“Sur la recommandation et la réutilisation de ressources éducatives sous licence”.

Contact : Patricia Serrano Alvarado mail, page web.

Encadrantes :

- MC HDR Patricia Serrano Alvarado, Lab LS2N mail, page web.
- CR CNRS Margo Bernelin, Lab. Droit et Changement Social (DCS), mail, page web.

En collaboration avec l’expertise de Colin de la Higuera, Chaire UNESCO RELIA, mail

Mots clés : Ressources éducatives sous licence, interface utilisateur, Machine learning, Web sémantique, graphe de connaissances.

Axe(s) AiBy4 : IA pour les humains.

Équipes impliquées : LS2N-GDD, Chaire UNESCO RELIA, DCS.

Demande de financement : bourse de thèse complète.

Interdisciplinarité : Patricia Serrano Alvarado (MC en informatique) et Margo Bernelin (CR en droit) ont entamé une collaboration interdisciplinaire en 2019 dans le contexte de la thèse de Benjamin Moreau et le projet régional DataSanté. Margo a apporté un point de vue juridique aux travaux sur la compatibilité des licences de Benjamin Moreau (thèse dirigée par Patricia Serrano Alvarado). Elles poursuivent leur collaboration dans le projet CominLabs CLARA et le projet d’amorçage de la MSH Ange-Guépin MétaDroit.

Impacts local et national : Le projet CLARA inclut comme partenaires la Chaire UNESCO RELIA et la Direction du Numérique pour l’Éducation du Ministère de l’Éducation nationale. Dans les deux cas les questions traitées dans CLARA, et donc intimement liées avec le sujet de thèse proposé, sont importantes : l'Unesco fait des Ressources Éducatives Libres (REL) une priorité pour l'ODD 4, et la Chaire travaille depuis 2017 sur différents projets permettant une meilleure visibilité des RELs, à travers de l'indexation, de l'analyse automatique par différents algorithmes d'intelligence artificielle, et la construction de grands corpus de RELs. Pour la DNE-MEN un objectif important est de permettre aux enseignants de mieux partager leurs cours. Et en cela, le projet CLARA et ce sujet de thèse sont intéressants. Cela offre également des perspectives très concrètes d'expérimentation et de mise en œuvre des résultats.

Résumé : Ce sujet de thèse s’inscrit dans le cadre du projet CLARA qui vise à faciliter la création de ressources éducatives sous licence basées sur des ressources existantes. Le travail dans CLARA se concentre sur la création d’un graphe de connaissances reliant de ressources éducatives sous licence et l’interrogation du graphe avec de requêtes SPARQL dont les contraintes peuvent être assouplies. Les travaux de cette thèse seront consacrés quant à eux à améliorer la qualité d’expérience des enseignants. En particulier, nous cherchons à contribuer avec (a) un système de recommandation de ressources éducatives combinant des techniques du Machine learning et du Web sémantique, (b) une solution pour proposer une licence juridiquement conforme aux licences des ressources utilisées la moins restrictive possible, et (c) une interface Web qui offre aux enseignants la meilleure expérience utilisateur possible.

Contexte et motivation de la thèse

Lorsqu'un enseignant souhaite faire un nouveau cours, il va généralement consulter des manuels ou des livres mais aussi des ressources éducatives sur le Web qui pourraient être réutilisées. Il existe de nombreuses ressources utiles et pertinentes sur le Web (diapositives, vidéos, figures, texte, code, etc.), mais les trouver et les organiser dans un plan de cours est un défi. De plus, l'enseignant peut faire face à des problèmes de droits d'utilisation car il est illégal de combiner des ressources si leurs licences ne sont pas compatibles avec la licence du nouveau cours. Il s'agit alors du non-respect du droit d'auteur, aussi nommé contrefaçon, lequel est puni par le code de la propriété intellectuelle (art. L. 335-2 s.). L'utilisateur en infraction pourra alternativement voir sa responsabilité contractuelle engagée sur le fondement de la licence non respectée. Ainsi, l'analyse des clauses contractuelles des licences pour la construction d'un cours est un véritable enjeu et, face au manque de temps disponible des enseignants pour ce faire, son automatisation est nécessaire. En effet et idéalement, l'analyse des ressources disponibles ainsi que la vérification de leurs licences devrait être rapide.

Le projet CLARA[1], financé par le LABEX CominLabs, vise à faciliter la création de nouvelles ressources éducatives sous licence à partir de ressources existantes[2]. Il existe de nombreuses ressources pédagogiques sous licence et réutilisables qui ne peuvent pas être découvertes car elles ne sont pas bien *connectées*. Les annotations sémantiques lisibles par machine nous permettront de connecter et d'enrichir les ressources éducatives grâce à des ontologies bien connues[3]. Les correspondances sémantiques apporteront une valeur ajoutée considérable lors de la recherche de ressources pertinentes. L'interrogation du graphe avec des requêtes SPARQL dont les contraintes peuvent être relaxées, grâce à la sémantique, permettra de garantir des réponses non vides. De plus, nous visons à enrichir les ressources en les connectant via des liens identifiés par des algorithmes de Machine learning (tels que TF-IDF pour trouver les KNN). Ainsi, *basée sur un plan de cours, CLARA proposera un ensemble pertinent de ressources éducatives ayant des licences compatibles.*

Cependant, le succès des contributions du projet CLARA dépendent de leur facilité d'utilisation, c'est-à-dire, de la qualité d'expérience des enseignants qui les utilisent. Sans connaissances en informatique ou juridiques, un enseignant devrait pouvoir bénéficier aisément des apports de CLARA. Il doit pouvoir : (a) fournir facilement un plan de cours (thèmes à traiter, niveau de la formation, domaine du cours, licence du cours, etc.) et obtenir des recommandations de ressources pertinentes pour son cours, (b) savoir quelle licence pourra protéger son nouveau cours conformément aux licences des ressources utilisées, et (c) affiner et visualiser les résultats de ses recherches afin de sélectionner au mieux les ressources qu'il juge pertinentes.

---

[1] https://project.inria.fr/clara/

[2] Ressources du projet européen X5GON https://platform.x5gon.org/products/feed/, ou ressources du Ministère de l'Education Nationale comme https://edubase.eduscol.education.fr/, http://www.sup-numerique.gouv.fr, https://www.canal-u.tv/.

[3] Nous utiliserons les standards du web sémantique (OWL, RDF, RDFS, RDF-star, SPARQL) afin de construire un graphe de connaissances de ressources éducatives enrichies et liées. Ce graphe de connaissances sera relié à d'autres graphes de connaissances comme celui de DBpedia et Wikidata.

Objectifs

L'objectif général de cette thèse est de rendre les ressources pédagogiques accessibles aux enseignants afin de faciliter la création de nouveaux cours. Nous envisageons de proposer :

1. de solutions de recommandation de thèmes et de ressources éducatives le plus pertinentes possibles et cela de manière efficace et rapide,
2. une solution pour proposer une licence conforme aux ressources utilisées la moins restrictive possible,
3. une interface Web qui apporte aux enseignants la meilleure expérience utilisateur possible.

Afin de recommander à l'enseignant les ressources pédagogiques les plus pertinentes, nous envisageons de marier les techniques de Machine Learning avec le Web sémantique [1][2]. L'idée est d'injecter des informations sémantiques dans les systèmes de recommandation. Cela pourrait améliorer la qualité des recommandations mais aussi leur interprétabilité. Par exemple, nous envisageons d'annoter sémantiquement les modèles des ressources éducatives issus des techniques telles que TF-IDF, ce qui permettra de calculer une similarité sémantique entre les ressources afin de pouvoir recommander à l'enseignant des ressources (les KNN par exemple) basés sur une distance sémantique [3].

Concernant l'objectif 2, l'enjeu est de fournir une ou plusieurs licences conformes juridiquement aux licences des ressources à combiner. Les travaux à développer seront dédiés à la description sémantique des licences, la vérification de leur compatibilité et la proposition d'une licence la moins restrictive possible et conforme aux licences des ressources utilisées.

En ce qui concerne l'objectif 3, nous avons l'intention de fournir une interface Web qui apporte la meilleure expérience utilisateur possible. Pour cela, nous aurons besoin d'interagir avec les enseignants tout au long du développement de nos contributions afin de masquer au mieux les technologies utilisées dans le projet (OWL, RDF, RDFS, SPARQL, etc.). L'objectif est de produire des interfaces utilisateur intuitives, ergonomiques et bien adaptées aux pratiques et habitudes des enseignants.

Les figures de l'Annexe A, montrent deux pages d'une maquette de l'application Web recherchée dans cette thèse. L'enseignant pourra saisir le domaine de son enseignement (informatique, mathématiques, etc.) et une liste de thèmes. Il pourra également filtrer les ressources qui l'intéressent par licence, langue, auteur, format du média, etc. Il pourra affiner la recherche de ressources en base à des recommandations et sauvegarder progressivement les ressources qui l'intéressent le plus.

Les résultats de nos travaux de recherche seront utilisés pour recommander des ressources liées aux thèmes donnés par l'enseignant. L'outil aidera l'enseignant avec l'auto-complétion et la vérification de chaque thème saisi pour permettre de ne saisir que des thèmes dont les ressources existent dans le graphe de connaissances. L'enseignant aura aussi la possibilité d'explorer visuellement les concepts liés sémantiquement aux thèmes renseignés avec la possibilité de compléter et de modifier les thèmes choisis. Il pourra également renseigner la licence qu'il souhaite utiliser pour protéger son cours. Ces fonctionnalités pourront être utilisées à tout moment et même avant d'envoyer une requête, puisque seul un ensemble de

thèmes est nécessaire. Ensemble, ces fonctionnalités permettront de limiter le nombre de requêtes (SPARQL) envoyées au graphe de connaissances. Ces requêtes peuvent nécessiter des techniques de raisonnement ou de relaxation de contraintes, ce qui pourrait entraîner des temps d'attente importants pour l'utilisateur.

Travaux existants

Sur la recommandation de ressources éducatives

Diverses expérimentations basées sur l'IA ont été réalisées sur le contenu des ressources éducatives collectées dans le cadre du projet X5GON. Nous en citons ici quelques-unes. Basé sur un score Pagerank, Wikifier [4] aide à déterminer les concepts les plus importants dans un document. Les concepts sont annotés sémantiquement à l'aide des ontologies DBpedia et Wikidata. [5] est un outil utilisant Wikifier afin d'annoter des ressources pédagogiques. Il vise à guider un utilisateur dans son apprentissage en fragmentant les ressources et en annotant leurs fragments. Ces annotations donnent alors à l'utilisateur un aperçu des sujets abordés dans une ressource ou dans ses fragments et permettent la navigation vers d'autres ressources. Utilisant les réseaux de neurones récurrents siamois, [6] propose d'ordonner des ressources pédagogiques de manière à identifier une progression d'apprentissage pédagogique. Enfin, en utilisant trois techniques différentes de représentations sémantiques des documents (Doc2Vec [7], TF-IDF et Wikifier) il est possible de déduire les concepts impliqués dans les ressources éducatives, et ainsi calculer les k plus proches voisins[4] de chaque ressource grâce à des mesures de distance comme le cosinus.

Notre objectif est de proposer un système de recommandation utilisant des connaissances sémantiques. Pour cela nous pouvons utiliser les techniques utilisées dans X5GON pour calculer les modèles des ressources éducatives. Tout d'abord, il est possible de sémantiser les distances déjà calculées dans X5GON afin de les intégrer dans le Web des données. Il est également possible d'utiliser les annotations sémantiques obtenues avec Wikifier pour calculer une distance sémantique entre deux ressources. Cette distance ne tiendrait pas seulement compte de la similarité des vecteurs de concepts mais aussi de la similarité sémantique entre les différents concepts, c'est-à-dire l'ontologie. Enfin, il est également possible de sémantiser les ressources elles-mêmes et ainsi de pouvoir les intégrer au Web des données. Ces trois apports du Web sémantique pourraient permettre de calculer une distance plus fidèle entre les ressources et ainsi essayer d'améliorer la qualité des recommandations.

Sur les licences lisibles par machine et leur compatibilité

L'analyse automatique des licences est possible grâce à des licences lisibles par machine. Les langages d'expression tels que CC REL[5] ou ODRL[6] [8] permettent une description fine des licences. D'un point de vue informatique, des solutions existent pour établir la compatibilité entre licences automatiquement [9][10]. CaLi [9] est un modèle basé sur un treillis qui positionne automatiquement une licence sur un ensemble de licences en termes de

[4] POC des KNN: https://wp3.x5gon.org/#/overpsychologyview/134366
[5] https://creativecommons.org/ns
[6] https://www.w3.org/TR/odrl-model/

compatibilité et de conformité. L'originalité de CaLi est de passer par une relation de restrictivité afin d'ordonner partiellement les licences en termes de compatibilité et de conformité.

D'un point de vue juridique, établir la compatibilité entre des licences est une tâche complexe [11], nécessaire et paradoxalement peu investiguée par les juristes qui préfèrent dédier leur étude à un type de licence en particulier, plutôt qu'à comparer des licences différentes entre elles [12]. Il est pourtant crucial de s'intéresser à la compatibilité, ou plutôt à l'incompatibilité juridique des licences, tant leur multiplication dans le champ de la science ouverte est créatrice de frein à la circulation des informations plutôt que d'opportunités nouvelles [13]. Comme le souligne De Filippi "l'interopérabilité légale" devient tout aussi centrale que "l'interopérabilité technique" entre les licences et les ressources couvertes [14]. En effet, en l'absence de compatibilité entre les licences des ressources exploitées, l'utilisateur se retrouve en infraction avec le droit d'auteur.

Dans ce cadre, la présente recherche permettra d'étendre CaLi afin de développer un volet théorique concernant le concept de compatibilité juridique des licences. Les travaux doctoraux prendront ici appui sur le droit des contrats et sur les instruments juridiques dédiés aux partages des données à l'image de la directive de l'Union européenne n°2019/790 du 17 avr. 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique, afin de proposer une analyse renouvelée de la notion de compatibilité juridique. D'autre part, la recherche s'attachera aux raisons expliquant la complexité d'établir la compatibilité juridique entre les différentes clauses contractuelles de ces documents et montrera qu'elles peuvent être les solutions techniques pour y remédier, faisant ici le pont entre le droit et l'informatique.

Les réflexions menées sur la compatibilité juridique entre les licences nous permettront d'apporter de nouvelles contributions sur la compatibilité automatique entre licences. Ainsi, en nous basant sur l'approche CaLi, nous espérons proposer un mécanisme capable de produire la licence la moins restrictive possible, juridiquement valable et conforme à un ensemble de licences. Cela nous permettra de fournir à l'enseignant une licence pour protéger son cours en format lisible par machine mais aussi en langage naturel.

Sur les interfaces utilisateur

Dans le projet CLARA, à partir d'informations fournies par le projet X5GON, nous sommes en train d'intégrer des ressources éducatives existantes dans un graphe de connaissances. Les requêtes interrogeant un graphe de connaissances peuvent être très complexes (elles peuvent contenir des jointures, des unions, des filtres, des contraintes optionnelles, un contexte, etc.). Il faut connaître le schéma du graphe et le vocabulaire utilisé (c'est-à-dire les ontologies décrivant les données) ainsi que le langage de requêtes SPARQL. Un défi important pour le Web des données est de savoir comment faciliter l'accès et l'exploration des graphes de connaissances aux utilisateurs finaux [15].

A notre connaissance, aucune interface existante ne permettrai aux enseignants d'explorer le graphe de connaissances des ressources éducatives à partir d'un plan de cours. Parmi les approches existantes, Sparklis [16] combine de manière intégrée différents paradigmes : recherche par facettes, générateurs de requêtes et interfaces en langage naturel. Dans nos

travaux, nous nous inspirerons de ce type de contribution afin de concevoir une interface adaptée aux utilisateurs finaux que sont les enseignants.

Plan de travail

Le tableau suivant montre la planification des travaux de cette thèse. La première partie sera consacrée à la proposition d'un système de recommandation hybride utilisant des techniques du Machine learning et du Web sémantique. La deuxième partie portera sur le travail qui permettra de toujours proposer une licence conformément aux licences des ressources utilisées. La troisième partie est répartie dans le temps de manière à ce que des tests avec les utilisateurs (c'est-à-dire les enseignants) soient effectués pour apporter un feedback sur nos contributions.

Planning en mois

Tâches	Année 1				Année 2				Année 3			
	T3	T6	T9	T12	T15	T18	T21	T24	T27	T30	T33	T36
Sur la recommandation de resources éducatives												
Etude de techniques basées sur le contenu permettant d'analyser diverses mesures de similarité entre resources éducatives.												
Proposition d'une méthode de recommandation de resources hybride basée sur de techniques de Machine Learning et Web sémantique.												
Implémentation et expérimentation.												
Sur la compatibilité de licences												
Analyse de la combinaison des licences d'un point de vue juridique.												
Proposition d'une méthode de création automatique de licences conformément aux licences des resources utilisées dans un cours et valide juridiquement.												
Implémentation et expérimentation.												
Sur l'expérimentation et test												
Implémentation d'une interface utilisateur permettant de recommander de resources aux enseignant en base à des plan de cours.												
Expérimentation et test de cas d'utilisation.												
Manuscrit de thèse et articles												
Rédaction d'articles de recherche.												
Manuscrit de thèse et soutenance.												

Références

[1] Vito Walter ANELLI, Vito BELLINI, Tommaso DI NOIA, and Eugenio DI SCIASCIO. "Knowledge-Aware Interpretable Recommender Systems." Studies on the Semantic Web 47 (2020): 101-124.

[2] Vito Walter ANELLI. "Knowledge-Enabled Recommender Systems in the Linked Data Era." PhD, Politecnico de Bari, Department of Electrical and Information Engineering (2019).

[3] Alexandre PASSANT. "Measuring semantic distance on linking data and using it for resource recommendations." Association for the Advancement of Artificial Intelligence Spring Symposium Series (2010).

[4] Brank JANEZ, Gregor LEBAN, and Marko GROBELNIK. "Annotating documents with relevant wikipedia concepts." Proceedings of SiKDD Conference on Data Mining and Data Warehouses (2017).

[5] Sahan BULATHWELA, Stefan KREITMAYER, and María PÉREZ-ORTIZ. "What's in it for me? augmenting recommended learning resources with navigable annotations." Proceedings of the 25th International Conference on Intelligent User Interfaces Companion (2020).

[6] Victor CONNES, Colin DE LA HIGUERA, and Hoel LE CAPITAINE. "What should I learn next? Ranking Educational Resources." Annual Computers, Software, and Applications Conference. IEEE (2021).

[7] Quoc LE, and Tomas MIKOLOV. "Distributed representations of sentences and documents." International conference on machine learning. PMLR (2014).

[8] Iannella RENATO, and Serena VILLATA. "ODRL information model 2.2." W3C Recommendation (2018).

[9] Moreau BENJAMIN, Patricia SERRANO-ALVARADO, Matthieu PERRIN, and Emmanuel DESMONTILS. "Modelling the compatibility of licenses." European Semantic Web Conference. Springer, Cham (2019).

[10] Guido GOVERNATORI, Antonino ROTOLO, Serena VILLATA, and Fabien GANDON. "One license to compose them all." International semantic web conference. Springer, Berlin, Heidelberg (2013).

[11] Margo BERNELIN. "The compatibility of open/free licenses: a legal imbroglio." International Journal of Law and Information Technology 28.2 (2020): 93-111.

[12] Mélanie DULONG DE ROSNAY. "Traduction et localisation des licences Creative Commons" Net.lang : Réussir le cyberespace multilingue (2012): 239-244.

[13] Niva ELKIN-KOREN. “What Contracts Cannot Do: The Limits of Private Ordering in Facilitating a Creative Commons”, 74 Fordham L. Rev. 375 (2005).

[14] Primavera de FILIPPI. “Copyright Law in the Digital Environment : Private Ordering and the regulation of digital works”, Academic Publishing GmbH, 2012.

[15] Jakub KLÍMEK, Petr ŠKODA, and Martin NEČASKÝ. "Survey of tools for linked data consumption." Semantic Web 10.4 (2019): 665-720.

[16] Sébastien FERRÉ. "Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language." Semantic Web 8.3 (2017): 405-418.

Annexe A. Maquette d'interface Web.

Figure 1. Maquette d'application Web pour la recherche de ressources éducatives sous licence. Des nouveaux thèmes seront proposés grâce à des techniques d'IA (*Topics related to yours*).

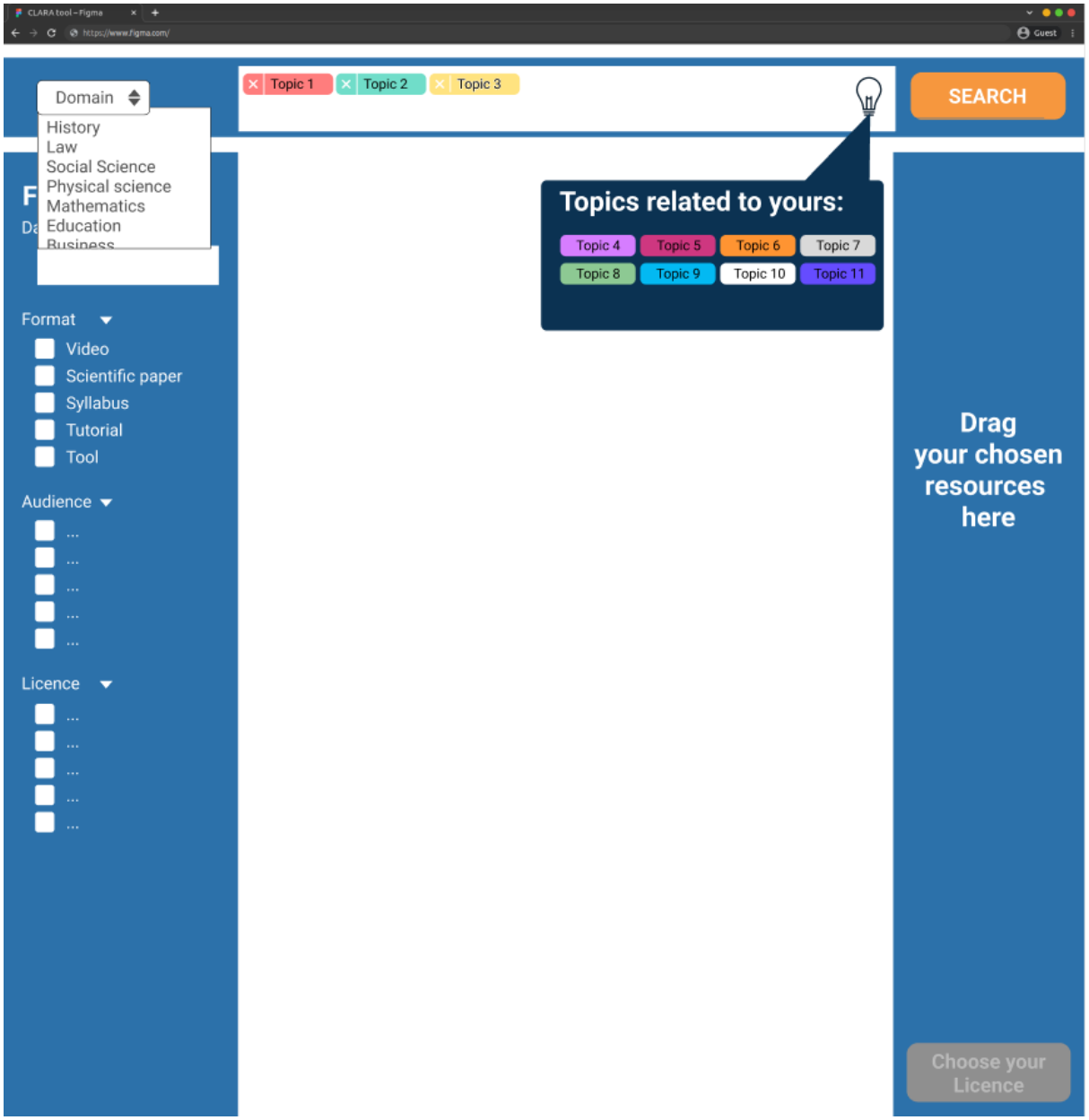

Figure 2. Page de la maquette Web montrant : un ensemble de ressources éducatives résultat d'une recommandation (milieu), de ressources gardées de côté par l'enseignant (droite), et une visualisation sur la navigation parmi les concepts liés aux thèmes données par l'utilisateur (haut).