CapData Opéra : faciliter l'interopérabilité des données des maisons d'opéra

E Peyre, F Amarger, N Chauvat

CapData Opéra : faciliter l'interopérabilité des données des maisons d'opéra

E. Peyre[1], F. Amarger[2], N. Chauvat[2]

[1] Réunion des Opéras de France, ROF
[2] Logilab

eudes.peyre@rof.fr
prenom.nom@logilab.fr

Résumé

Le projet CapData Opéra, mené à l'initiative de la ROF[1] *utilise les technologies du Web sémantique comme fondement d'une solution de structuration et de diffusion des données culturelles capable de répondre aux enjeux de développement des publics, de soutien à la création artistique et d'accès à la culture.*
Cette solution de mutualisation permet d'interroger les données produites par plusieurs acteurs du domaine pour, par exemple, connaître la programmation et la circulation d'une œuvre ou d'une production entre plusieurs maisons d'opéra.
Pour faciliter la publication des données produites par chaque maison d'opéra, la ROF propose une ontologie du domaine, des référentiels, une infrastructure de publication, des outils et de l'accompagnement humain.
Dans cet article, nous présentons les objectifs et les moyens mis en œuvre pour fédérer des données hétérogènes, nous faisons un retour sur expérience en abordant les aspects techniques et la gestion, et nous décrivons les résultats actuels et les perspectives de ces travaux.

Mots-clés

ROF, Opéras, RDF, Interopérabilité

Abstract

The "CapData Opéra" project, initiated by ROF (Réunion des Opéras de France - French Opera Association) and supported by the French Ministry of Culture, uses Semantic Web technologies to share cultural data with the public and the artistic community.
The aim is to aggregate data produced by various domain actors to make it globally searchable. This highlights previously invisible data, such as the exchange of creative works and performances between opera houses. To achieve this, an ontology has been designed to define a common vocabulary and implement data interoperability objectives. This ontology is aligned with schema.org, and we are working to align additional models. A set of SHACL rules has been created to validate the data before publication. A dedicated tool, Rodolf, has been developed to monitor the RDF publishing process. This tool is used to execute the process and track which sources have been uploaded to the SPARQL endpoint, including upload times and any errors encountered. Exporting RDF data can be challenging for institutions unfamiliar with Semantic Web technologies, so a dedicated Software Development Kit (SDK) has been developed to assist web developers in exporting CapData RDF data even if they lack experience in this area.
In this presentation, we aim to share with the SWIB community the objectives and solutions we have found to federate heterogeneous data. We will present feedback on this project, focusing on technical and management aspects, and then describe the results we have achieved and the future of this project.

Keywords

ROF, Opera, RDF, Interoperability

1 Introduction

Les structures et les établissements culturels du spectacle vivant ont connu à la fin des années 1990 une profonde révolution liée aux transformations induites par le développement d'Internet et du numérique. Deux décennies plus tard, les enjeux liés à la diffusion et aux échanges des données produites prennent une importance majeure.
Si les stratégies et solutions développées par les politiques culturelles et les acteurs de ce secteur, dont les maisons d'opéra, ont été menées de manière relativement homogène et précoce en réponse aux enjeux de démocratisation, de création, de valorisation et de médiation auprès des publics, les problématiques liées à la gestion et au partage des données produites n'ont émergé que plus récemment.
Pour mettre en valeur leur programmation et interagir efficacement avec leur public, les services au sein des maisons d'opéra sont amenés à manipuler quotidiennement de nombreux outils numériques, tel que des CMS (*Content Management System*) avec lesquels sont gérés les informations diffusées sur leur site web, des logiciels dédiés à la gestion de billetterie, à la production ou encore de multiples réseaux sociaux. La faible interopérabilité de ces systèmes, couplée à la mise en place de modélisations spécifiques au sein de chaque établissement, conduit à une complexité im-

1. Réunion des Opéras de France

portante dès lors que les données doivent être échangées ou croisées avec d'autres acteurs.
Par exemple, une requête permettant d'obtenir la programmation lyrique ou chorégraphique des maisons d'opéra pour la saison 2023/2024 reste aujourd'hui sans réponse satisfaisante. Ceci est dû au fait que les données, et surtout les modélisations des données, ne sont pas uniformisées.
Dans ce contexte, la Réunion des Opéras de France (ROF), réseau national des maisons d'opéra, scènes et compagnies lyriques, développe au sein de la mission ressources et développement numérique, le projet CapData Opéra. Porté en réponse aux enjeux et besoins identifiés auprès de ses membres et politiques culturelles, ce projet vise au développement d'une solution mutualisée et hautement réplicable afin de favoriser l'interopérabilité, l'échange de données et leur valorisation auprès des publics.
Dans cet article, nous aborderons le sujet du partage de données et de l'intéropérabilité en mettant en lumière les différentes facettes du projet CapData Opéra. Nous débuterons par une présentation du projet et son contexte, en expliquant les objectifs visés, le choix de l'architecture et les solutions déployées pour répondre aux besoins détectés auprès des maisons d'opéra. Nous détaillerons ensuite les outils spécifiquement développés pour faciliter la réalisation de ce projet, en soulignant le rôle essentiel de l'ontologie dans le processus de création et de mise en œuvre. Enfin, nous conclurons en évoquant les perspectives futures du projet, dont l'importance d'une approche partagée, transversale et multi-échelles.

2 Contexte et besoins

Chaque maison d'opéra produit et diffuse sur son site web, ses réseaux sociaux et auprès de la presse et partenaires, sa programmation artistique et des ressources médias (vidéos, photographies, audios et textes) à destination des publics. En leur sein, les services, dont ceux de production ou de communication, produisent des données et des métadonnées de programmation et médias qui peuvent être similaires, mais qui sont souvent saisies à de multiples reprises et stockées dans des bases de données ne permettant que trop faiblement l'échange d'information (bases silotées et faiblement interopérables).
L'absence d'une stratégie généralisée et commune de standardisation, d'identification des données et de mise en place de référentiel au sein des maisons d'opéra et plus largement des lieux de programmation du spectacle vivant, représente un véritable frein au développement de la découvrabilité et à la diffusion des créations artistiques et des contenus auprès des publics.
Il n'existe en effet pas d'identifiant normalisé pour la gestion des productions de spectacle vivant, alors que le secteur du livre utilise l'ISBN [2] au niveau international et l'industrie musicale dispose de l'ISRC [3] utilisé par exemple pour identifier les morceaux diffusés sur les plateformes en ligne[9]. La généralisation de l'usage d'identifiants publics, comme l'ISNI [4] pour les artistes, permettrait par exemple de simplifier la diffusion des mentions et la gestion des droits lors d'une diffusion, ainsi que de développer la transparence lors de la diffusion des médias auxquels ils participent, comme l'ont illustré les expériences de diffusion des données de programmation des maisons d'opéra sur les espaces « #Culture chez nous », ou au sein de l'application du Pass Culture.
La diffusion des représentations programmées par les maisons d'opéra passe aujourd'hui par une succession de saisies manuelles : les services sont invités à saisir pour chaque réutilisation leurs offres d'événement, site web, billetteries, agendas culturels, applications. Malgré l'existence d'une API pour le Pass Culture, la faible structuration des données et interopérabilité au sein des systèmes conduit les services une nouvelle fois à une saisie majoritairement manuelle. L'enquête préliminaire a fait ressortir une moyenne de six doublons de saisie pour les données de programmations, traduisant un fort besoin, pour les équipes et pour la visibilité des contenus, de mise en place d'une solution efficace et puissante pour l'exposition des données. La faible exposition des données des maisons d'opéra sur les services innovants de diffusion musicale représente également un frein considérable à l'émergence de nouvelles expériences susceptibles de répondre aux besoins et usages du public.
Le développement de données structurées et leur exposition depuis les établissements culturels apparaissent donc comme un levier extrêmement puissant pour accroître la visibilité des contenus lyriques et chorégraphiques au sein des nouveaux modes de diffusion, dont les plateformes de *streaming*, enceintes connectées et services innovants. Ce même constat apparaît dans les « freins structurels technologiques à dépasser » de la « Mission exploratoire sur les métavers »[1] commandée par le ministère de l'Économie, des Finances et de la Relance, le ministère de la Culture et le secrétariat d'État chargé de la Transition numérique et des Communications électroniques.
Outre les problématiques d'échanges transversaux, de doublons de saisies des données, l'analyse comparative des services existants a permis d'observer trois freins supplémentaires à l'exposition et la diffusion des données culturelles :

1. Faute de donnée standardisée et exposée depuis les maisons d'opéra, une majorité de services numériques externes est contrainte de demander aux services des maisons une saisie manuelle supplémentaire ou un export spécifique non standardisé des données. Ce fonctionnement entraîne une surcharge importante de travail pour les équipes et limite le rayonnement des données culturelles.
2. Dans le cas minoritaire où le service externe propose une API pour collecter les données produites par les maisons, l'absence de standardisation des données impose des développements spécifiques, faiblement réplicables et un coût financier pour chaque structure. Ce constat fait ressortir un coût important pour les finances publiques sans bénéfices

2. Voir `https://www.isbn-international.org/`
3. Voir `https://isrc.ifpi.org/`
4. Voir `https://isni.org/`

de réplicabilité et de ruissellement à l'ensemble des établissements culturels.

3. Pour combler ces problématiques d'absence de standardisation et d'exposition des données, plusieurs agrégateurs et services utilisent, la technique du *scraping* pour collecter les données de programmation. Cette technique, à l'impact environnemental négatif, n'est pas satisfaisante et implique des développements spécifiques non pérennes pour chaque établissement culturel.
4. Enfin, la protection de la souveraineté des établissements culturels sur les données qu'ils produisent apparaît comme prioritaire. Celle-ci passe par le développement de leur capacité à disposer et à exposer leurs données en toute autonomie.

La mise en place d'une solution mutualisée a pour objectif d'optimiser les coûts d'investissement et de fonctionnement, tout en favorisant son déploiement au sein des maisons d'opéra et potentiellement des établissements intéressés du secteur du spectacle vivant.

3 Projet CapData Opéra

Initié en 2022, le projet CapData Opéra est porté par la ROF en partenariat avec l'Opéra National de Bordeaux, le groupe de travail numérique de la ROF, et le réseau TMNlab. Une première expérimentation a été réalisée dans le cadre de l'appel à projets "Découvrabilité en ligne des contenus culturels francophones"[7] en 2023. Forte de cette première phase, la ROF s'est associée à 6 maisons d'opéras et au réseau TMNlab pour lancer le projet CapData Opéra - France 2030[12], qui s'est inscrit dans le Programme d'investissements d'avenir (PIA4) - *"Expériences augmentées du spectacle vivant"*, une opération soutenue par l'Etat et opérée par la Caisse des Dépôts. Ce nouveau chantier vise à déployer et à industrialiser la solution à plus grande échelle. Ces deux projets font appel à plusieurs prestataires et expertises techniques afin d'assurer la mise en place de la solution de valorisation des données auprès des publics. Ils visent à proposer des solutions déployables au sein des maisons d'opéra participantes, qui peuvent déverrouiller les principaux freins à l'échange et la réutilisation des données. Chaque étape de la chaîne de circulation des données, de leur structuration à leur exposition et réutilisation, fait l'objet d'un travail spécifique et de la mise en place de solutions hautement réplicables, incluant des outils techniques, un accompagnement des partenaires et des prestataires, ainsi qu'une documentation appropriée.

L'exemple des maisons d'opéra illustre les défis rencontrés dans la gestion des données inter et intra sectorielle. Historiquement, chaque maison d'opéra a élaboré son propre système d'information, caractérisé par des schémas et des formats de données hétérogènes.

Depuis 2004, les technologies du Web Sémantique, promues par le World Wide Web Consortium (W3C)[5], offrent une voie vers une interopérabilité accrue grâce aux standards de la famille RDF (Resource Description Framework)[11]. L'adoption d'un format tel que le RDF, et l'utilisation d'une ontologie, ou modèle de données commun, facilitent l'intégration des données en permettant à chaque contribution de s'aligner sur cette ontologie unifiée, assurant ainsi l'interopérabilité et la circulation des données.

Outre la facilitation de l'échange de données, cette approche soutient la souveraineté de chaque maison d'opéra sur ses propres données, en lui permettant de contrôler la manière dont elles sont partagées. Dans ce sens, le choix de l'architecture sélectionnée vise à répartir de manière équilibrée et dans le respect de la souveraineté de chaque partenaire l'enjeu de responsabilité de publication. Les maisons participantes publient elles-mêmes leurs données au format RDF, se rapportant spécifiquement à leurs besoins et stratégies de diffusion. Ces données sont ensuite collectées au sein d'un entrepôt SPARQL. La gestion mutualisée de cet entrepôt favorise l'interrogation des informations issues de l'ensemble des maisons d'opéra participantes au projet.

Cette démarche met en lumière l'importance du Web Sémantique dans le renforcement de l'interopérabilité entre systèmes d'information hétérogènes. Elle souligne également le rôle crucial de standards ouverts et partagés, comme le RDF, dans la construction d'un écosystème de données cohérent et efficace, bénéfique à l'ensemble du secteur culturel.

3.1 L'ontologie

Le développement d'une ontologie commune a pour objectif de favoriser le partage, les réutilisations et la découvrabilité des contenus culturels auprès des publics. Préalables à l'étape d'exposition des données et d'élaboration de connecteurs pour la diffusion des données depuis les systèmes d'informations respectifs des maisons participantes, les travaux ontologiques permettent également de définir le périmètre de connaissances partagées entre les maisons d'opéra et plus largement avec le secteur culturel. Sa mise en place permet de simplifier les échanges, tant au sein des services des maisons d'opéra qu'auprès des collectivités publiques et des industries culturelles et créatives (ICC).

La réalisation d'un état de l'art sur les ontologies existantes et en capacité de répondre aux besoins des maisons d'opéra et des politiques culturelles a permis de détecter deux ontologies candidates.

La première est l'ontologie schema.org[6], qui présente de nombreux avantages. Tout d'abord, son approche englobante via l'adoption de définition large des concepts, apparaît comme très efficace pour répondre à de nombreuses situations, tout particulièrement pour le partage et la description des événements (schema.org/Event). Elle semble donc particulièrement adaptée pour la diffusion et la découvrabilité des dates de représentations. De plus sa documentation et son utilisation par les principaux moteurs de recherche pour l'indexation des contenus du web l'ont rendu très populaire auprès des équipes de communication et prestataires en charge des sites web.

5. Voir `https://www.w3.org/`

6. `https://schema.org`

Trois écueils nous ont conduits à poursuivre la phase de recherche ontologique. Premièrement, bien que cette ontologie décrive, de manière détaillée, un événement, une organisation ou même une œuvre, il apparaît que les étapes préliminaires et nécessaires à l'élaboration d'un spectacle ne sont que partiellement décrites. La notion de production, essentielle au sein du spectacle vivant, car englobant l'ensemble des actions menant à la représentation, telles que la conception des décors et des costumes ou la gestion des distributions, sont absentes de schema.org. Deuxièmement, le concept de producteur apparaît trop large pour une description fine et essentielle du rôle et de l'implication juridique de chaque partie prenante dans l'élaboration d'une production. Troisièmement, l'absence de traduction se révèle être un frein dans la capacité de représenter finement la vision développée par les politiques culturelles et de la représentation de la diversité.

La seconde ontologie que nous avons considérée est IFLA-LRM[15], qui est une référence en termes de gestion de connaissances dans le monde de la Culture de manière générale. Les concepts représentés sont bien plus proches de ce que nous souhaitons représenter pour le projet CapData Opéra. Si la notion d'œuvre y apparaît comme centrale, l'ontologie est néanmoins élaborée pour répondre aux objectifs de la gestion des ressources bibliographiques, ce qui ne nous a pas semblé parfaitement adapté aux besoins et à la description du spectacle vivant.

Les spectacles vivants, incluant le théâtre, la danse, la musique live, et d'autres formes d'art performance, nécessitent en effet des informations et des métadonnées détaillées et intrinsèques à la description d'une production artistique ou d'un spectacle, par exemple de ses décors, costumes, montages, aux effectifs et compositions des formations, voir également des publics.

Les recherches sur l'ontologie ont également mis en lumière l'existence d'initiatives et travaux de recherches similaires à l'international, dont ceux du groupe de "Performing Arts Information Representation Community Group", néanmoins celui-ci semble inactif depuis quelques années. Cet intérêt à l'échelle internationale et visible lors de la journée "Rendez-vous France-Québec sur la découvrabilité des contenus culturels francophones" [4] de l'édition 2023 du MTL Connecte, a permis d'entamer une réflexion sur les enjeux et la pertinence d'une action coordonnée, voire mutualisée, du chantier ontologique.

Nous avons fait le choix de développer l'ontologie CapData Opéra afin de proposer une description des connaissances qui répond aux besoins détectés auprès des maisons d'opéra et plus largement du secteur du spectacle vivant. Nous l'avons voulue complémentaire des autres modèles et avons mis en place directement dans l'ontologie des alignements vers l'IFLA-LRM et le schema.org. Nous permettons ainsi une représentation fidèle aux besoins du domaine, tout en rendant aisée l'utilisation de l'ontologie de référence dans le domaine culturel et de l'ontologie de référence pour l'indexation par les moteurs de recherche sur le Web.

Pour faciliter la réutilisation de cette ontologie par les maisons d'opéra, nous la documentons et la publions à l'URL `https://ontologie.capdataopera.fr`. Cette page permet de représenter les différentes versions de l'ontologie (actuellement la version 1.7) et la date de publication. Il y a, pour chaque version, une documentation générée par Widoco[5] et une représentation graphique générée avec WebVOWL[10]. De plus, les URI utilisées dans l'ontologie pointent sur ce site, ce qui favorise la négociation de contenu et la récupération des formats HTML ou RDF suivant la requête HTTP.

En parallèle de cette étape, les référentiels et vocabulaires contrôlés, utilisés au sein de l'ontologie, ont fait l'objet d'une exposition sur l'entrepôt SPARQL dédié au projet. Cette action participe à accroître l'interopérabilité et la découvrabilité via l'usage de définitions et de vocabulaires partagés au sein du réseau et plus largement du secteur culturel.

Enfin, des règles SHACL de vérification sont présentes sur le même site, avec une documentation adaptée [7], pour permettre aux maisons d'opéra de valider leurs données.

L'un des avantages du processus de modélisation et des outils mis en place autour de la publication de l'ontologie est sa grande agilité. Nous avons pu, en effet, confronter la modélisation de manière concrète aux besoins des équipes des maisons participantes, prestataires, services externes et utilisateurs. Une modification de l'ontologie était rapidement intégrée dans un outil d'export de données en RDF grâce à la documentation disponible et aux outils de vérification des données. La réalisation manuelle des alignements a permis d'améliorer de manière itérative le modèle. Cette approche nous a permis de nous rendre compte très rapidement des écueils et de pouvoir corriger l'ontologie de manière agile afin d'obtenir une version stabilisée et intégrable au sein des connecteurs et des applications développées en parallèle.

3.2 Le suivi de production

Nous considérons, comme présenté en introduction de cette section, que la maison d'opéra gère la publication de ses propres données RDF. Cette publication prend la forme d'un fichier RDF disponible à une URL donnée contenant l'intégralité des données. Ce fichier est mis à jour régulièrement par un export régulier de la part des maisons d'opéra. Pour permettre l'interrogation de toutes ces données, nous souhaitons les récupérer pour les publier dans un entrepôt SPARQL dédié. Pour cela, un script s'exécute tous les jours pour récupérer ces fichiers mis à disposition derrière les URL de chaque maison. Un ensemble de traitements de nettoyage de données sont appliqués, comme l'effacement des espaces avant et après les valeurs littérales, la détection de l'utilisation d'une valeur à la place d'un identifiant de référentiels (par exemple le code pays), etc. Ces traitements sont appliqués systématiquement à chaque récolte des données. De plus, certaines données sont alignées sur les référentiels fournis par la ROF afin de faciliter l'interopérabilité. Enfin, les données sont validées en utilisant les règles SHACL liées à l'ontologie. De cette manière nous obtenons quatre graphes différents pour chaque maison d'opéra lors

7. générée grâce à `https://shacl-play.sparna.fr/play/doc`

d'une récolte de données :

— un graphe contenant les données initiales ;
— un graphe contenant les données nettoyées ;
— un graphe contenant les alignements avec les données de la ROF ;
— un graphe contenant les triplets du rapport de validation SHACL [8].

Ces graphes sont ensuite envoyés dans un entrepôt SPARQL mutualisé pour toutes les maisons d'opéra et administré par la ROF (`https://sparql.capdataopera.fr/`).

De cette manière, toutes les données sont mises au même endroit et différents graphes nommés permettent de récupérer les données qui nous intéressent. De plus, il devient trivial de faire des requêtes entre plusieurs maisons d'opéra à partir du moment où toutes les données sont sur le même entrepôt. Le choix de cette architecture répond également à une analyse approfondie des coûts de maintenance à moyen et long terme. L'étude de faisabilité du projet a en effet mis en lumière le coût non soutenable que représentait la mise en place d'API au sein de chaque maison d'opéra pour la gestion de la diffusion et la récupération des données. Une approche non mutualisée engendrait une démultiplication des coûts financiers pour les établissements et les collectivités publiques.

Afin de mettre en place une boucle rétroactive bénéfique pour les maisons d'opéra, nous avons écrit des requêtes SPARQL pour qu'elles récupèrent leurs données, nettoyées, alignées et validées. Il leur est donc possible de réinsérer ces données dans les systèmes d'information d'origine pour augmentant la qualité de leurs données. Par exemple en ajoutant pour une personne les identifiants ROF, ISNI ou ARK issus de l'alignement.

4 Les outils

Au cours du projet, nous avons poursuivi une démarche visant à industrialiser toute la chaîne de production et nous avons constaté des manques parmi les outils disponibles sous une licence de logiciel libre. Nous avons essayé de cartographier cette situation à travers l'approche SemGraph [9] qui, pour chaque étape de la chaîne, suggère des outils possibles. Nous avons développé ou fait évoluer certains outils quand nous l'avons jugé utile et que cela était dans nos moyens.

4.1 Publication de l'ontologie

Lorsque nous avons souhaité publier l'ontologie pour la rendre disponible sur le Web, nous avons identifié des portails comme OntoPortal[8] pour héberger nos modèles et avons envisagé de simplement les publier derrière un serveur Web standard. Néanmoins, nous avons souhaité pouvoir gérer plus finement les versions, avoir une documentation et de la négociation de contenu qui permettent d'accéder à la documentation directement, tout cela intégré à nos

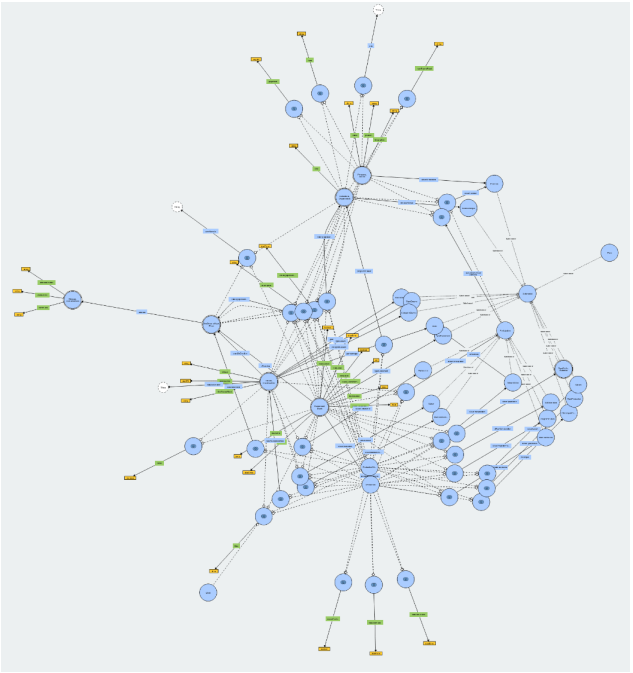

FIGURE 1 – L'ontologie CapData Opéra représentée avec WebVOWL

outils de déploiement continu habituels. Nous avons développé un script qui permet, lorsque l'on met à jour un entrepôt de gestion de version de code qui contient une ontologie, de générer la documentation, les règles SHACL, la documentation des règles SHACL et qui publie tout cela derrière un serveur Web automatiquement. Le déploiement continu permet de mettre à jour le site `https://ontologie.capdataopera.fr/`.

Le script utilise différents outils pour permettre la génération de tous ces éléments. Tout d'abord, nous utilisons Widoco[5] pour la génération de la documentation à partir des métadonnées de l'ontologie et de sa modélisation. Le rendu graphique est très lisible et proche de ce qui existe déjà dans d'autres projets, ce qui en fait une documentation simple à prendre en main. Cet outil utilise WebVOWL[10] pour avoir une représentation graphique de l'ontologie. Comme on peut le voir sur la figure 1, cette représentation permet d'avoir un aperçu global de ce qui est présent dans l'ontologie, et donc elle est particulièrement utile pour la découverte de l'ontologie, mais il est difficile d'en détecter les détails.

Nous avons utilisé la génération de documentation SHACL en utilisant l'outil proposé dans SHACL-play [10]. La documentation générée (comme nous pouvons l'observer sur la figure 2) permet de se rendre compte très facilement de ce qui est attendu et avoir un rapport valide lors de la validation SHACL des données.

Tous ces outils sont utilisés dans notre processus d'intégration continue proposé dans GitLab [11] et le résultat est dé-

8. `https://www.w3.org/TR/shacl/#validation-report`
9. Voir `https://semgraph.logilab.fr`
10. `https://shacl-play.sparna.fr/play/doc`
11. `https://docs.gitlab.com/ee/ci/`

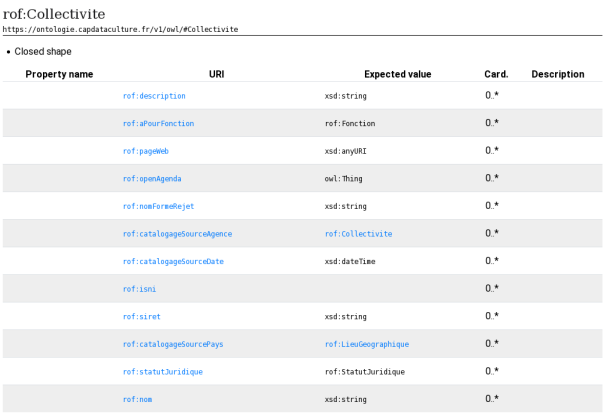

FIGURE 2 – Documentation SHACL

ployé en utilisant les GitLab Pages [12]. Chaque fois qu'une modification dans l'ontologie est effectuée, tout le processus est automatiquement relancé grâce à l'intégration continue et le résultat est accessible grâce au serveur Web proposé dans les GitLab Pages.

Ce processus d'intégration et de déploiement continu pour la publication d'ontologie est un réel atout qui peut être utilisé dans d'autres projets dès lors qu'une ontologie doit être maintenue.

Un besoin qui a été prégnant tout le long du projet a été de pouvoir vérifier ce qui a été exporté dans l'entrepôt SPARQL. Nous avons commencé à explorer les données exportées par l'intermédiaire d'un certain nombre de requêtes SPARQL pour voir le résultat. Cette solution a vite montré ses limites, car il n'a pas été simple d'écrire les requêtes SPARQL permettant de tout voir facilement et rapidement. Nous avons alors utilisé l'outil SparqlExplorer [13] qui permet de parcourir l'ensemble des données d'un entrepôt SPARQL pour découvrir les données qui y sont présentes.

La figure 3 présente la page d'accueil du SparqlExplorer une fois que l'on a spécifié l'entrepôt SPARQL à explorer. Nous pouvons observer la liste des classes et le nombre d'instances associés à chacune des classes, et un champ de recherche, qui permet de chercher parmi les littéraux.

Lorsque nous cliquons sur une URI, nous affichons la vue présentée sur la figure 4. Cette vue permet de lister l'ensemble des triplets concernant cette URI et de pouvoir filtrer ces triplets (ici un filtre a été appliqué avec la valeur "nom"). De cette manière, il est possible de parcourir les triplets pour observer ce qui a vraiment été exporté et donc de s'assurer que le résultat correspond bien à ce qui est attendu.

Nous avons intégré l'outil YASGUI[13] pour interroger l'entrepôt SPARQL grâce à une interface plus pratique à utiliser que l'interface proposée par Virtuoso. Cette interface, visible sur la figure 5, comporte une option pour partager un lien vers une requête SPARQL. Ce lien a beau-

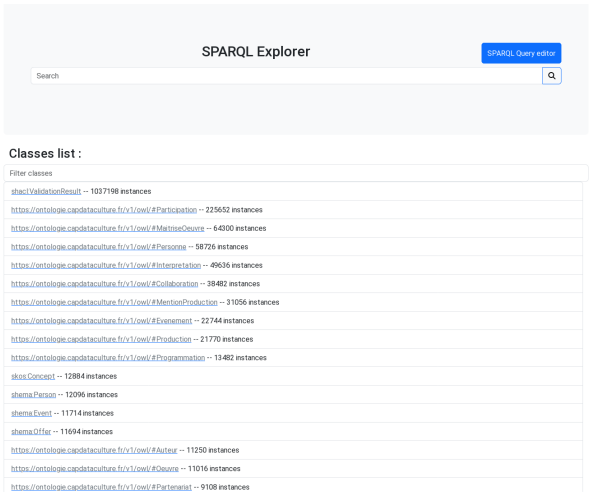

FIGURE 3 – Page d'accueil du SparqlExplorer

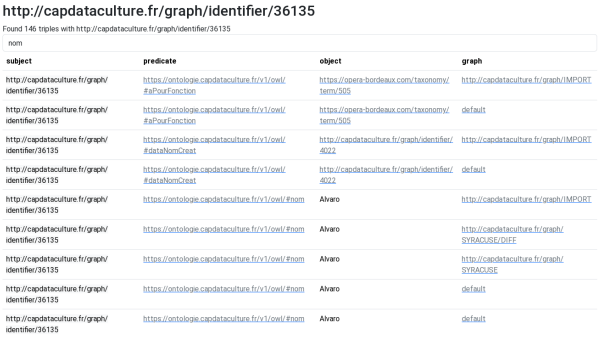

FIGURE 4 – Liste de triplets dans le SparqlExplorer

12. https://docs.gitlab.com/ee/user/project/pages/
13. https://sparqlexplorer.app/

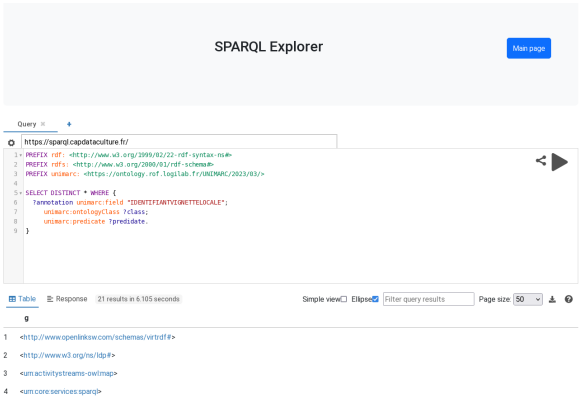

FIGURE 5 – YASGUI dans le SparqlExplorer

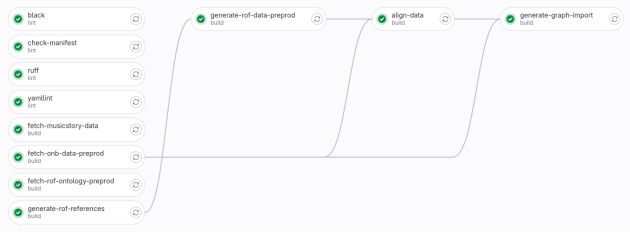

FIGURE 6 – Exécution des différentes étapes de récupération des données

coup été utilisé durant le projet et a grandement simplifié les échanges et les collaborations pour communiquer entre partenaires la présence au l'absence de certaines informations dans le graphe.

4.2 Assemblage et publication du graphe

Une fois que chaque maison d'opéra a publié ses données à une adresse URL de son choix, il faut les récupérer pour les agréger dans le même entrepôt SPARQL. Cette récupération s'effectue quotidiennement pour s'assurer d'avoir des données à jour dans l'entrepôt. Pour cela, nous avons mis en place une tâche récurrente avec le mécanisme d'intégration continue de notre forge logicielle [14]. Nous pouvons ensuite suivre l'exécution de chaque tâche et regarder le résultat obtenu à chaque étape.

La figure 6 permet de voir les différentes étapes d'une mise à jour de l'intégralité des données. Sur cet exemple, seules trois sources de données sont présentes : Les données de l'Opéra National de Bordeaux, les données de Music Story [15] et les données de la Réunion des Opéras de France (ROF). Cette dernière source requiert plusieurs étapes car nous interrogeons directement une API pour récupérer les données que nous transformons ensuite en RDF. Nous sommes en train d'étudier le transfert de cette responsabilité vers l'équipe qui gère les données de la ROF.

La solution trouvée ici permet de mettre en lumière l'importance de l'adoption de la solution. Dès qu'un fournisseur de données n'adopte pas les technologies préconisées

14. GitLab `https://docs.gitlab.com/ee/ci/`

15. Pour la valorisation notamment sur les plateformes de streaming `https://music-story.com/fr/`

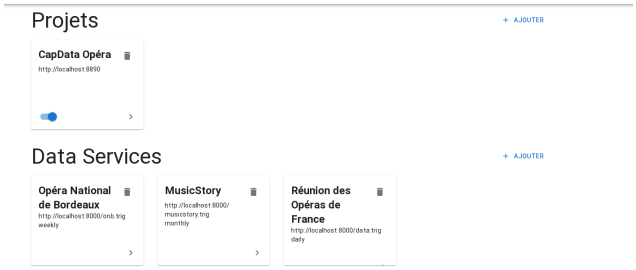

FIGURE 7 – Exemple de l'application de suivi de production

par le projet, cela demande un traitement dédié et spécifique pour cette source. Par exemple, ici, nous avions accès à une API, que nous avons utilisée pour générer le graphe RDF contenant l'intégralité des données. Ce traitement n'a pas été simple à mettre en place à cause de la complexité de l'API [16]. Ce travail a nécessité beaucoup d'échanges avec l'équipe en charge de la gestion des données. Il est apparu que les développements sont beaucoup plus fluides lorsque chaque gestionnaire de données gère la transformation en RDF de ses propres données.

Afin de faciliter les développements et de permettre aux fournisseurs de données d'être les plus autonomes possible, nous avons initié le développement d'une application de suivi de production qui permettra de suivre l'état de chaque récolte spécifiquement et d'avoir accès aux journaux d'import pour savoir comment il s'est déroulé. L'application enverra les données dans l'entrepôt SPARQL si la récolte s'est correctement déroulée. De cette manière, les maisons d'opéra seront autonomes dans la publication de leurs données. Elles pourront ajouter elle-même la source de données dans l'application de suivi de production et corriger les erreurs qui seront remontées dans les journaux suite à la vérification de la conformité des données en utilisant les règles SHACL. Cette application donnera une vision claire de ce qui a été importé dans l'entrepôt SPARQL. Elle constitue une étape importante dans la phase d'industrialisation du projet CapData Opéra.

Comme nous pouvons le voir dans la figure7, nous pouvons définir un projet, ici "CapData Opéra" et différentes sources ("Opéra National de Bordeaux", "Music Story" et "Réunion des Opéras de France").

La figure 8 montre un exemple de l'ajout d'une recette dans l'application de suivi de production. Cette recette permet d'identifier une source de données à importer pour le projet, l'URI du graphe dans lequel nous souhaitons envoyer les données et le processus à appliquer sur les données. Ce processus pourra être modifié dans le code pour permettre des traitements particuliers, comme transformer du CSV, ou utiliser une API, etc. Des erreurs d'import pour une source ne bloqueront pas l'import des autres sources, ce qui permet une plus grande flexibilité. Il sera alors possible d'intégrer

16. Basée sur une modélisation UNIMARC[2]

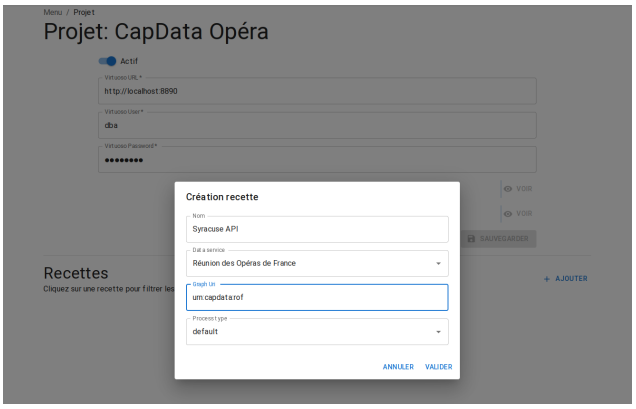

FIGURE 8 – Ajout d'une recette dans l'application de suivi de production

un plus grand nombre de maisons d'opéras plus facilement. Le graphe final est publié dans un entrepôt RDF interrogeable en SPARQL grâce au logiciel Virtuoso [17] et parcouru avec un navigateur web en utilisant SparqlExplorer, comme présenté dans le chapitre précédent.

5 Responsabilités et périmètres

Nous avons présenté le processus de publication des données des différentes maisons d'opéra pour permettre l'interopérabilité entre ces données. L'architecture et les outils présentés facilitent l'autonomisation du processus, tout en assurant la pleine souveraineté des maisons d'opéra et structures participantes sur leurs données. La mise en place d'une solution mutualisée et la coordination des développements des outils développés dans le cadre du projet permet de répartir les responsabilités.

Tout d'abord dans le processus de publication des données, la mission numérique de la ROF en lien avec les besoins détectés auprès des groupes de travail dont celui des référents numériques et des échanges et des réflexions avec les partenaires, assure la maintenance de l'ontologie. Cette maintenance nécessite de considérer les besoins en modélisation des différentes maisons d'opéra, de les transcrire dans l'ontologie et de faire un suivi de versions de l'ontologie pour communiquer autour des changements. C'est pour cela que nous avons adopté pour la gestion rapide de l'ontologie un déploiement continu et documenté. De plus, une gestion de version de l'ontologie est intégrée directement dans l'URI de celle-ci, ce qui permet de pouvoir communiquer chaque fois qu'une modification importante dans l'ontologie a lieu en annonçant la publication d'une nouvelle version.

Un certain nombre de référentiels sont publiés par la ROF directement dans l'entrepôt SPARQL du projet en utilisant le vocabulaire SKOS (comme la liste des fonctions lors d'une participation d'une personne à une production). Il est important de pouvoir proposer des référentiels uniques pour tout le projet afin d'assurer une interopérabilité importante entre les données des différentes maisons d'opéra.

17. https://virtuoso.openlinksw.com/

Enfin, il est également nécessaire de proposer une infrastructure permettant d'héberger les outils proposés, dont principalement l'ontologie, l'application de suivi de production et l'entrepôt SPARQL. La coordination menée par le réseau ROF permet d'assurer une continuité et la stabilité des services tout en adoptant une approche mutualisée.

Au-delà de l'aspect technique, l'accompagnement et la constitution d'un socle commun de connaissances auprès des maisons participantes et des prestataires se sont révélés être une clé majeure dans la conduite et la réussite du projet. Les simples publications de l'ontologie et de l'entrepôt SPARQL seraient complètement insuffisantes pour assurer leur intégration et déploiement.

La constitution d'une documentation adaptée de l'ontologie et la communication s'avère stratégique pour le déploiement de la solution. Plusieurs présentations publiques ont eu lieu au cours de l'année 2023. Dans ce sens, nous avons souhaité développer une application de suivi de production la plus simple possible. Nous organisons aussi des réunions, ateliers et sessions de travail pour faciliter au maximum l'appropriation des concepts et des technologies que nous avons mis en œuvre ici. Les technologies du Web Sémantique restent méconnues par de nombreux prestataires du domaine culturel, la mise en place de salons de messagerie instantanée favorisant les échanges et questions, se révèlent être des leviers particulièrement puissants. Près de 5 salons dédiés à la gestion et coordination des chantiers ont été mis en place, générant plus de 10 000 interactions sur l'année 2023.

Tout en favorisant la mutualisation, le choix d'une architecture permet une grande agilité, assure le respect de la souveraineté et la pleine responsabilité des maisons, partenaires et services externes participant au projet. Ils conservent une complète autonomie dans le choix d'ouverture, de diffusion, de réutilisation et de l'enrichissement de leurs données. L'Opéra National de Bordeaux a ainsi la possibilité d'exposer ses données de programmation tout en enrichissant sa base avec des données identifiées et exposées au sein de l'entrepôt par un ou plusieurs membres des structures participantes.

6 Conclusion

Élaboré sur la base des standards du web sémantique, le projet CapData Opéra a déployé une solution mutualisée fondée sur une architecture et des outils hautement réplicables. L'expérimentation et les premiers résultats liés confirment la pertinence de l'approche choisie en vue de simplifier l'échange, la gestion et la découvrabilité des données des maisons d'opéra et des autres structures participantes.

En complète adéquation avec les politiques culturelles, l'interopérabilité des données répond à de multiples besoins détectés auprès des maisons d'opéra et enjeux du spectacle vivant, dont le développement de la découvrabilité des œuvres, des artistes et plus largement des arts lyriques et chorégraphiques auprès des publics. D'autres acteurs culturels nous semblent en effet être sur la même ligne que la

nôtre. Nous pouvons citer le monde des marionnettistes avec qui nous sommes en contact, le monde du théâtre qui porte des initiatives comme la publication des données des Registres de la Comédie Française[6], le ministère de la Culture qui conduit des ateliers dans le cadre de la deuxième génération de la feuille de route "Politique données et contenus culturels"[14] ou encore les réflexions du groupe de travail "Ouverture des données"[3] animé par le réseau du TMNlab.

La modélisation d'une ontologie et son adoption au sein de systèmes d'information hétérogènes sont des actions complexes. L'expérimentation souligne les rôles essentiels de la coordination et de l'accompagnement des établissements partenaires et prestataires. Si la question de l'héritage des logiciels et des processus humains existants est aujourd'hui bien connue, elle nécessite une attention toute particulière pour l'intégration de nouveaux modèles.

Le projet a été l'occasion d'éprouver un certain nombre d'outils du Web Sémantique et d'identifier les fonctionnalités manquantes ou les besoins pour lesquels les outils restent à concevoir. Nous avons mis en place la génération de la documentation de l'ontologie et son déploiement continu par l'intermédiaire d'un entrepôt de code, une application de suivi de production et un outil de navigation dans le graphe final. Nous prévoyons d'améliorer ces outils, mais surtout de les rendre plus génériques pour qu'ils puissent être utilisés dans d'autres projets.

L'expérimentation menée dans un premier temps avec l'Opéra National de Bordeaux est en cours d'industrialisation et de déploiement auprès de six maisons d'opéra dont l'Opéra National de Bordeaux, Théâtre du Châtelet, l'Opéra de Rennes, l'Opéra Comique, l'Opéra national Capitole Toulouse et l'Opéra de Limoges. Cette approche permet d'affiner progressivement les différents chantiers et outils : ontologie, connecteurs, applications, documentations et services dédiés à la valorisation.

Nous avons commencé à étudier la possibilité de valoriser ces données aux travers de services de valorisation dédiés, par exemple via un prestataire permettant de faire le lien avec les plateformes de streaming, ou encore un système de gestion d'agenda partagé pour la publication automatique des événements.

L'expérimentation, le constat d'un besoin présent partagé par un grand nombre d'établissements du spectacle vivant et plus largement du secteur culturel et l'émergence d'initiatives similaires à l'internationale, soulignent le besoin et la pertinence d'une démarche coordonnée des recherches et actions.

L'adoption d'une approche coordonnée met en lumière le chantier essentiel de la gouvernance. Les modèles de fonctionnement internationaux de l'IFLA ou du schema.org sont ainsi précieux d'enseignement. Les opportunités offertes par une telle approche sont nombreuses, tant pour la mutualisation des coûts financiers, le développement d'outils partagés en capacité de simplifier et d'assurer de manière pérenne leurs adoptions et déploiement, tout en répondant aux besoins et enjeux transversaux du secteur culturel.

Références

[1] Adrien Basdevant, Camille François, and Rémi Ronfard. *Rapport de la mission sur le développement des métavers*. PhD thesis, Ministère de la Culture (France), 2022.

[2] Permanent UNIMARC Committee et al. Unimarc authorities format manual. 2023.

[3] Groupe de travail TMNlab. Living lab - Ouverture des données, 2023.

[4] Mission franco-québécoise sur la découvrabilité en ligne des contenus culturels francophones. Table ronde "Normaliser la diversité des données culturelles : est-ce possible ? Rendez-vous France-Québec", Montréal Connecte. `https://www.youtube.com/watch?v=3HbgAUUNUiw`, 2023.

[5] Daniel Garijo. Widoco : a wizard for documenting ontologies. In *The Semantic Web–ISWC 2017 : 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, pages 94–102. Springer, 2017.

[6] Charline Granger and Fabien Amarger. Les registres de la comédie-française sur le web de données liées : de l'hétérogénéité de données vers des données quantitatives en rdf. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle APIA@ PFIA2023*, number 2023, pages 63–71. AFIA-Association Française pour l'Intelligence Artificielle, 2023.

[7] Direction générale des médias et des industries culturelles (DGMIC). Découvrabilité en ligne des contenus culturels francophones, 2022.

[8] Clement Jonquet, John Graybeal, Syphax Bouazzouni, Michael Dorf, Nicola Fiore, Xeni Kechagioglou, Timothy Redmond, Ilaria Rosati, Alex Skrenchuk, Jennifer L Vendetti, et al. Ontology repositories and semantic artefact catalogues with the ontoportal technology. In *International Semantic Web Conference*, pages 38–58. Springer, 2023.

[9] Julie Knibbe. Les données dans la musique : Enjeux et stratégies d'investissement. 2023.

[10] Steffen Lohmann, Vincent Link, Eduard Marbach, and Stefan Negru. Webvowl : Web-based visualization of ontologies. In *Knowledge Engineering and Knowledge Management : EKAW 2014 Satellite Events, VISUAL, EKM1, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers. 19*, pages 154–158. Springer, 2015.

[11] Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107) :6, 2004.

[12] Eudes-Emmanuel Peyre and Groupe de travail numérique ROF. Capdata Opéra - France 2030. `https://www.rof.fr/rof/capdata-opera.aspx`, 2022.

[13] Laurens Rietveld and Rinke Hoekstra. The yasgui family of sparql clients 1. *Semantic Web*, 8(3) :373–383, 2017.

[14] Ministère de la Culture (SNUM) Service du numérique. Ateliers préliminaires à la deuxième génération de la feuille de route "Politique des données et contenus culturels", 2024.

[15] Maja Žumer. Ifla library reference model (ifla lrm)—harmonisation of the frbr family. *KO Knowledge Organization*, 45(4) :310–318, 2018.