Comment construire un algorithme de recommandation alternatif aux modèles dominants chez les GAFAM ?

Samuel Gantier

COMMENT CONSTRUIRE UN ALGORITHME DE RECOMMANDATION ALTERNATIF AUX MODÈLES DOMINANTS CHEZ LES GAFAM ?

SAMUEL GANTIER

Depuis une dizaine d'années, un nombre pléthorique de plates-formes de vidéos à la demande propose une offre audiovisuelle et cinémato-graphique abondante. En 2021, sur le seul territoire français, le CSA (aujourd'hui Arcom – Autorité de régulation de la communication audiovisuelle et numérique) recensait 78 plates-formes disponibles sur abonnement (SVOD) et/ou permettant un visionnage à l'acte (VOD)[1].

Une offre cinématographique surabondante sur Internet

L'avènement du modèle d'affaires sur abonnement, encouragé par des financements du Centre national du cinéma (CNC) et du programme européen Media, ainsi que par la baisse des coûts techniques, a permis l'émergence d'une grande variété d'offres cinématographiques tous genres confondus : cinéma patrimonial, série, court-métrage, fiction art et essai, documentaire d'auteur, film d'animation… Comme dans d'autres secteurs des industries culturelles, le marché de la vidéo à la demande se caractérise par

la présence d'un petit nombre d'acteurs qui dominent la filière de manière hégémonique en proposant des offres généralistes (Netflix, Disney+, Prime Video, MyCanal…). À l'autre extrémité, on trouve un grand nombre de petites plates-formes indépendantes qui se partagent des niches spécialisées (Mubi, UniversCiné, LaCinetek, Benshi, Tënk, Brefcinema…). Ces plates-formes cinéphiles s'inscrivent dans la logique de l'économie de la « longue traîne » dans la mesure où elles proposent à leurs abonnés des œuvres *a priori* peu rentables mais qui, en étant diffusées en ligne, trouvent l'opportunité d'être vues par un public dispersé géographiquement.

Différents modes de recommandation des films

Aujourd'hui, quiconque souhaite visionner légalement un film sur Internet doit faire face à un double défi : tout d'abord adopter une plate-forme parmi une concurrence accrue, puis fouiller des catalogues de plusieurs milliers de titres. Face à cette surabondance de choix, les plates-formes développent des systèmes d'aide à la décision afin de renforcer l'engagement de leur audience. Il est possible de distinguer trois modes de recommandation pour aider l'utilisateur dans ses choix de visionnage.

Le premier niveau de recommandation est éditorial. C'est le cas notamment sur les plates-formes de niches cinéphiles qui revendiquent une qualité de curation des œuvres par des équipes de programmateurs spécialisés. Cette prescription des œuvres est au cœur des discours d'accompagnement et présente l'avantage de prolonger une longue tradition cinéphilique préexistante en festival ou ciné-club. Cette forme de médiation experte a toutefois l'inconvénient d'être uniquement *top-down* (de l'équipe éditoriale vers son public). Elle s'adresse ainsi de manière identique à tous les usagers, quels que soient leurs attentes, besoins, pratiques culturelles ou contexte d'usage.

Un deuxième niveau de recommandation est social. Il s'agit d'une prescription communautaire qui met en avant les appréciations de goût des usagers par l'intermédiaire de systèmes de notations, commentaires, indices de popularité, etc. Cette pratique, largement répandue sur le Web collaboratif, pose la question de la confiance accordée à un tiers pour s'approprier un avis et donner du crédit à une recommandation. Sans un « contrat de confiance » entre la personne qui recommande et son interlocuteur, la prescription s'avère généralement infructueuse.

Un troisième niveau de recommandation consiste enfin à déployer des dispositifs algorithmiques qui permettent de filtrer de manière automatisée et systématique des catalogues de plusieurs milliers de titres. Mais qu'est-ce au juste qu'un algorithme de recommandation ? Comment ces dispositifs pourraient-ils aider

à améliorer la rencontre entre la création cinématographique et son public ? Si l'on y regarde de plus près, quels sont les apports et les biais de ces intermédiaires qui orientent nos choix de consommation culturelle ?

De manière générale, un algorithme se définit comme une suite finie et non ambiguë d'instructions permettant de résoudre un problème. Pour ce faire, un jeu de données entrantes (*input*) est soumis à des opérations de calcul afin de produire un résultat (*output*) censé apporter une solution au problème initial. Pour des raisons industrielles, commerciales ou de marketing, certains éditeurs de logiciels dissimulent le fonctionnement de l'algorithme qu'ils proposent. D'où l'appellation largement répandue de « boîte noire », qui évoque une technologie informatique opaque et trop complexe à appréhender pour un non-spécialiste. Pour un spécialiste en algorithmie, il est fréquent que la méthodologie employée ne soit pas accessible directement et qu'il faille recourir à une enquête spéculative de rétro-ingénierie.

Dans le langage courant, l'algorithme désigne de manière métonymique tout le système de recommandation, alors qu'il ne représente en réalité qu'un élément du dispositif social et technique, qui se décompose en différents niveaux intriqués (du recueil des données à l'implémentation dans l'interface utilisateur). Pour un grand nombre de professionnels de la culture, les algorithmes de recommandation nuiraient *de facto* à la diversité culturelle, quelles que soient la manière dont ils sont conçus et les valeurs auxquelles adhèrent leurs concepteurs. Ainsi, les questions le plus souvent soulevées sur l'éthique des algorithmes concernent la protection de la vie privée, la transparence dans le traitement des données, l'applicabilité des résultats, l'équité de traitement entre individus ou encore la supervision du système automatisé par un humain. Ces questionnements éthiques sous-entendent que l'humain délègue une partie de ses responsabilités à la machine pour établir une tâche dans un secteur d'activité donné. Mais comment garantir que la supervision humaine d'un expert soit pertinente et non porteuse également de biais de jugement ? De façon générale, les études soulignent qu'il est tout aussi important d'interroger les biais dans la construction des données que le processus de calcul informatique en lui-même. Phénomène que les informaticiens désignent trivialement par l'expression « *garbage in, garbage out* », soulignant ainsi que, faute de données cohérentes et fiables à l'entrée, la sophistication du calcul algorithmique est vouée à l'échec pour construire du sens à la sortie.

Différentes approches de conception des algorithmes de recommandation

Les systèmes de recommandation constituent un dispositif automatisé de filtrage et de tri d'informations désignant une ou plusieurs caractéristiques d'un document (page Web, biens culturels ou marchands…). Selon la nature des métadonnées exploitées, il est possible de définir les dispositifs de recommandation selon trois grands types de démarches.

Premièrement, les approches centrées sur les données de navigation des utilisateurs, appelées également « filtrage collaboratif », se basent sur l'hypothèse que les personnes ayant aimé des contenus identiques par le passé possèdent un goût similaire et seront enclines à apprécier les mêmes contenus dans le futur. Basées sur l'ensemble des interactions des utilisateurs, ces méthodes présentent l'avantage de recommander des objets complexes sans avoir à les analyser en amont. L'un des exemples les plus connus est l'algorithme *Item-to-item Collaborative Filtering*, développé en 1998 par Amazon et qui a participé à sa croissance fulgurante dans le secteur de l'e-commerce. Dans le champ culturel, certains acteurs utilisent des métadonnées telles que critiques, évaluations ou commentaires sur les blogs et médias sociaux pour créer leurs jeux de données. Toutefois, on peut se demander s'il est véritablement pertinent de recommander un bien culturel, fortement porteur d'identité sociale et symbolique, avec les mêmes technologies que celles de l'e-commerce. De plus, l'explicabilité et l'adhésion du public au dispositif de recommandation semblent une condition préalable indispensable pour permettre son appropriation sociale.

Deuxièmement, les démarches centrées sur le contenu exploitent des métadonnées qualifiant de larges catalogues de documents qu'elles mettent en correspondance avec un profilage de l'utilisateur. Ces approches nécessitent un travail préalable d'indexation sémantique des contenus et impliquent d'adopter des catégorisations discriminantes, comme c'est le cas par exemple avec les genres et sous-genres musicaux, littéraires, audiovisuels ou cinématographiques. Elles présentent l'avantage de ne pas utiliser les données de consultation des utilisateurs, mais nécessitent en contrepartie un coût important d'ingénierie pour générer des jeux de données fiables et pertinents. En évitant de recourir à une personnalisation de la recommandation, l'usage d'algorithmes de recommandation de contenu à contenu tend à limiter l'entretien des utilisateurs dans des « bulles de filtres » forgées sur l'antériorité de leur navigation. En outre, l'explicitation de la recommandation à l'utilisateur est facilitée par l'affichage simultané de la source du calcul algorithmique (le contenu de départ) et de l'objet de la recommandation (le contenu prescrit).

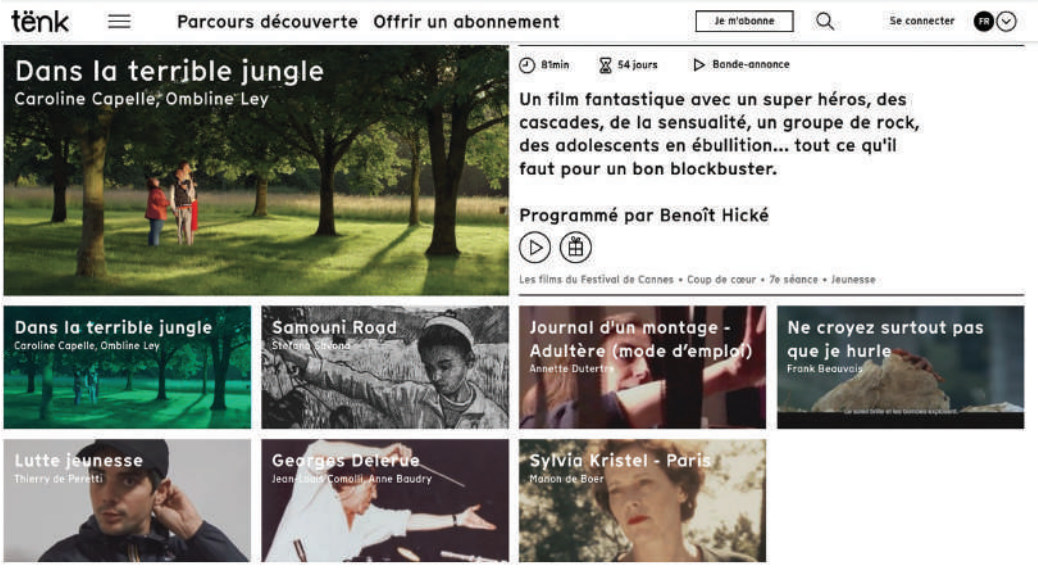
Troisièmement, les approches hybrides combinent démarches centrées sur l'utilisateur et sur le contenu. C'est le cas notamment du célèbre algorithme de Netflix qui associe

l'analyse des catégories de films (genres et sous-genres audiovisuels) et les traces de visionnage des abonnés. Ces techniques sont dynamiques dans la mesure où les données qui entrent dans le système varient en fonction des usages et améliorent ainsi la précision des résultats, comme le mettent en avant les promoteurs de l'intelligence artificielle avec le *deep learning*. Dans tous les cas, que les métadonnées soient collectées à l'insu de l'utilisateur ou fabriquées spécifiquement pour cette application, l'algorithme nécessite une base d'apprentissage importante avant de pouvoir délivrer une série de recommandations pertinentes.

Les algorithmes peuvent-ils aider à promouvoir la visibilité du cinéma d'auteur ?

Les plates-formes qui offrent des dispositifs de recommandation ont tendance à accentuer la visibilité des contenus les plus populaires et à maintenir leurs utilisateurs dans des bulles de filtres hermétiques, voire à proposer des contenus moins convaincants que les explorations autonomes de catalogue par les usagers. De fait, ce type de recommandation est souvent opposé à une prescription éditoriale exigeante et éclairée. La recommandation algorithmique, prédictive et personnalisée, suscite généralement la méfiance des acteurs de la médiation culturelle qui lui reprochent son opacité et sa normativité. Ils expriment parfois leur crainte que leur expertise humaine soit remplacée par des machines. Or, les professionnels de la programmation cinématographique disposent justement des connaissances indispensables pour élaborer des métadonnées capables d'alimenter la recommandation algorithmique de manière qualitative. À l'avenir, les programmateurs impliqués au sein de plates-formes cinéphiles auraient tout intérêt à travailler étroitement avec des professionnels de la documentation afin de fabriquer, générer et structurer leurs propres métadonnées des catalogues de films qu'ils diffusent.

Fort du constat que seuls les grands groupes industriels disposent des moyens humains et matériels pour développer des outils de recommandation dédiés à leurs objectifs commerciaux, le programme de recherche-action AlgoDoc[2] propose de concevoir un dispositif de recommandation adapté aux besoins des plates-formes cinéphiles, quelque peu marginalisées puisque ne pouvant suivre le rythme effréné des



Page d'accueil de la plate-forme Tënk spécialisée dans le cinéma documentaire d'auteur.

innovations d'usage dans le secteur. Pour mener à bien cette étude expérimentale, un consortium recherche-industrie a été construit avec le laboratoire LaRSH de l'Université Polytechnique Hauts-de-France. La plate-forme Tënk a mis à la disposition de l'équipe de recherche un catalogue d'environ 2 200 films documentaires avec leurs métadonnées associées. La société Spideo a pour sa part donné accès à son moteur algorithmique. Cette entreprise d'intermédiation revendique dans son positionnement commercial une approche sémantique qui lui permet d'indexer de larges catalogues audiovisuels multi-genres, tel celui de son client historique, MyCanal.

Une expérimentation menée avec Tënk pour recommander le cinéma documentaire autrement que par la thématique abordée

Créée en 2016, la plate-forme Tënk revendique environ 20 000 utilisateurs en 2022. Son identité éditoriale promeut la défense de la création documentaire contemporaine et patrimoniale, et s'oppose aux programmes télévisés formatés qui excluent la majorité de la création contemporaine pourtant financée par des fonds publics. Le cinéma documentaire présente une grande variété de démarches artistiques, depuis les films ethnographiques

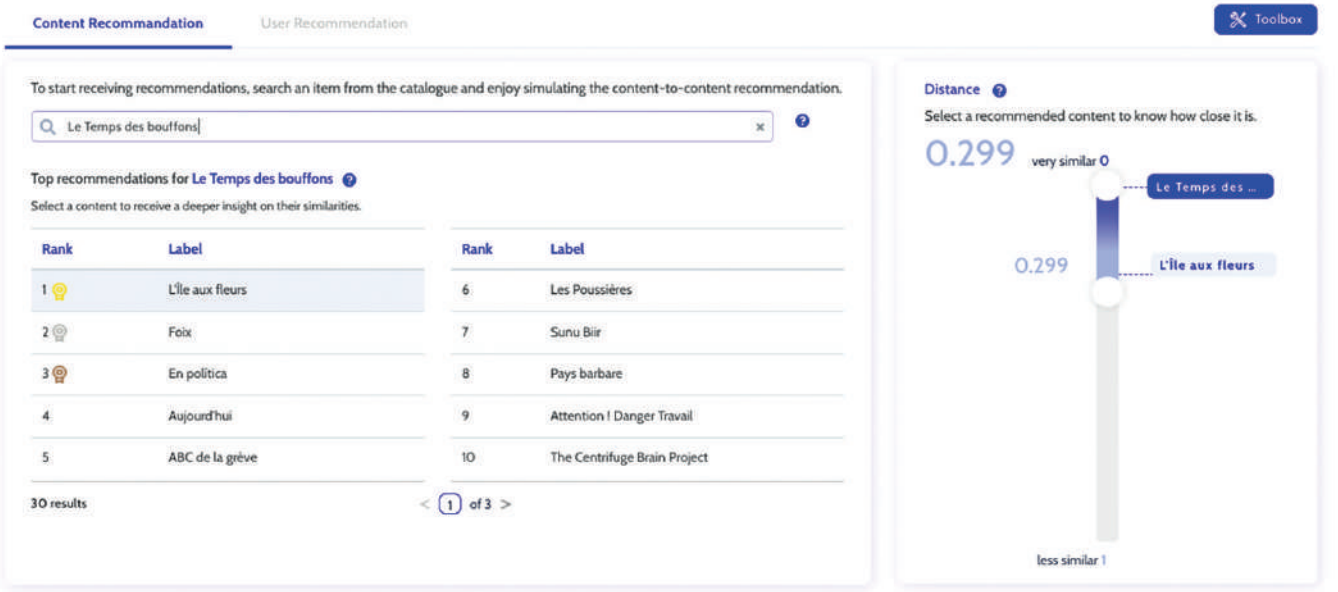


Tableau de bord du moteur algorithmique de Spideo montrant la distance qui sépare les films *Le Temps des bouffons* et *L'Île aux fleurs*.

qui observent en immersion un phénomène social jusqu'aux enquêtes d'investigation sur un fait de société, en passant par les récits autobiographiques intimistes, les démarches didactiques dotées d'une voix off pédagogique, les films historiques qui exploitent des fonds d'archives ou encore les essais expérimentaux basés sur de l'animation.

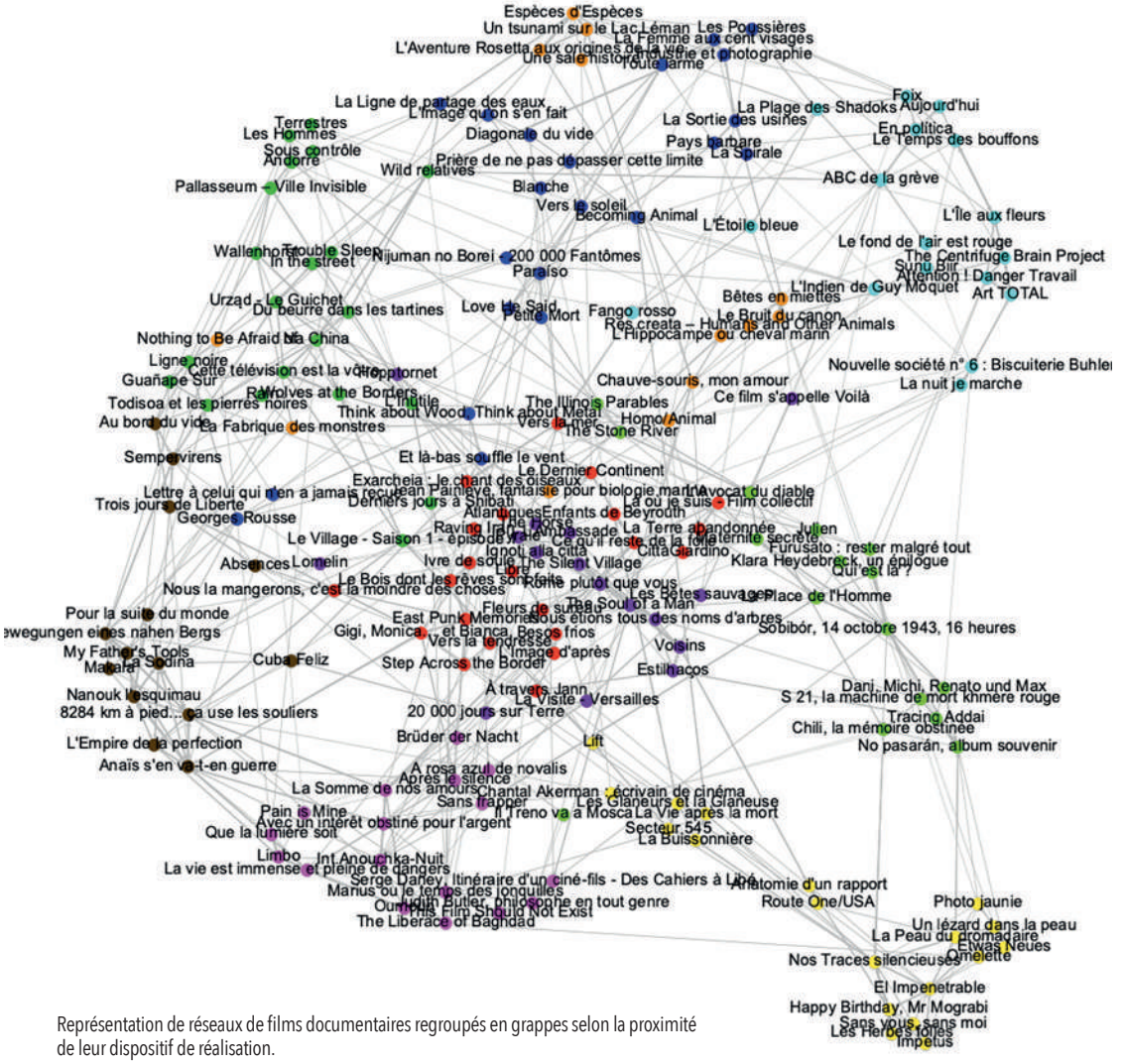
Pour la direction de Tënk, développer un algorithme de recommandation permet de « dépasser une première expérience de visionnage décevante ». Il s'agit de proposer une alternative si le spectateur n'a pas aimé le premier film visionné sur la plate-forme, le jugeant par exemple « trop pointu, voire élitiste ». Ce constat a conduit l'équipe de

recherche à imaginer un jeu de données qui puisse traduire la grande variété des démarches artistiques présentes dans le champ du documentaire. L'attention a donc été portée sur la description des dispositifs de réalisation qui caractérisent le cinéma documentaire. En effet, les plates-formes indexent généralement leur catalogue en qualifiant différentes facettes de métadonnées : auteur, réalisateur, comédiens, date, thématique, genre cinématographique, etc. Or, ces classifications s'avèrent inopérantes pour le segment du cinéma documentaire. La conception d'un algorithme alternatif consiste donc à créer une recommandation basée sur une nouvelle forme de catégorisation du cinéma documentaire alternative aux métadonnées fondées sur une classification thématique des films. L'enjeu est de faire correspondre la nature des données importées dans l'algorithme avec les spécificités cinématographiques et la richesse culturelle du catalogue. Dans cette optique, la médiation algorithmique cherche à s'inscrire en parfaite cohérence éthique, artistique et politique avec la promesse éditoriale de la plate-forme.

Concrètement, cette recherche expérimentale a débuté par la construction d'un thésaurus, c'est-à-dire d'un outil intellectuel en vocabulaire contrôlé, permettant d'indexer la grande variété des dispositifs de réalisation documentaire. À cette fin, 292 concepts ont été définis pour décrire les caractéristiques formelles de chacun des films du catalogue de Tënk. À titre d'exemple, le descripteur « archives audiovisuelles » permet de préciser si le film contient des archives amateurs, documentaires ou fictionnelles ; sur un autre plan, le descripteur « attitude de la personne filmant » indique si cette dernière a une posture agressive, complice, en action, en performance ou en retrait vis-à-vis de la personne filmée ; le descripteur « registre » précise quant à lui la démarche de l'auteur : biographique, autobiographique, dénonciatrice, didactique, onirique, persuasive, polémique, etc. Sur le plan opérationnel, un documentaliste-indexeur a choisi 10 descripteurs, parmi les 292 concepts recensés dans le thésaurus, afin de qualifier finement les traits caractéristiques du dispositif de réalisation pour chacun des documentaires issus du catalogue. Ces métadonnées ont ensuite été ingérées par le moteur algorithmique afin de calculer la distance vectorielle qui les sépare. Plus le score donné par l'algorithme est proche de 0, plus deux films sont considérés comme proches sur le plan formel, comme l'illustre le tableau de bord du moteur algorithmique présenté ci-contre.

En résumé, ce travail d'indexation du catalogue de Tënk a permis d'identifier des familles de films plus ou moins proches selon leurs spécificités formelles, et ce indépendamment du sujet abordé, du réalisateur ou de la période de réalisation. Le graphique suivant présente ainsi différentes « familles de films », regroupés par couleur en fonction des caractéristiques de leur dispositif de réalisation.

Ce dispositif de recommandation soulève l'intérêt d'équiper de manière alternative les plates-formes de vidéos à la demande cinéphiles, mais



Représentation de réseaux de films documentaires regroupés en grappes selon la proximité de leur dispositif de réalisation.

également toutes les institutions qui disposent de vastes fonds, collections ou catalogues de films documentaires. L'enjeu sous-jacent est de participer à la découvrabilité d'œuvres cinématographiques patrimoniales et contemporaines invisibilisées par la grande standardisation du marché télévisuel ces vingt dernières années. En perpétuant des usages sociaux lui préexistant, le recours à un algorithme de recommandation de contenu à contenu constitue un outil dont tous les acteurs de la médiation culturelle devraient pouvoir s'emparer pour valoriser leur catalogue. Le défi des prochaines années pour la médiation numérique des œuvres cinématographiques sur les plates-formes consistera à maîtriser les métadonnées qui alimentent les algorithmes de recommandation afin d'en faire une technologie conforme aux valeurs et aux missions de diffusion de la création cinématographique auprès d'un public élargi.

1. Selon l'étude menée par l'Hadopi et le CSA, *La Multiplication des services de vidéo à la demande par abonnement : stratégies de développement et impact sur les usages*, 9 mars 2021.
2. Programme Algorithme de recommandation de films documentaires financé par la région Hauts-de-France et piloté par le laboratoire LaRSH-DeVisu de l'Université Polytechnique Hauts-de-France.