com/ICA-EGAD/RiC-O`
- `https://ica-egad.github.io/RiC-O/`

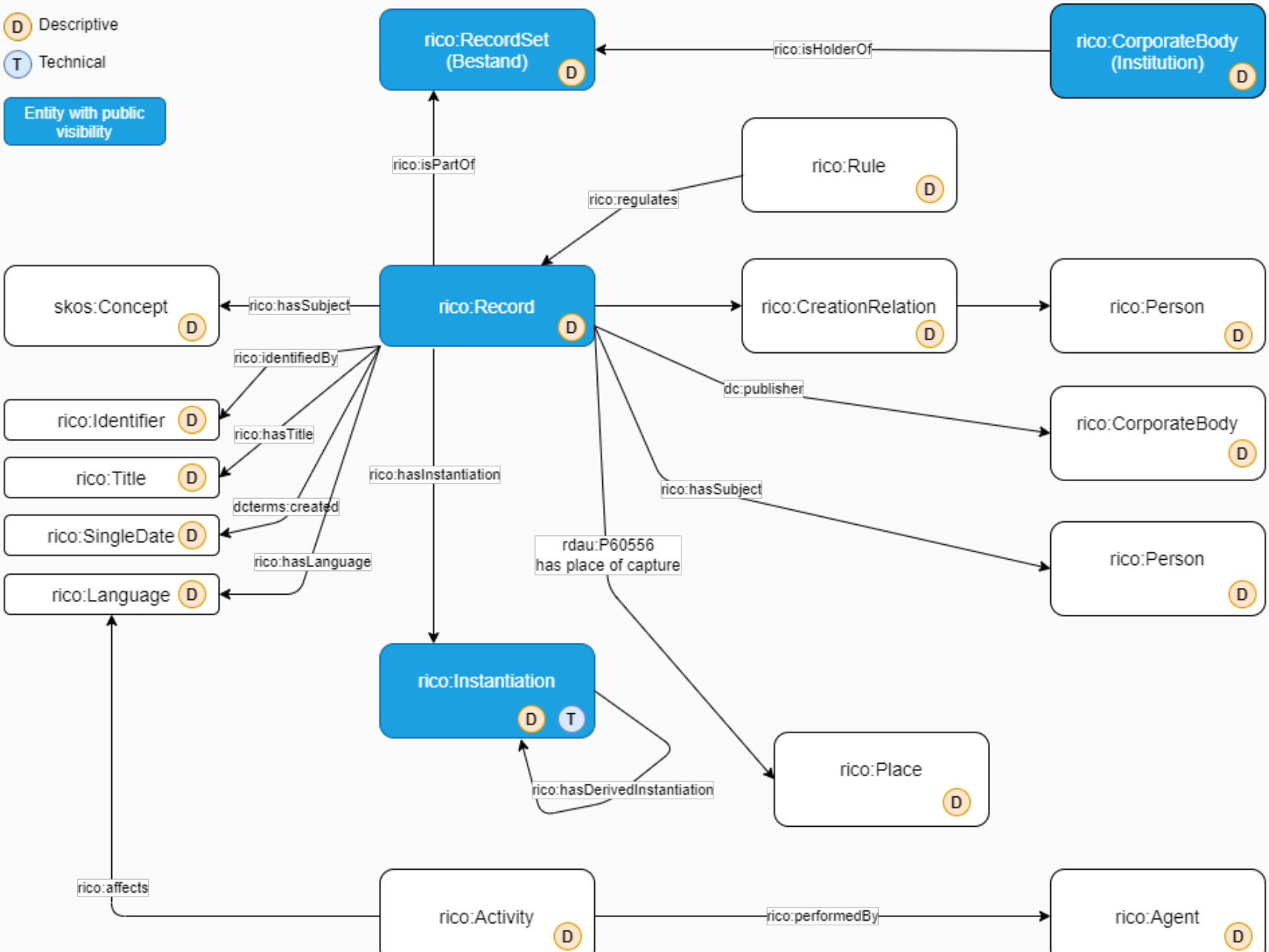
RiC-O in Memobase : classe → sous-classe

- rico :RecordResource → rico :RecordSet, rico :Record
- rico :Instantiation → physicalObject, digitalObject, thumbnail
- rico :Agent → rico :Person, rico :CorporateBody, rico :Mechanism
- rico :Date → rico :SingleDate, rico :DateRange, rico :DateSet
- rico :Place
- rico :Concept → rico :Language, rico :Title, rico :Identifier, rico :CarrierType
- rico :Relation → rico :CreationRelation, rico :recordResourceHoldingRelation
- rico :Rule
- rico :Event → rico :Activity

RiC-O in Memobase : propriétés

- rico :title
- rico :descriptiveNote
- rico :scopeAndContent
- rico :hasOrHadIdentifier
- rico :hasOrHadHolder
- rico :contidionsOfUse
- rico :publicationDate
- rico :hasCarrierType
- rico :technicalCharacteristics
- rico :performsOrPerformed
- ...

RiC-O


D Descriptive
T Technical
Entity with public visibility
rico:RecordSet (Bestand)
rico:isHolderOf
rico:CorporateBody (Institution)
rico:isPartOf
rico:Rule
rico:regulates
skos:Concept
rico:hasSubject
rico:Record
rico:CreationRelation
rico:Person
rico:identifiedBy
dc:publisher
rico:Identifier
rico:hasTitle
rico:CorporateBody
rico:Title
rico:hasInstantiation
rico:hasSubject
dcterms:created
rico:SingleDate
rdau:P60556
has place of capture
rico:Person
rico:hasLanguage
rico:Language
rico:Instantiation
rico:Place
rico:hasDerivedInstantiation
rico:affects
rico:Activity
rico:performedBy
rico:Agent

Limites et avantages du Web Sémantique

Problèmes et soucis en rapport au SW

Expressivité :

- **souci de perte d'expressivité** due à la nécessité d'utiliser des normes préétablies (ontologies standards) et suivre le consensus
- l'**explicitation** de la sémantique des données pose des défis et révèle toutes sortes d'incohérences et de lacunes dans les données sources qu'il faut corriger

Complexité :

- les **ontologies** sont très complexes (avec différents types et niveaux d'abstraction)
- pour récupérer les données il faut connaître **SPARQL**
- manque d'**outils simples** à utiliser

→ besoin de formation à ces technologies

Défis techniques : pas encore de **standardisation** de couches supérieures

Avantages du Web Sémantique

Les deux plus grandes avantages du SW sont que :

1. les **identifiants** sont uniques au niveau global (Web)
2. il y a des identifiants pour les sujets/objets, mais aussi pour les relations → chaque relation est précisément définie (*typed links*, **relations qualifiées** ou liens typés)

→ on est sûr de parler des même choses et de même types de relations

→ **pas d'ambiguïté** sur la signification des données

Les identifiants globaux permettent d'interconnecter de manière significative des données dispersées dans différentes sources → **réutilisation et augmentation de la connaissance** (plus de contextes).

Avantages du Web Sémantique

Expressivité :

- enrichissement conceptuel et amélioration de l'expressivité par une **sémantique explicite, auto-descriptive et formelle**
- possibilité de définir plusieurs hiérarchies (physique, conceptuel)
- indépendance par rapport aux langages naturels => échange d'informations à l'échelle globale
- connaissance du domaine consensuelle et réutilisable
- création d'un **espace sémantique global**
- sémantique interprétable par le machines

Avantages du Web Sémantique

Qualité des données :

- **désambiguïsation** de l'information et **transparence** des données
- **détection automatique** des erreurs, incohérences et lacunes
- la logique intégrée permet d'**évaluer** la rigueur du raisonnement

Avantages du Web Sémantique

Qualité des services :

- **accès direct** à toutes les données qui sont très granulaires
  - SPARQL-**endpoint de SAPA** : `https://www.performing-arts.ch/sparql`
  - **Memobase API** : `https://api.memobase.ch/`
- interopérabilité sémantique automatisée
- raisonnement automatique (*Machine Reasoning*)

Conclusion

Conclusion

Web Sémantique :

- représentation de la **sémantique** des données
- équilibre entre **compréhension formelle pour les machines** et **facilité d'expression pour les humains**
- une grande opportunité
- un retour sur investissement important

. . . mais encore beaucoup de travail à faire !

Conclusion

→ De Coulon, B. (2024). «Déploiement de la norme Records in Contexts pour la gestion des collections de la Fondation SAPA». *Revue électronique Suisse De Science De l'information* (RESSI), (24). https ://doi.org/10.55790/journals/ressi.2024.e1511

→ Plüss, Rebekka and Padlina, Roberta. «Wissensnetz der Zürcher Ehedaten des 16.–18. Jahrhunderts : Eine Anwendung von Semantic-Web-Technologien im Archiv», *ABI Technik*, vol. 42, no. 4, 2022, pp. 230-241. https ://doi.org/10.1515/abitech-2022-0043

Documentation, Code & Contact

Wiki	`https://ub-basel.atlassian.net/wiki/spaces/MD/overview?homepageId=47350146`
GitLab	`https://gitlab.switch.ch/memoriav`
API	`https://api.memobase.ch/`
Feedback	`memobase@memoriav.ch`
Contact	`roberta.padlina@memoriav.ch`