

UNIVERSITÉ CLAUDE BERNARD LYON 1

Rapport

Application web collaborative de gestion d'œuvres culturelles

Pour le 21/05/2018

Auteurs: Sarah Touatioui Audrey Seguin

Responsable: Fabien Duchateau

1 Introduction	2
1.1 Contexte	2
1.2 Problèmes	2
1.3 Objectifs	2
2 Pré-requis	3
2.1 Le web sémantique en bibliothèque	3
2.2 Ressources externes	3
3 Mise en oeuvre	4
3.1 Environnement de développement	4
3.2 Conception du modèle	4
3.3 Intégration de données	5
3.4 Entity linking	6
4 Conclusion	7
5 Annexes	8

1 Introduction

Ce projet a été effectué dans le cadre de l'UE de Projet d'Orientation en Master de l'Université Claude Bernard. Notre projet nous a été confié par l'un de nos enseignants Monsieur Fabien Duchateau qui nous a encadrées durant cette période.

1.1 Contexte

De nombreuses associations gèrent des collections d'œuvres culturelles (essentiellement bibliographiques) sur des thématiques très spécifiques. Cependant, leur budget ne permet généralement pas de disposer d'un outil de gestion de ces œuvres. De plus, ces outils de gestion sont généralement destinés à un public expert en catalogage (libraires), et comportent des fonctionnalités et un modèle de données complexes. Ces collections sont donc isolées et ont une visibilité limitée.

Ce projet sera appliqué aux œuvres culturelles en Espéranto. Les espérantistes sont organisés en groupes locaux (en général, un groupe par ville), et chaque groupe local dispose généralement d'une collection de livres à disposition de ses membres. L'idée est de créer une bibliothèque virtuelle pour l'Espéranto, en permettant aux groupes de cataloguer simplement leurs livres, de rechercher des livres de façon intuitive et de demander à emprunter un livre.

1.2 Problèmes

Différents problèmes se posent pour les espérantistes. Avoir une vision d'ensemble sur les œuvres en lien avec l'espéranto est difficile car aucun outil n'existe pour effectuer un tel catalogage. Quelques institutions, comme UEA (association universelle d'espéranto) ou le musée de l'espéranto à Vienne, proposent une collection d'œuvres. Mais ces quelques collections ne fournissent pas une vision globale, en particulier parce que les collections sont très hétérogènes et fortement incomplètes (par rapport aux collections détenues par les groupes). Cette hétérogénéité se retrouve à différents niveaux : les œuvres ne sont pas toutes disponibles sous le même format, elles n'ont pas le même modèle de données, ce qui soulève des conflits d'intégration de données à résoudre afin d'effectuer un catalogage complet.

1.3 Objectifs

Ce projet a pour objectif le développement d'un prototype d'une application collaborative, qui permet à différents groupes espérantistes de gérer des collections de livres. Pour cela, il nous faut faire l'intégration d'œuvres existantes à partir de sources externes. En effet, il est préférable d'avoir une collection initiale, ce qui évite aux groupes de saisir l'intégralité de leur collection. Les sources étant diverses et n'existant pas toutes sous la même formes, il nous faudra également uniformiser toutes les métadonnées. Grâce à cette application, les utilisateurs devront pouvoir gérer les collections d'œuvres et les partager.

2 Pré-requis

Dans le Web sémantique, une ontologie est un langage plus expressif qu'un schéma RDF. Une ontologie cherche à s'appuyer sur des modélisations de ressources du Web à partir de représentations conceptuelles et a pour objectif de permettre à des programmes de faire des inférences dessus. Elle est vue comme un ensemble de connaissances, on y retrouve également le vocabulaire et les relations sémantiques, avec quelques règles simples d'inférence et de logiques relatives à des sujets particuliers. Une ontologie doit être conçue comme un modèle réutilisable afin de partager des connaissances et de constituer une base commune pour l'échange et l'interprétation d'informations.

2.1 Le web sémantique en bibliothèque

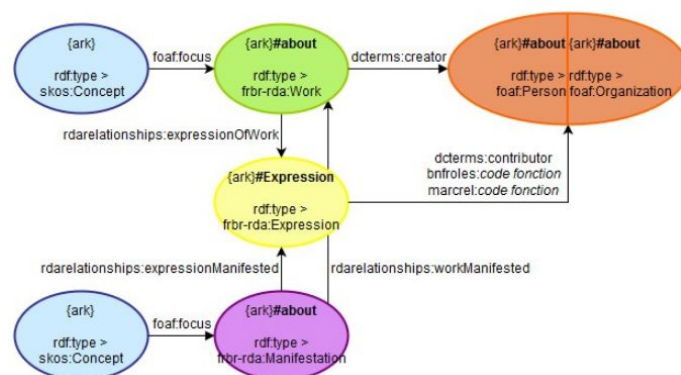
Plusieurs ontologies sont disponibles dans le domaine bibliographique afin de représenter les œuvres culturelles. Les principes FRBR définissent des spécifications pour l'œuvre, les agents qui y ont contribué, les différents versions, les caractéristiques physiques des exemplaires, etc. FRBRcore, FRBRer et FRBROo sont des implémentations de FRBR en RDF. RDA est un vocabulaire RDF qui s'inspire des principes FRBR. Enfin BIBFRAME et bibliotek-o sont des alternatives à FRBR, qui simplifient notamment les principes de base.

Le choix de l'une ou l'autre des librairies dépend de ses propres préférences ainsi que de l'utilisation que l'on veut en faire. Sur recommandation de notre encadrant, nous avons donc choisi d'utiliser la librairie bibliotek-o (dérivée de bibframe). Ce choix dérive du fait que cette librairie est plus simple de compréhension que d'autres telles que FRBROo qui sont elles, assez complexes pour les personnes n'étant pas des bibliothécaires.

2.2 Ressources externes

La bibliothèque nationale de France (BnF) a pour mission de collecter, conserver, enrichir et communiquer le patrimoine documentaire national.

Seule une partie du catalogue de la BnF est disponible sous forme de données ouvertes et liées (LOD) : *data.bnf.fr*. Ces données sont représentées sous forme de graphe RDF (suivant le modèle FRBRer), et elles peuvent être interrogées via un endpoint sparql. (<http://data.bnf.fr/opendata>)



L'Association Universelle d'Espéranto (UEA) est aujourd'hui la plus importante association internationale consacrée à l'espéranto.

UEA Katalog est un catalogue mis à disposition du grand public et qui recense les oeuvres en lien avec l'espéranto. Cependant, ce catalogue n'est pas disponible sous forme de dump ou interrogeable par des requêtes. L'association fournit une archive contenant une page HTML pour chacune des 8000 oeuvres. (<http://katalogo.uea.org/?inf=1670>)

Titolo	Esperanta antologio	
	Poemoj 1887-1981	
Kategorio	Antologioj	
Prezo	9.00 €, triona rabato ekde 3 ekz.	
Eldonloko, jaro	Rotterdam, 1984 (2a eld)	
Eldoninto	UEA	
Klarigoj	Pli ol 850 originale verkitaj poemoj de 163 poetoj donas superrigardon pri la tuta poezia kulturo de Esperanto; kun postparolo de W. Auld kaj 36 paĝoj da biografiaj notoj.	
Kontribuantoj	Red. William Auld	
ISBN/ISSN	9290170301	
Formato	12+888 paĝoj, 22 cm	
Karakterizoj	Ilus. Bind. Kromkovrilo. Jubilea Kolekto n-ro 1	
Recenzoj	Esperanta Antologio de Gaston Waringhien	Aldonu

3 Mise en oeuvre

3.1 Environnement de développement

Dans cette section nous décrivons les programmes principaux que nous avons utilisé pour la conception et la mise en oeuvre du modèle de données.

Protégé, un système open source gratuit, développé en java par l'université Stanford. Utilisé dans le domaine du web sémantique, il permet la création d'ontologies dans divers formats : RDFS, OWL. Pour privilégier la réutilisation d'ontologies déjà existantes, Protégé permet aussi de charger d'autres ontologies depuis un fichier interne ou sur internet.

Jena, un framework Java libre et open source pour la construction d'applications Web sémantique et de données liées. Dans jena, nous trouvons principalement:

Un **modèle RDF**, un ensemble de tuples pour qui des méthodes lui sont fournies afin de créer des ressources des propriétés et des littéraux. Le modèle peut être interrogé pour une simple sélection, un ajout ou une suppression.

Un **dataset RDF**, un ensemble de données RDF. Il contient un graphique par défaut, étant l'union de tous les graphiques nommés. Concrètement, un dataset est un répertoire de fichiers, utilisé pour stocker un jeu de données RDF. Dans une requête SPARQL, un modèle de requête est comparé au graphique par défaut.

Un **reasoner**, permet la déduction d'un ensemble de données, à partir d'un modèle contenant toutes les données RDF et d'un schéma contenant le vocabulaire du modèle.

Un **infModel**, un modèle d'inférence. Il est créé en attachant une instance d'un raisonnement à un ensemble de données RDF, un modèle. Ainsi ce modèle contiendra toutes les données de base ainsi que celles déduites grâce aux règles d'inférences.

3.2 Conception du modèle

Dans cette section nous décrivons les étapes que nous avons suivies pour concevoir l'ontologie.

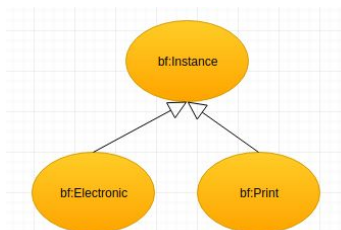
Pour développer l'ontologie, nous utilisons une approche itérative. Nous commençons par aborder l'ontologie de manière générale, nous la complétons de manière plus détaillée. Ensuite, nous revenons dessus pour discuter des points forts et des points améliorables et nous effectuons des modifications si nécessaire. Tout au long de ce processus, nous passons par les étapes suivantes:

Etude des besoins: Dans cette étape, nous soulignons ce que l'ontologie doit pouvoir fournir comme information. Elle doit principalement permettre la gestion des livres et des utilisateurs de l'application.

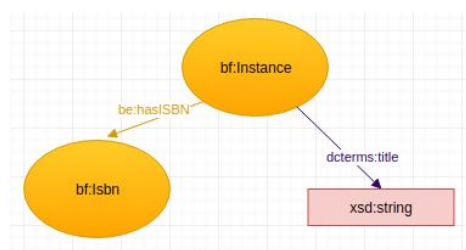
Réutilisation des ontologies et vocabulaires existants: après spécification des besoins, nous cherchons parmi les ontologies déjà existantes celles qui les satisfont le plus, et nous privilégions donc au maximum la réutilisation de celles-là. De même pour les vocabulaires, il est recommandé de réutiliser les concepts existants quand c'est faisable. Pour la gestion des livres nous utilisons le vocabulaire défini par bibliotek-o: <http://bibliotek-o.org/ontology>. Quant à la gestion des administrateurs, membres et responsables de groupe nous utilisons le vocabulaire défini par vCard: <http://www.w3.org>.

Enumérer les termes importants dans l'ontologie: pour chaque terme, nous définissons sa signification exacte ainsi que le concept qu'il représente. Nous établissons aussi une liste de toutes ses propriétés. Par exemple, dans notre cas une instance d'un livre représente une édition de celui-là. Une *instance* est caractérisée par son titre, auteur, éditeur, date et lieu de publication et un isbn, etc.

Définition des classes: une fois toutes les définitions des termes établies, nous passons à la définition des classes et de la hiérarchie des classes. Un livre peut par exemple avoir deux exemplaires de version différentes, une version électronique et une version papier. Ainsi ces deux classes "*electronic* et *print*" dérivent de la classe *instance*.



Définition des propriétés de classes: après définition des classes, nous définissons leurs propriétés. Il existe deux sortes de propriétés, des *Object properties* et des *data properties*. Une propriété d'objet relie une ressource à un objet, alors qu'une propriété de données relie une ressource à un littéral. Par exemple, une instance peut être reliée à un objet définissant son isbn par la propriété *be:hasISBN* et reliée à un littéral définissant son titre par la propriété *dcterms:title*.

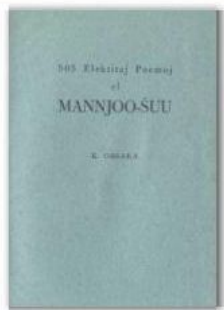


3.3 Intégration de données

Dans l'idéal, nous souhaitons intégrer les catalogues de UEA et de la BnF.

Pour la BnF, nous avons constaté que dans le catalogue sémantique (*data.bnf.fr*), les oeuvres en espéranto ne sont pas encore présentes. Pour les intégrer, il faut utiliser les données stockées au format MARC, un format de catalogage des années 1960 qui est difficilement compréhensible par des non-bibliothécaires. L'intégration des données de la BnF a donc été abandonnée.

Concernant UEA, le catalogue est fourni sous forme de fichiers HTML (un par oeuvre, voir figure de droite). Les fichiers HTML semblent avoir été générés avec une même application, et il est donc relativement facile de les parser automatiquement. Chaque fichier possède un certain nombre d'informations sur l'oeuvre qu'il décrit, par exemple le titre, le nom de l'auteur, le nombre de pages, une catégorie, etc. Il y a beaucoup moins d'informations sur UEA que ce que l'on pourrait trouver dans un catalogue de bibliothèque, mais ces données sont suffisantes car principalement celles qui intéressent les espérantistes.

Titolo	505 elektitaj poemoj el Mannjoo-ŝuu	
Kategorio	Antologioj, Poezio tradukita	
Prezo	5.10 €, sesona rabato ekde 3 ekz.	
Eldonloko, jaro	Tokio, 1958	
Eldoninto	JEI	
Klarigoj	El la unua japana poezia antologio, 8a jc. p.K.	
Tradukisto, lingvo	Kenji Ossaka / El la japana	
Formato	212 paĝoj, 13 cm	
Recenzoj	Recenzoj mankas.	Aldonu

Pour intégrer les données UEA, nous avons récupéré les oeuvres par catégories. Pour chaque catégorie, nous trouvons une liste de liens vers les livres qu'elle contient. Il suffit alors de parser la liste, récupérer tous les liens référençant ces livres et ensuite lancer une requête GET pour chacun des liens. Pour chaque requête, le serveur répond par un fichier html qu'il faut parser à nouveau afin de récupérer les données concernant le livre.

Antologioj

[Laŭkategoria indekso de la Katalogo](#)

(Momente nehavebla) **33 Rakontoj**. La Esperanta novelarto. Div. aŭtoroj. J. Régulo. La Laguna. 1964. 328 paĝoj. Prezo: 18.00 €

Klarigo: Klasika antologio de la Esperanta prozo.

>> Pliaj informoj: <http://katalogo.uea.org?inf=3390>

39 modernaj poetoj. Antologio de la postmilita bosnia-hercegovina poezio. Tradukis L. Režek. ELBIH. Sarajevo. 1985. 110 paĝoj. Prezo: 4.50 €

Klarigo: Okdek naŭ poemoj.

>> Pliaj informoj, mendilo: <http://katalogo.uea.org?inf=1823>

505 elektitaj poemoj el Mannjoo-ŝuu. Tradukis Kenji Ossaka. Tokio. 1958. 212 paĝoj. Prezo: 5.10 €

Klarigo: El la unua japana poezia antologio, 8a jc. p.K.

>> Pliaj informoj, mendilo: <http://katalogo.uea.org?inf=1724>

Al la mondo donace / Svetu na dar. Antologio de la serba porinfana poezio / Antologija srpske poezije za decu. Tradukis Z. Martinov. Strategia. Beograd. 1996. 80+80 paĝoj. Prezo: 6.00 €

Klarigo: 57 poemoj de 22 poetoj.

>> Pliaj informoj, mendilo: <http://katalogo.uea.org?inf=5268>


- Fichier html représentant la catégorie Anthologie -

3.4 Entity linking

L'entity linking est un domaine de recherche qui consiste à identifier de façon non ambiguë une entité, notamment en la reliant à un référentiel. Par exemple, les mentions "JRR Tolkien" ou "John Ronald Reuel Tolkien" d'un texte pourraient être reliées à la base de connaissances WorldCat, et plus particulièrement à l'entité <https://www.worldcat.org/wcidentities/lccn-n79005673>. L'entity linking est un processus important dans le contexte du web sémantique.

Dans ce projet, il est intéressant de relier les oeuvres ou personnes rencontrées dans les notices à un référentiel. Cette tâche est relativement complexe pour plusieurs raisons : d'une part, les oeuvres et personnes espérantistes sont peu connues, et il y a de fortes probabilités qu'aucun lien ne puisse être détecté. D'autre part, les oeuvres espérantistes sont soit des traductions d'oeuvres populaires (eg, "les 1001 nuits") soit des oeuvres uniquement écrites dans cette langue. Il existe aussi de nombreuses oeuvres qui sont à propos de l'espéranto (eg, manuel d'apprentissage de la langue ou histoire de l'espéranto). Ces particularités compliquent la tâche de découverte automatique d'un lien.

Pour faire l'entity linking, nous avons choisi la source Wikidata que nous avons interrogé via des requêtes GET sur le lien <https://query.wikidata.org/sparql>. Pour lier une entitée, livre, à celles de wikidata nous avons adopté un algorithme qui commence par segmenter le titre de cette entité en plusieurs mots. Ensuite, chercher sur wikidata tous les livres dont le titre contient l'ensemble de ces mots dans n'importe quel ordre. Le serveur wikidata renvoie alors tous les URIs identifiant ces livres, à partir desquels on peut récupérer des liens vers wikipedia dans différentes langues.



Esperanta antologio

Isbn: <https://www.biblioesperanto.com/isbn/9290170301>


Langue: [esperanto](#)

Voir aussi

UEA

Voir aussi: <http://katalogo.uea.org?inf=1670>

Wikidata

Titre:	Esperanta antologio
Source:	http://www.wikidata.org/entity/Q3978259
Auteur:	William Auld
Editeur:	Stafeto
Date de publication:	1958-01-01T00:00:00Z
Langue:	Esperanto
Articles Wikipedia	eo fa 

4 Conclusion

Un utilisateur peut effectuer une recherche d'un livre par: isbn, titre, auteur ou année d'édition.

- Une recherche par ISBN donne au maximum un résultat.
- Une recherche par titre, auteur ou année d'édition peut ne donner aucun résultats ou plusieurs.

Il est possible d'effectuer une recherche d'un administrateur, responsable de groupe ou membre à partir de son identifiant.

Il est possible également d'ajouter un administrateur, un responsable ou un membre de groupe.

Pour chaque livre, il est possible de vérifier si il est présent dans wikidata. Si c'est le cas nous pouvons récupérer les informations de ce livre disponibles sur wikidata ainsi que les liens wikipédia décrivant ce livre dans plusieurs langues. Il serait d'ailleurs intéressant de pouvoir enrichir wikipédia en y ajoutant les entités que nous avons en locales et qui ne s'y trouvent pas.

Voici quelques perspectives d'évolution pouvant être utiles pour une future utilisation de l'application:

- Pouvoir modifier directement les informations des livres (en possédant les droits nécessaires).
- Voir la localisation du livre souhaité sur une carte.
- Implémenter un système d'emprunt pour que les utilisateurs puissent demander à emprunter les livres qui les intéressent.

Références

Protégé: <https://protege.stanford.edu/>

Jena: <https://jena.apache.org/>

endpoint sparql BNF: <http://data.bnf.fr/sparql/>

Modèle de données de la BNF: <http://data.bnf.fr/fr/opendata>

UEA catalogue: <http://katalogo.uea.org/>

5 Annexes

Ontologie:

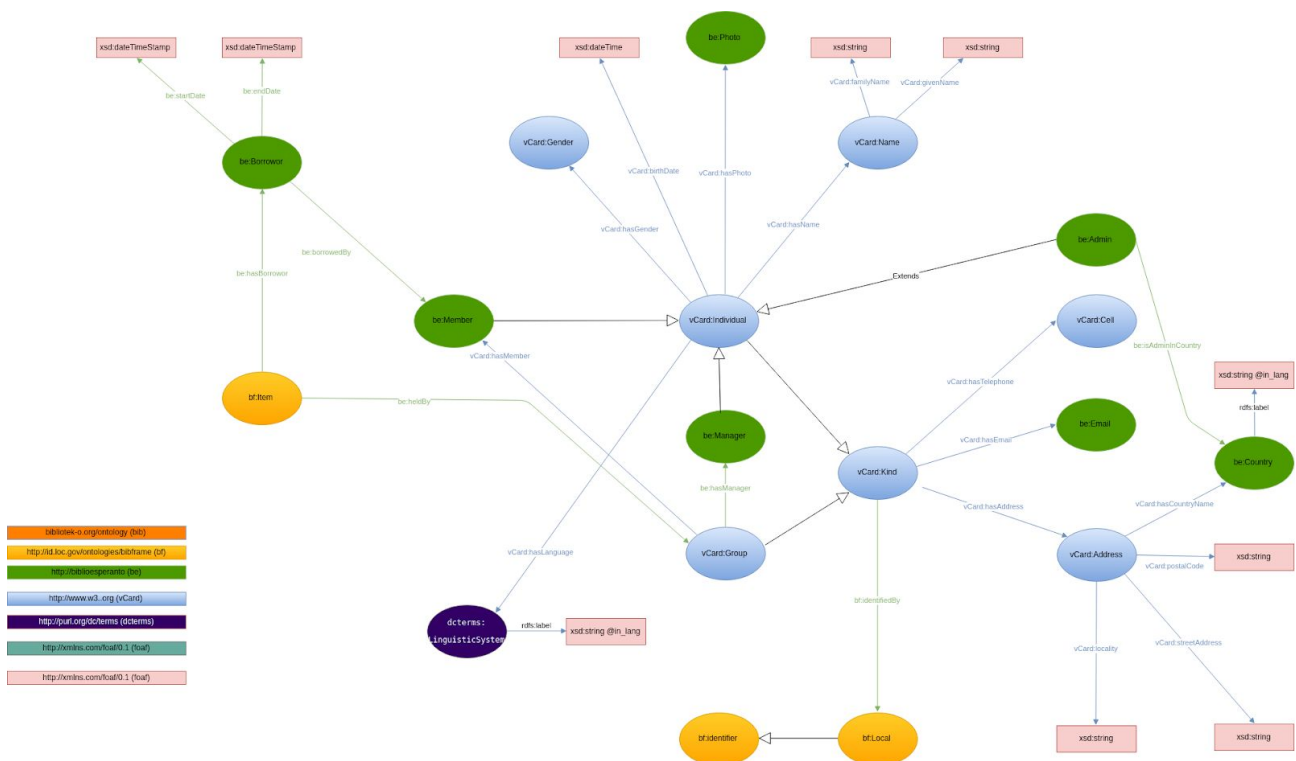
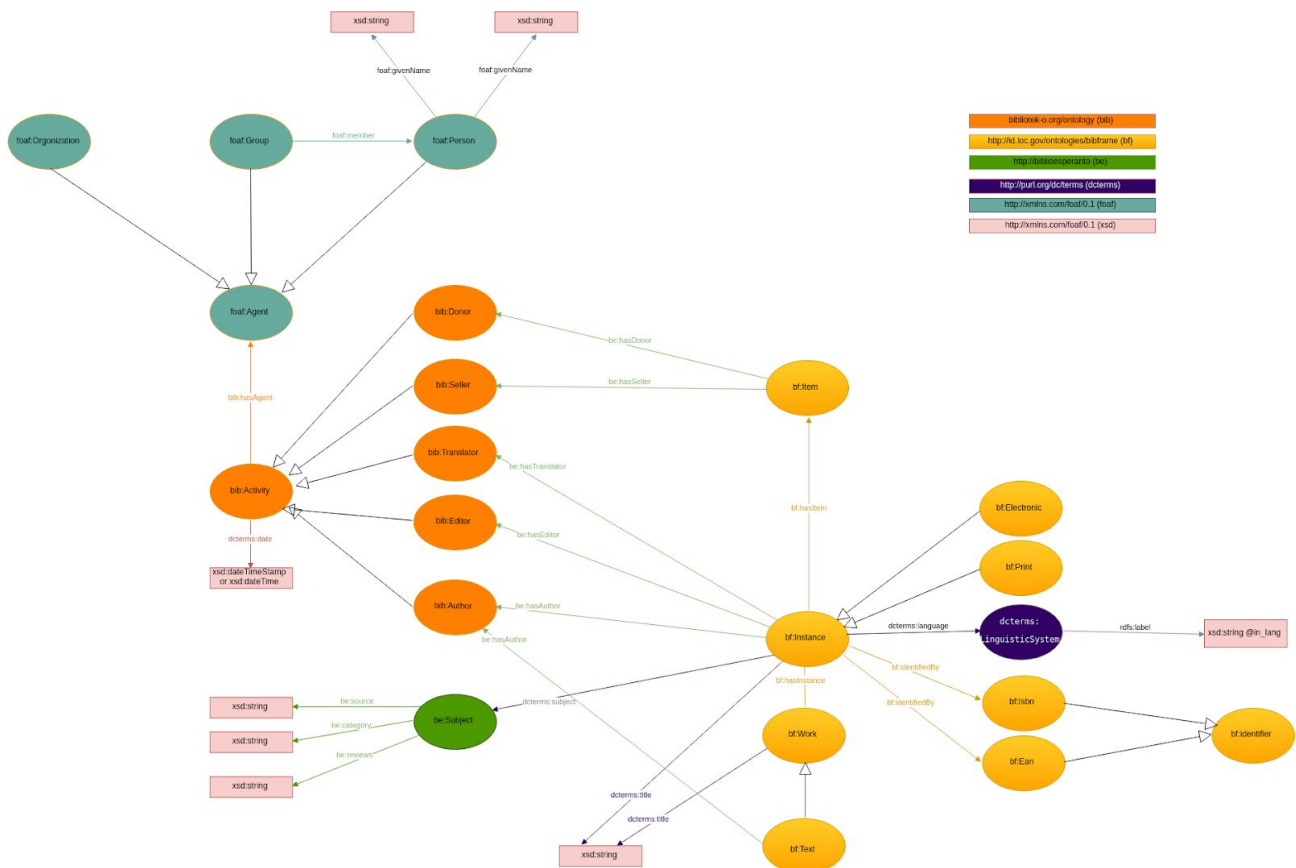
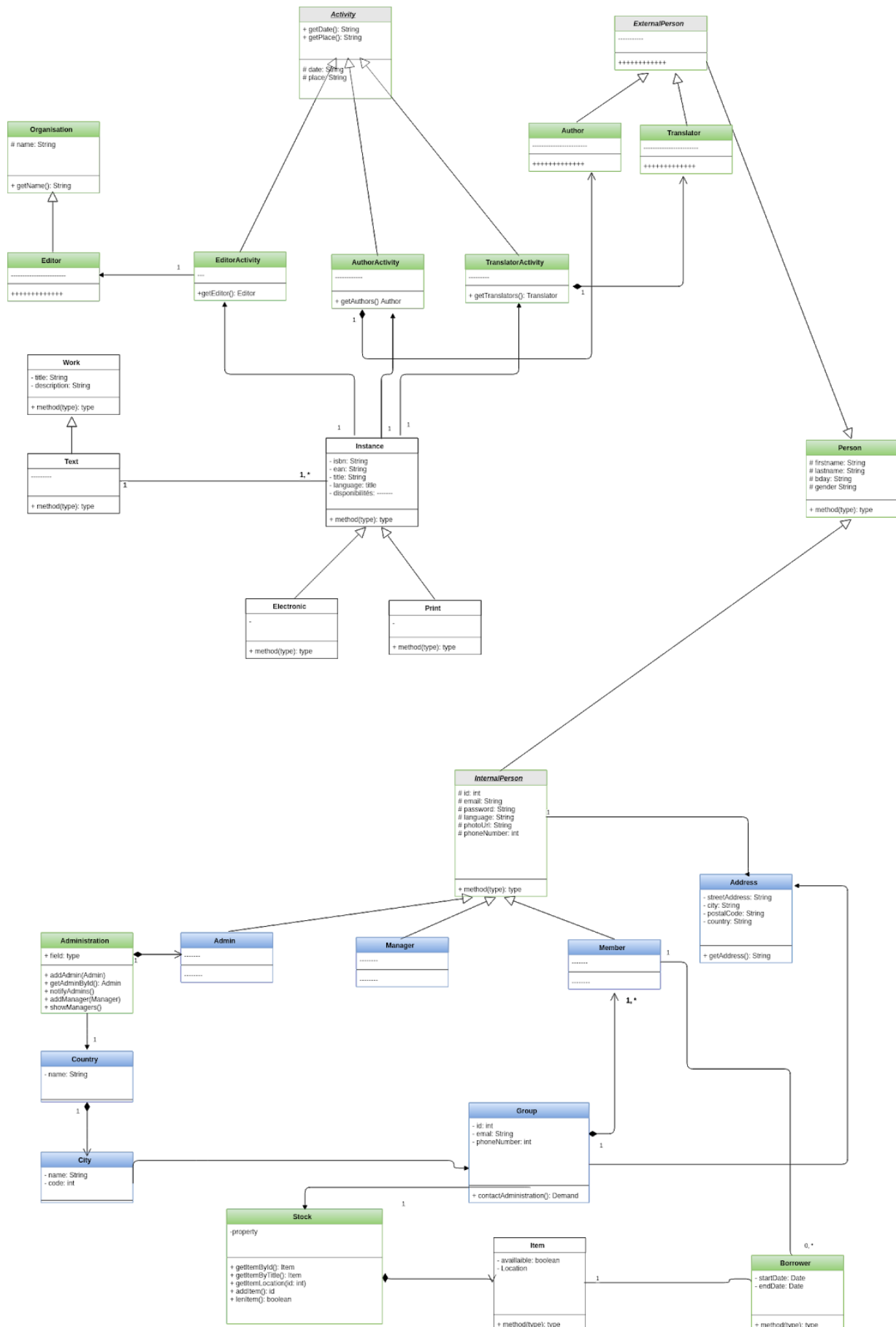


Diagramme de classes:




Recherche par titre:


BiblioEsperanto

Rechercher une oeuvre


Titre ▼ Esperanta Search



Esperanta antologio
ISBN: <https://www.biblioesperanto.com/isbn/9290170301>
Voir +



Kroatia Esperanta poemaro
ISBN:
Voir +



Nova Esperanta krestomatio
ISBN: <https://www.biblioesperanto.com/isbn/9290170433>
Voir +

Recherche par isbn:

BiblioEsperanto

Rechercher une oeuvre

ISBN ▼ 9290170301 Search



Esperanta antologio
ISBN: 9290170301
Voir +

Quelques liens:

Grammaire de be: <https://drive.google.com/open?id=1P3hDyWxMAGvg5rxBYdqJSiZ3VZ1PM7bz>

Maquettes: <https://drive.google.com/file/d/1FRfBC1Ldlf9TpcWcvROncxVIWDLOgsM/view?usp=sharing>