Redécouvrir les théâtres de la Foire et de la Comédie-Italienne avec les bases THEAVILLE et RECITAL

Françoise Rubellin, Guillaume Raschia

Redécouvrir les Théâtres de la Foire et la Comédie-Italienne avec les bases THEAVILLE et RECITAL

Françoise Rubellin et Guillaume Raschia

Théâtres sans privilège et privilège du numérique

L'histoire du théâtre français du XVIII[e] siècle a longtemps été focalisée sur la Comédie-Française, laissant de côté la Comédie-Italienne (à part les pièces de Marivaux), les théâtres de la Foire (on désigne ainsi les spectacles données à Paris aux foires Saint-Germain, Saint-Laurent et Saint-Ovide), les théâtres de société et d'éducation etc. Fort heureusement la fin du XX[e] siècle a connu de grandes avancées dans le domaine, pour lesquelles l'avènement du numérique a joué un rôle fondamental. Barry Russell créa le premier site dédié aux théâtres forains (foires.net), puis lança le site CESAR (Calendrier Electronique des Spectacles d'Ancien Régime) avec David Trott et Jeffrey Ravel ; David Trott ébaucha un site sur les théâtres de société (avec Marie-Emmanuelle Plagnol-Diéval et Dominique Quéro). Ces entreprises révélèrent une nouvelle manière d'aborder le théâtre particulièrement féconde, d'autant qu'elles nécessitaient un effort collaboratif inédit ; malgré le décès prématuré de Barry Russell en 2003 et de David Trott en 2005, elles ont opéré une révolution dans l'historiographie des spectacles d'Ancien Régime.

La méconnaissance des théâtres non privilégiés, c'est-à-dire qui n'avaient pas été dotés de monopole par Louis XIV (contrairement à la Comédie-Française et à l'Opéra), était due en partie à l'absence de données : qu'il s'agisse de chronologies, de pièces, de documents d'archives, il fallait établir des calendriers, localiser et éditer des manuscrits, transcrire un grand nombre de documents de tous ordres… Les deux bases de données que nous présentons ici, l'une succinctement (THEAVILLE), l'autre en détail (RECITAL), permettent une meilleure connaissance de ces théâtres, sous l'angle de la musique pour l'une, sous l'angle de l'organisation financière et logistique pour l'autre.

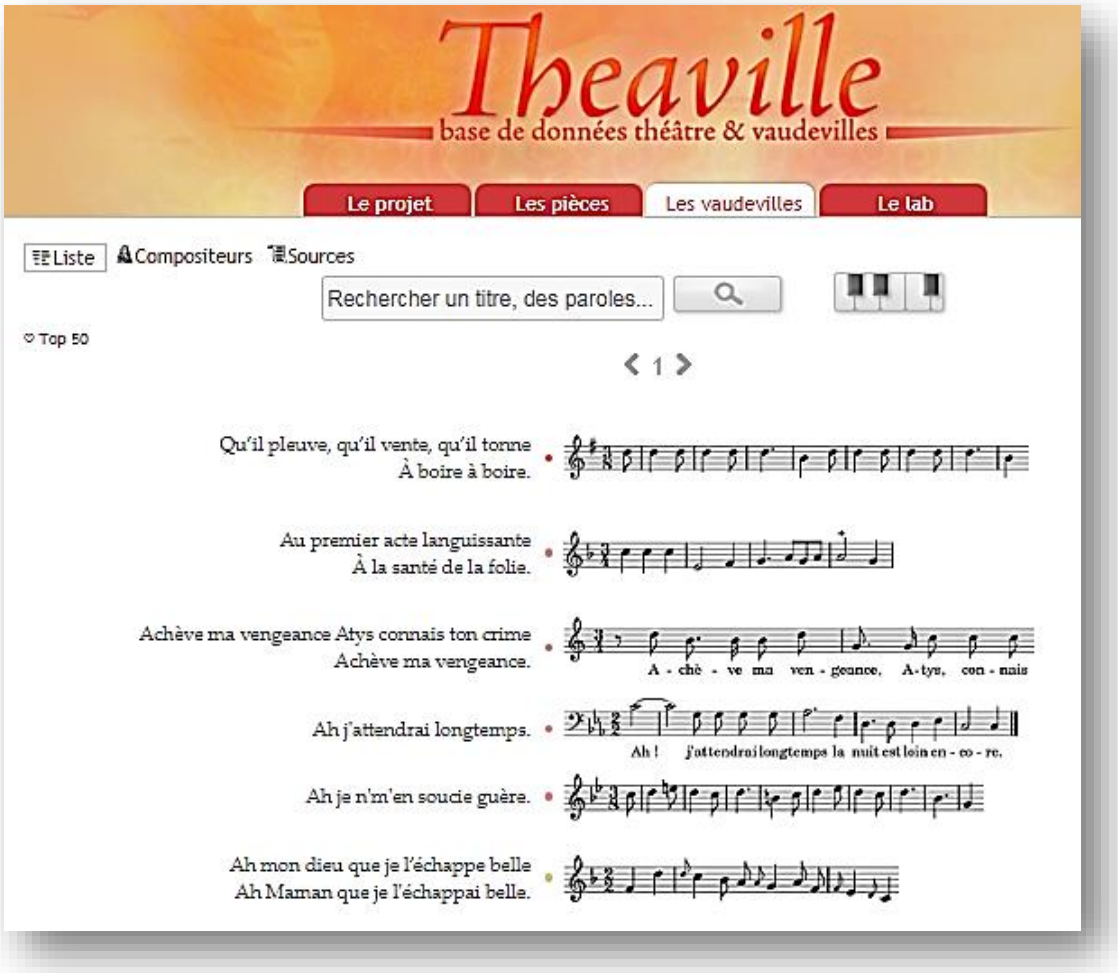
THEAVILLE : une base pour réconcilier texte et musique

Le CETHEFI (Centre d'études des théâtres de la Foire et de la Comédie-Italienne, basé à l'Université de Nantes, dans le laboratoire L'AMo) obtint de l'ANR le financement d'un projet centré sur les parodies d'opéra (POIESIS, 2008-2012) qui permit la création d'une base de données, THEAVILLE, contenant près de 250 textes de parodies dramatiques d'opéra, et plus de 2000 fichiers sons de vaudevilles (airs populaires – pour la plupart - employés dans ces comédies). Cette base de données est en perpétuelle augmentation (theaville.org) et a fait l'objet d'autres présentations[1] ; rappelons-en seulement le principe.

L'exemple en lien ici permet de voir qu'on peut connaître les différents titres d'un même vaudeville, qu'on peut l'entendre en fichier son, extraire le fichier LilyPond, le fichier PDF, être renvoyé à la partition sur Gallica ou GoogleBooks, voir toutes les variantes de l'air, sa fréquence d'utilisation dans les pièces (c'est-à-dire dans les 250 parodies d'opéra figurant

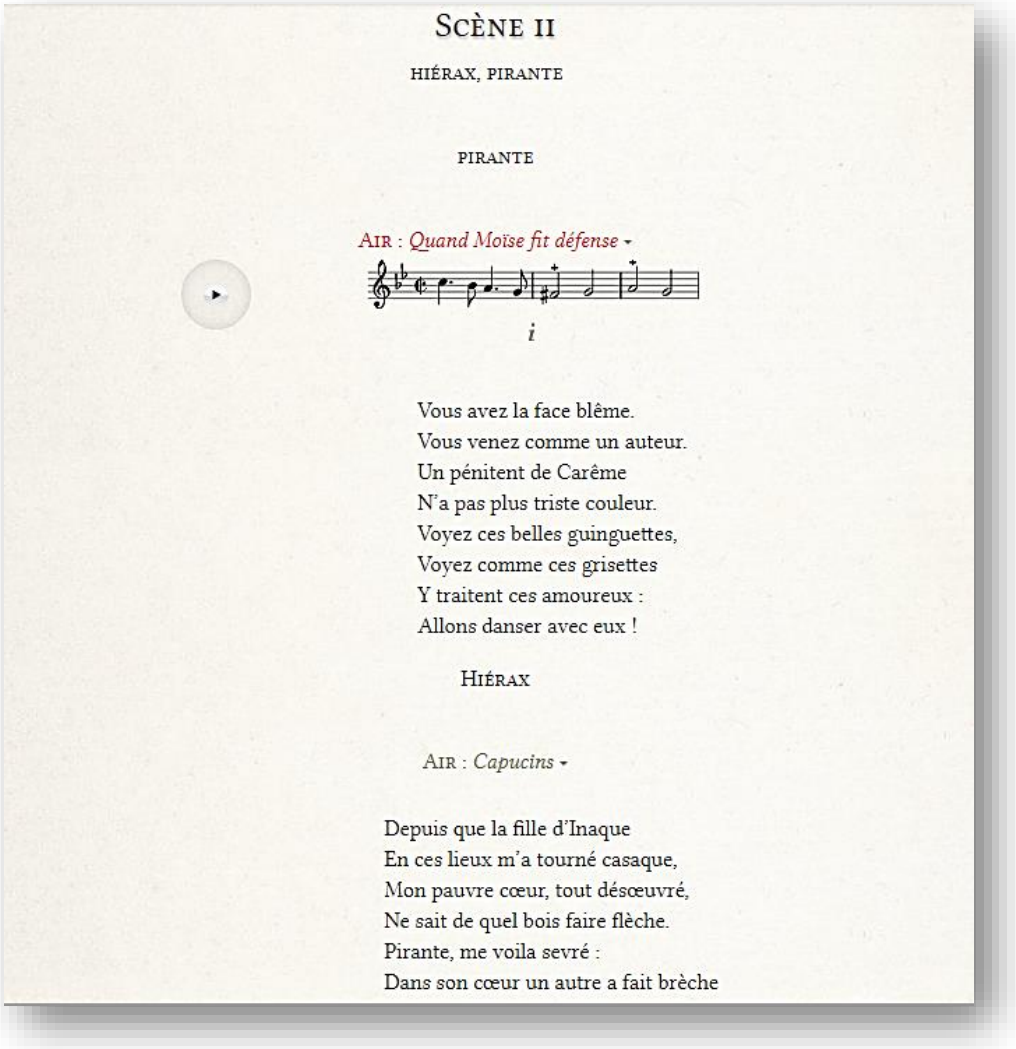
[1] Voir notamment « Entretien avec Françoise Rubellin » dans *Espaces des théâtres de société. Définitions, enjeux, postérité*, Valentina Ponzetto, Jennifer Ruimi éd., Presses Universitaires de Rennes, 2020.

dans la base), et les premières paroles connues. On doit à Florent Coubard le développement et l'amélioration de la base depuis plusieurs années. Quelques exemples de ses inventions : sous l'onglet vaudeville il a fait figurer un petit clavier de piano (à côté du titre à chercher) qui permet de retrouver un air dont on ignore les paroles, quelle que soit sa tonalité ; il a introduit un top 50 des airs les plus employés, qui permet aussi de faire varier les périodes interrogées ; il a mis en place la possibilité de marquer des vaudevilles avec des signets, afin de constituer une sorte de classeur personnel de ceux qu'on a besoin de consulter ou d'utiliser, etc.



Capture d'écran du site Theaville, onglet vaudevilles.

À condition de s'inscrire (en haut à droite de l'écran, inscription gratuite), on peut avoir accès aux pièces éditées dans le site. Ce sont toutes les parodies d'opéra dont il reste le texte aujourd'hui. Lorsqu'on passe la souris sur le titre de l'air, s'il devient rouge on peut cliquer et voir apparaître une petite partition, que la flèche de gauche permet d'entendre. Le « i » sous la portée donne accès à toutes les informations concernant le vaudeville (ses différents titres, ses variantes mélodiques, ses sources, sa fréquence d'utilisation dans les parodies d'opéra etc.)



SCÈNE II

HIÉRAX, PIRANTE

PIRANTE

AIR : *Quand Moïse fit défense*

i

Vous avez la face blême.
Vous venez comme un auteur.
Un pénitent de Carême
N'a pas plus triste couleur.
Voyez ces belles guinguettes,
Voyez comme ces grisettes
Y traitent ces amoureux :
Allons danser avec eux !

HIÉRAX

AIR : *Capucins*

Depuis que la fille d'Inaque
En ces lieux m'a tourné casaque,
Mon pauvre cœur, tout désœuvré,
Ne sait de quel bois faire flèche.
Pirante, me voila sevré :
Dans son cœur un autre a fait brèche

Exemple tiré du Trompeur trompé *de Fuzelier (1733), parodie d'*Isis *de Quinault et Lully.*

RECITAL : une base pour découvrir le théâtre par les registres comptables

Un second projet financé par l'ANR à partir de 2014 a permis une autre approche de ces théâtres. Ce projet, dirigé par Françoise Rubellin et intitulé CIRESFI (Contrainte et Intégration : pour une Réévaluation des Spectacles Forains et Italiens), comportait au départ trois axes d'étude :

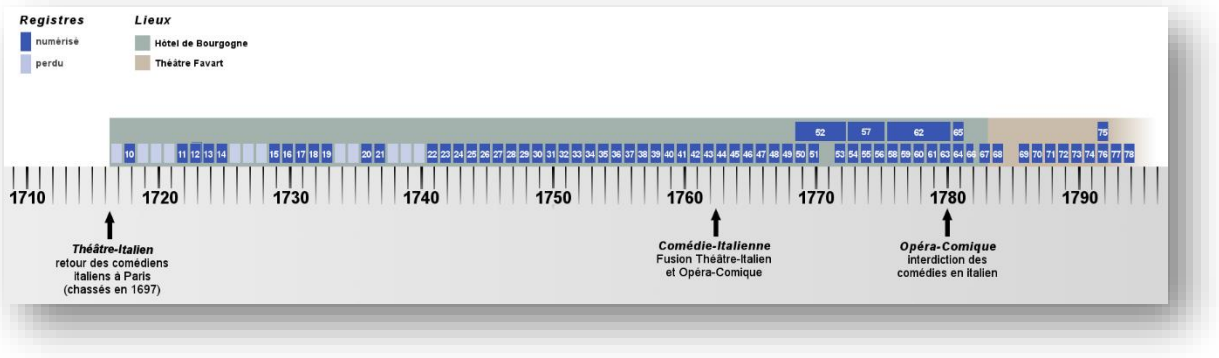
- l'intégration et l'institutionnalisation progressive des théâtres forains (l'opéra-comique) et des Italiens,
- l'émergence de formes esthétiques innovantes induites par la contrainte et la concurrence avec les théâtres détenteurs d'un privilège,
- l'économie du spectacle, en particulier son organisation, sa production, la composition sociale du public, les mécanismes de rémunération des auteurs, acteurs et autres personnels de spectacle, etc.

Pour la réalisation du troisième objectif, en ce qui concerne la Comédie-Italienne, il a donné lieu à la constitution d'une base de données à partir d'une plateforme de crowdsourcing imaginée pendant le programme, nommée RECITAL (**Re**gistres de la **C**omédie **Ital**ienne). Cette

base exploite la source primaire que constitue le fonds des registres de la Comédie-Italienne de 1717 à 1793, conservé à la Bibliothèque-Musée de l'Opéra (rattachée à la Bibliothèque nationale de France, sous une cote étonnante : TH-OC (pour Théâtre de l'Opéra-Comique). Or ce n'est qu'à partir de 1762 que la Comédie-Italienne absorbe une partie de la troupe de l'Opéra-Comique de la Foire. Nous avons choisi de travailler de 1717 (premier volume conservé) jusqu'à 1793. Cet ensemble a fait l'objet d'une première numérisation financée par un programme de recherche antérieur. Désormais les fac-similés des registres de la Comédie-Italienne sont intégralement disponibles sur Gallica. Une convention a été signée entre l'université de Nantes et la BnF pour céder au projet CIRESFI un droit d'exploitation des fac-similés, en échange de liens hypertextes sur la plateforme RECITAL vers les sources Gallica, et d'une livraison (non exclusive) à la BnF des données collectées.

Les registres sous grille à la Bibliothèque-Musée de l'Opéra (d.r.)

Dans ces 64 registres comptables (1 par saison théâtrale ; certains registres, en bleu clair dans la frise ci-dessous, sont perdus), représentant un volume de 25 250 pages, sont consignées chaque jour les représentations avec leurs recettes et leurs dépenses ; figurent aussi des récapitulations mensuelles et annuelles, de manière très variable.



Frise représentant la série des registres (cote Bibliothèque-Musée de l'Opéra, TH-OC 10 à 78) : en dessus de la ligne principale les cinq registres de vérification (©FlorentCoubard)

S'agissant pour une grande part de données numéraires (montants de recettes et de dépenses), l'interprétation en valeur absolue n'est pas signifiante, tandis que leur agrégation, leur comparaison et leur étude longitudinale à l'échelle d'un siècle (le XVIII^{e}) contribuent à améliorer notre compréhension de l'économie du spectacle. Dans cette perspective, et au-delà de nos propres expériences, il est intéressant de rendre accessibles ces données pour permettre à d'autres d'apporter un regard nouveau sur ces données (nouvelle méthode d'analyse, points de vue différents, etc.) et d'encourager le croisement des bases de données pour la mise en relation de tous les faits relatifs à l'histoire des spectacles : tout ceci ayant pour but de concourir à la production de connaissance.

Une interdisciplinarité équilibrée

L'équipe CIRESFI/RECITAL comprend des chercheurs en littérature et histoire du théâtre, des historiens, des musicologues, des chercheurs en informatique (science des données, représentation de connaissance, reconnaissance de formes), ainsi que des ingénieurs en informatique. Les questionnements scientifiques sont élaborés par les experts en histoire du théâtre, puis traduits en modèle de données par les chercheurs en informatique une fois que la source (les copies numérisées des registres de la Comédie-Italienne) est connue. Ce modèle de données est ensuite implémenté par les ingénieurs, et validé/corrigé par les chercheurs en sciences humaines. Ce processus est répété dans une boucle vertueuse puisque la validation du modèle suscite généralement de nouvelles questions. Dans cette boucle de rétroaction, nous avons observé à de nombreuses reprises l'effet normatif de la modélisation sur les questionnements scientifiques : ici la nécessité d'une catégorisation rigoureuse, là le choix impérieux d'une interprétation plutôt qu'une autre, etc.

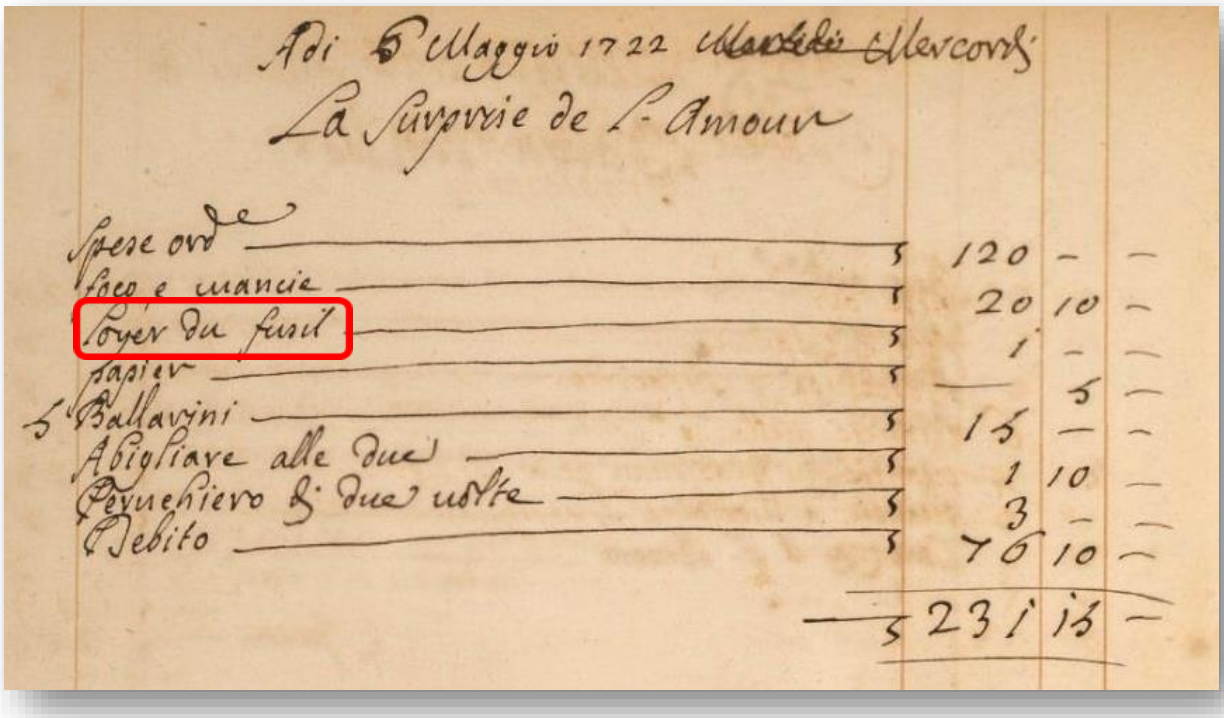
Les choix proprement techniques (langage de programmation, modèle de base de données, format d'échange, architecture technique, etc.) font l'objet de discussions entre ingénieurs et chercheurs en informatique. La présence de ces derniers est une caractéristique essentielle de CIRESFI ; nous avons observé dans bien des projets l'absence de chercheurs en informatique pour traiter des sujets relevant des humanités numériques. A contrario, nous prétendons que cet équilibre des compétences (scientifique et technique) ouvre la voie à de nouvelles méthodes de résolution des problèmes devenus habituels en humanités numériques (représentation des connaissances, ouverture et interopérabilité des données, intégration multi-sources, traitement du langage naturel, traitement d'images, nettoyage des données, propagation de l'incertitude, visualisation et exploration, analyse descriptive, analyse prédictive, etc.).

Le fusil et le lion, ou comment repenser la mise en scène

La base de données RECITAL répond d'abord à un besoin évident : il est impossible d'appréhender de manière livresque et linéaire plus de 25000 pages de registres comptables. Mais quel intérêt peuvent-ils présenter ? Bien des informations de ces registres doivent être expliquées pour devenir des données compréhensibles. C'est l'un des enseignements de ce travail de constitution de base de données : il s'agit en fait d'une inépuisable source de questionnements nouveaux.

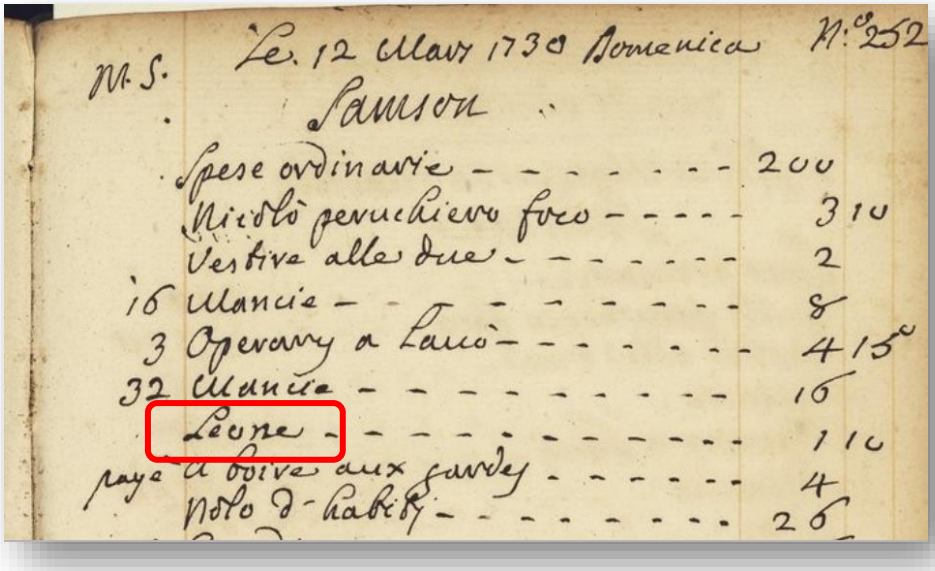
Ainsi l'utilisateur peut se demander : tel reliquat budgétaire sert-il une caisse noire ou s'agit-il d'une erreur de comptabilité ? Comment se fait-il que les places du parterre ne se vendent plus à partir de telle année ? Cette pièce a été déprogrammée puis reprogrammée quelques temps plus tard, selon une curieuse chronologie, pourquoi ? Qu'est-ce qu'un *escomba*, dépense qui revient régulièrement toutes les premières années ? Quel est l'usage de cet accessoire dans cette représentation ? Par exemple un fusil dans *La Surprise de l'amour* de Marivaux ? Pourquoi un lion dans *Samson* coûte-t-il seulement une livre dix sols ?

C'est alors que le savoir des « experts » est nécessaire : l'escomba n'est autre que la graphie italianisée du nom de famille Lescombat, nom de la concierge du Théâtre-Italien (trouvaille de Marion Danlos)…


Adi 5 Maggio 1722 Mercordi
La Surprise de L. Amour
Spese ord — 120 — —
Foco e mancie — 20 10 —
Loyer du fusil — 1 — —
Papier — — 5 —
Ballavini — 15 — —
Abigliare alle due — 1 10 —
Peruchiero p. due volte — 3 — —
Debito — 70 10 —
231 15 —

Page du registre TH-OC-11 mentionnant un fusil dans la pièce de Marivaux

Pour comprendre le fusil chez Marivaux, il faut se reporter à la fin du premier acte (scène IX), quand Arlequin arrive « *avec un équipage de chasseur* et s'écrie en voyant Colombine « me voilà justement sur le chemin du tigre », reprenant la métaphore que Lélio lui a expliquée au précédent : « Le tigre, c'est un caractère perfide retranché dans l'âme de ta maîtresse » (I, 2).


M.S. Le 12 Mars 1730 Domenica N.° 252
Samson
Spese ordinarie ---------- 200
Nicolò perruchiero foco ---- 310
Vestiva alle due ---------- 2
16 Mancie ---------- 8
3 Operary a Lauò ---------- 4 15°
32 Mancie ---------- 16
Leone ---------- 110
paye à boire aux gardes ------ 4
Nolo d'habits ---------- 26

Page du registre TH-OC-16 mentionnant un lion dans Samson

Quant au lion signalé dans les dépenses de la représentation de *Samson* (pièce créée en italien par Riccoboni en 1717, traduite et adaptée en français par Romagnesi en 1730), une livre dix sols ne représente pas une grande dépense ; elle est d'ailleurs réitérée à chaque reprise de la pièce. Un lion ne peut ni s'acheter ni se louer à Paris, et encore moins être mis en scène. Serait-ce alors un costume de lion ? Une machine ? Mais pourquoi repayer à chaque fois ?
Une réponse est fournie par la lecture de Voltaire, à l'article *Samson* des *Questions sur l'Encyclopédie*, lorsqu'il évoque la pièce : « On la représenta sur le théâtre français de la comédie prétendue italienne […] Dans cette pièce sublime, Arlequin, valet de Samson, se battait contre un coq d'Inde, tandis que son maître emportait les portes de la ville de Gaza sur ses épaules. » Ce lion du registre ne serait-il pas le parodique coq d'Inde (dindon) qui pouvait être mis à mort chaque soir, dans une action burlesque d'Arlequin, et qui ne figure pas dans le texte conservé de la pièce ? Ou bien désigne-t-on ainsi l'acteur qui pouvait être déguisé en lion, combattu par Samson (joué par Romagnesi) ? Nous remarquons que dix ans plus tard, lors des reprises de 1740 de *Samson*, le prix du lion passe à 3 livres. Et l'anecdote rapportée par d'Origny incite à pencher pour le volatile : « On raconte que Samson ayant été représenté à B... le dindon s'échappa de l'endroit où il était enfermé, précisément à la fin du premier quatuor de Lucile que l'on donnait pour petite pièce, et s'envola dans une loge […] » (Antoine d'Origny, *Annales du théâtre italien, depuis son origine jusqu'à ce jour*, Paris, Vve Duchêne, 1788, t. I, p. 115)
Ce ne sont là que quelques exemples montrant comment le fait de construire une base de données à partir de registres comptables peut conduire à des comparaisons et des hypothèses imprévues au départ.

Transformer des registres en données : deux approches technologiquement ambitieuses
L'existence des copies numérisées a permis d'envisager l'étape de collecte de données et de constitution de la base comme sous-objectif du programme CIRESFI. Dans la tâche de transformation des fac-similés en données structurées, nous avons suivi deux pistes innovantes et renoncé à la méthode traditionnelle. Cette dernière consiste à présenter des

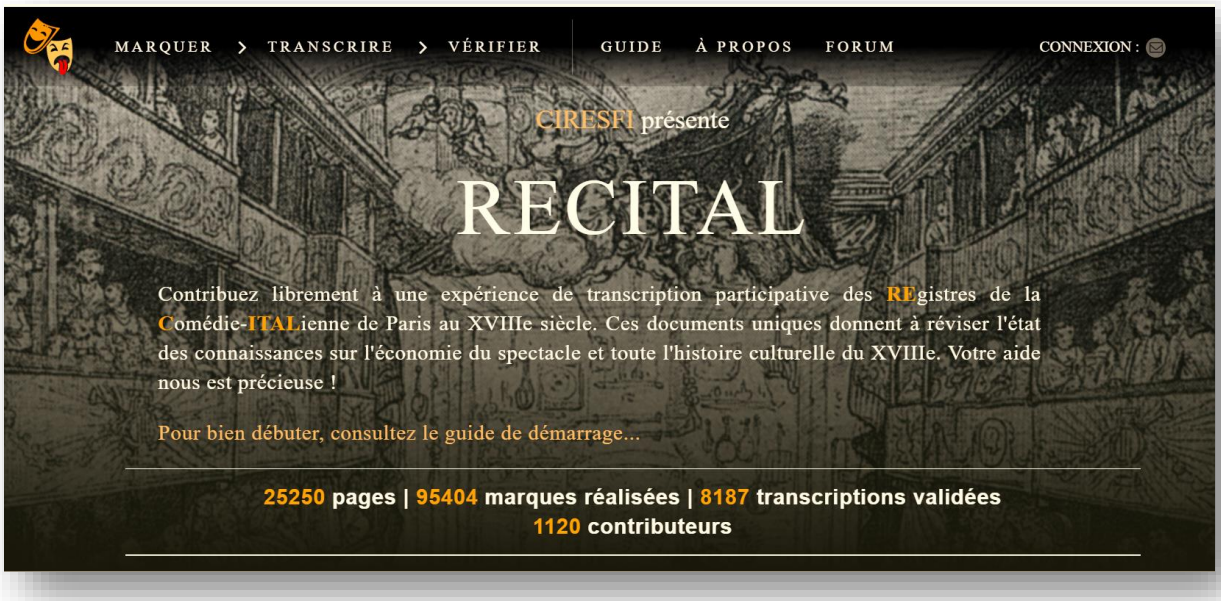
formulaires/interfaces de saisie parfaitement calibrés pour la source de données considérée, et bien alignés sur le modèle de données retenu. Une cohorte d'experts est alors convoquée pour réaliser cette tâche de saisie, en séparant le corpus en autant de sous-corpus que nécessaires. À l'issue de la saisie, une seconde phase manuelle de certification est mise en œuvre, afin de corriger les erreurs de saisie et confirmer la fiabilité des données collectées. Mais face à l'ampleur de la tâche (25 250 pages de corpus) et à la complexité du modèle de données, nous avons jugé inconcevable cette opération exclusivement manuelle et réservée à quelques experts du domaine. Il a donc été collectivement décidé d'expérimenter deux approches technologiquement ambitieuses en nous appuyant sur les compétences des équipes du Laboratoire des Sciences du Numérique de Nantes (LS2N).

La première approche, menée par l'équipe IPI (Image, Perception, Interaction) avec Harold Mouchère, Christian Viard-Gaudin en relation avec l'équipe TALN (doctorante Adeline Granet et post-doc Geoffrey Roman Jimenez) consiste à faire appel aux algorithmes de reconnaissance de forme (apprentissage automatique) pour extraire du texte (et des nombres) à partir des documents numérisés. Cette perspective, prometteuse en termes d'économie d'énergie, a été freinée par ce qu'on appelle « les caractéristiques inamicales » du corpus. En effet, l'ensemble des 25 250 pages de registres, partiellement dégradé par le temps, présente une hétérogénéité de formes, de scripteurs, de règles comptables, et même de langues (français et divers dialectes italiens), rendant impossible pour les technologies actuelles d'automatiser la tâche de reconnaissance d'écriture manuscrite. Partant, nous avons réduit l'ambition de cette piste de travail à trois objectifs :

1. la catégorisation des pages pour séparer les pages vierges des pages de compte quotidien ou celles de récapitulatif mensuel,
2. le cadrage (ou *zoning*), permettant de définir la zone de l'image qui contient une information spécifique (une date, un titre, etc.), et
3. la recherche par l'exemple (*word spotting*), offrant la possibilité de retrouver dans les 25 250 pages toutes les occurrences d'un terme dont on fournit une vignette en exemple.

Les deux premiers objectifs (catégorisation et cadrage) contribuent très exactement à l'élaboration de la base de données, comme nous le verrons au paragraphe suivant. Le troisième objectif (recherche par l'exemple) propose un mode d'exploration du corpus documentaire. Plus généralement, la démarche d'apprentissage automatique à partir des documents numérisés préserve de manière intrinsèque le lien entre la (méta)-donnée extraite et le document source. Les deux questions récurrentes et non triviales de ce type d'approche sont : comment évalue-t-on la fiabilité de ce qui est produit ? Sur quels exemples entraîne-t-on le dispositif à apprendre ? Une réponse possible se trouve dans l'articulation de la démarche d'apprentissage automatique avec la seconde voie empruntée pour résoudre le problème de collecte des données.

Cette seconde voie, empruntée par l'équipe DUKe (Data User Knowledge) avec Guillaume Raschia, les post-doc Marouane Hachicha et Benjamin Hervy, et Olivier Aubert, ingénieur et docteur en informatique) est celle de la production participative (ou *crowdsourcing*). Puisque ni l'expert du domaine, ni la machine ne sont en mesure de réaliser intégralement cette tâche de collecte, nous avons décidé de la confier à l'intelligence collective. Le principe consiste à ouvrir une plateforme disponible 24/7 sur le Web, par laquelle des bénévoles vont pouvoir réaliser des micro-tâches, à leur rythme, en fonction de leur disponibilité, de leur envie, et de leur compétence, et ainsi contribuer à la réalisation de la tâche de collecte intégrale des données du corpus. La plateforme RECITAL se trouve à l'URL http://recital.univ-nantes.fr.



Fragment de la page d'accueil de recital.org (décembre 2019)

La nature des micro-tâches est quadruple :

1. catégoriser une page,
2. marquer (encadrer) chaque information présente sur une page, en spécifiant son type
3. transcrire le contenu d'une marque, et
4. vérifier une transcription déjà soumise.

Exemple d'une page sur laquelle apparaissent en couleur les marques tracées par les utilisateurs

À l'issue de cette collecte participative, la base de données comprendra toute la transcription du contenu des 25 250 pages de registres, structurée selon 133 catégories d'information dont chaque occurrence conserve le lien avec le document source par le biais de sa marque dans la page. Il est à noter que, en décembre 2019, la collecte est encore en cours. Tous les lecteurs de cet article sont chaleureusement invités à y participer, ne serait-ce que pendant quelques minutes…

Bien entendu, le choix de confier cette tâche à une cohorte de bénévoles anonymes n'est pas sans effet de bord : la fiabilité des données produites devient une préoccupation essentielle de la démarche. Pour y remédier, nous mettons en œuvre un principe de redondance qui permet de diversifier les réponses à une même micro-tâche, puis nous appliquons un algorithme de recherche de consensus entre plusieurs avis. Nous utilisons pour cela un mécanisme de vote à la majorité. En la matière, la littérature scientifique en informatique propose différentes techniques de calcul de consensus.

Concernant les transcriptions, le consensus est évalué soit sur les données brutes, telles qu'elles ont été proposées par les bénévoles, soit sur une version transformée/normalisée par des traitements automatiques du langage naturel. La perspective d'une transcription modernisée (par opposition à une transcription diplomatique) est pertinente dans la mesure où la donnée doit se conformer à un modèle pré-établi. En outre, la préservation du lien donnée-document garantit qu'il est toujours possible de retrouver la "version originale" d'une transcription. Ainsi pour les noms de jour, Mercredi, Mercoledi, Mercordi, Mercoldi seront uniformisés en « mercredi » ; Arlequin, Arlecchino, Arl, Arlech, Arleq en Arlequin.

Le lien document/donnée : quelle objectivisation ?

Le lien donnée-document peut sembler une contre-mesure au phénomène d'objectivisation. Néanmoins, la persistance de l'effet d'essentialisation réside dans la complexité de la chaîne

de traitements depuis le document source jusqu'à la donnée. Tentons d'en exposer les principes. Par consensus à la majorité et suite à la multiplication des avis de bénévoles pour traiter une micro-tâche (par exemple produire la transcription d'un titre de soirée), la plateforme RECITAL retient la (le groupe de) transcription(s) majoritaire(s) comme donnée définitive.
Mais à ce stade le travail n'est pas encore achevé ;+ il est encore nécessaire de transformer la donnée recueillie par page en une donnée propre à une soirée théâtrale, autrement dit, de convertir un modèle proche du support (registre, page, marque dans la page, transcription de la marque), en un modèle proche du champ d'étude (soirée, représentation, budget, recette, dépense, troupe, acteurs, etc.). Là encore, une batterie d'algorithmes est mise en œuvre pour réaliser cette conversion, exploitant des techniques de traitement du langage naturel, d'alignement de séquences, de segmentation/agrégation, etc.
Il est donc apparu assez clairement qu'il est nécessaire de documenter/publier l'ensemble de ces processus de transformation, et plus encore, de les inscrire dans la base de données même. Nous avons donc implémenté une version liminaire du schéma W3C PROV-O pour la représentation de la provenance. Parallèlement, nous avons muni chaque fragment de donnée d'un indicateur de fiabilité, calculé à partir de la chaîne de traitements qu'il a subis depuis la source documentaire, jusqu'à la base de données. Actuellement, ces problèmes nourrissent la réflexion des chercheurs en informatique du projet CIRESFI, qui étudient un modèle de base de connaissances historiques comprenant la gestion de l'incertitude et de la provenance.

Attirer les utilisateurs : le cas des plateformes de crowdsourcing

Dans la phase initiale de conception a été dressé un premier inventaire des questions à poser aux données, des modes d'accès et de parcours à envisager, des restitutions visuelles et graphiques à proposer. Néanmoins, et comme cela a déjà été souligné, l'étendue des questionnements scientifiques suscités par les données des registres étant quasi infinie, la forme ultime de diffusion de la base de données est en soi un défi intéressant pour les chercheurs en sciences du numérique du projet CIRESFI.
En outre, et il s'agit d'une spécificité du projet CIRESFI, la phase de crowdsourcing pour la collecte des données a eu comme double conséquence de promouvoir la base de données auprès d'un large public, avant sa mise à disposition dans la forme finale, et de générer de nouvelles questions au gré de la réalisation de tâches de transcription. Dès l'ouverture de RECITAL nous avons communiqué sur Facebook et sur Twitter ; les nombreux relais ont été un moyen de faire connaître le projet ; le premier de ces tweets totalise à ce jour 5297 vues :

Exemples de tweets encourageant la participation au crowdsourcing sur RECITAL

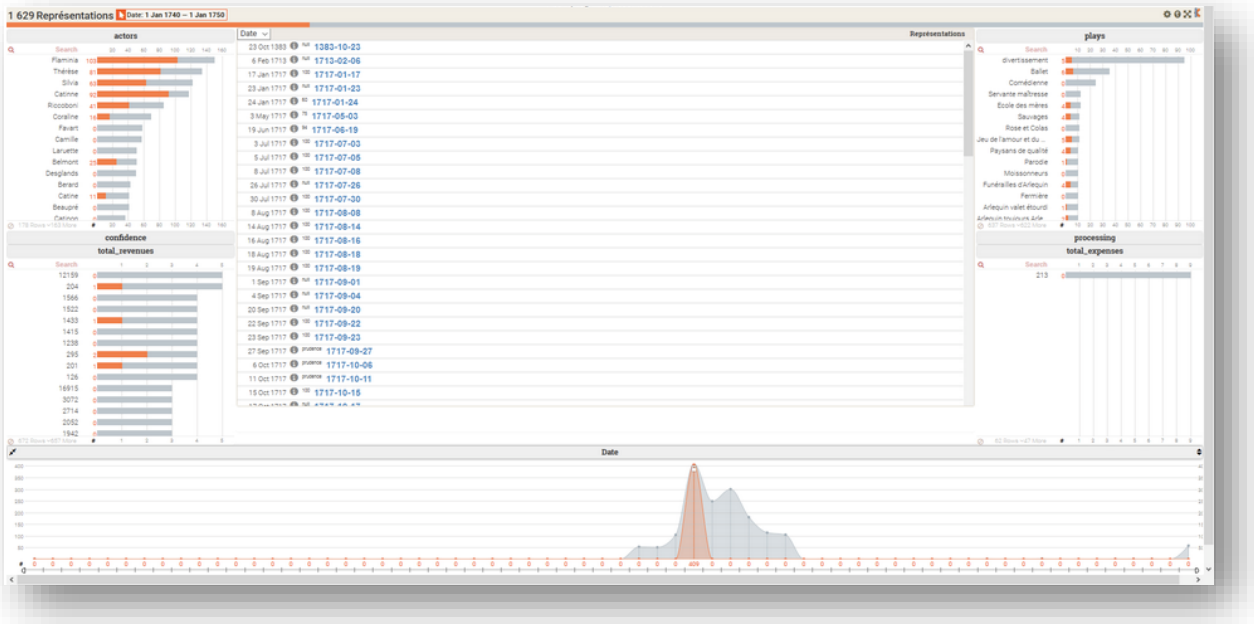
S'il est primordial d'atteindre un grand nombre d'utilisateurs, il peut être bien d'aider ceux d'entre eux qui se posent des questions techniques ou scientifiques. Amélie Renard a créé un guide d'introduction et des bulles d'aides signalées par un point d'interrogation. Le forum de RECITAL mis en place par Benjamin Hervy à partir d'octobre 2017 permet d'enregistrer les remarques des utilisateurs (pour y accéder il faut s'inscrire au forum sur le site de RECITAL), et d'y répondre (selon le temps disponible des experts bénévoles), ou bien de prendre en compte les problèmes pour améliorer la plateforme.
Comme pour les réseaux sociaux, il faut aussi fidéliser les utilisateurs qui concourent à enrichir la base de données. Nous avons mis en place une newsletter (dont on trouve un exemple ici) avec quatre rubriques (Les chiffres du mois / sciences et techno / le coin de l'histoire / prochain défi). Produit collaboratif, la newsletter nécessite la disponibilité d'un coordinateur (après le remarquable lancement dû à Benjamin Hervy, nous manquons actuellement de bonnes volontés pour continuer).
Les bases de données, selon nous, ne peuvent se passer de convivialité non numérique. Pour maintenir l'élan qui s'est manifesté lors du démarrage de RECITAL, nous organisons régulièrement des journées de transcription : une dizaine de bénévoles se retrouvent avec leur ordinateur portable, et passent du temps à transcrire… et à échanger, ce qui permet aussi de perfectionner l'outil. Enfin, certains enseignants comme Isabelle Ligier-Degauque à l'Université de Nantes ont eu l'idée d'intégrer dans leur enseignement de master recherche une initiation à la base de données comprenant un nombre d'heures à passer sur RECITAL, qui peuvent être vérifiées par un système mis au point à Polytech Nantes.

De la plateforme de crowdsourcing à la base de données
Souhaitant offrir le plus large choix possible d'options pour la diffusion de la base de données, nous avons identifié 4 modes complémentaires :

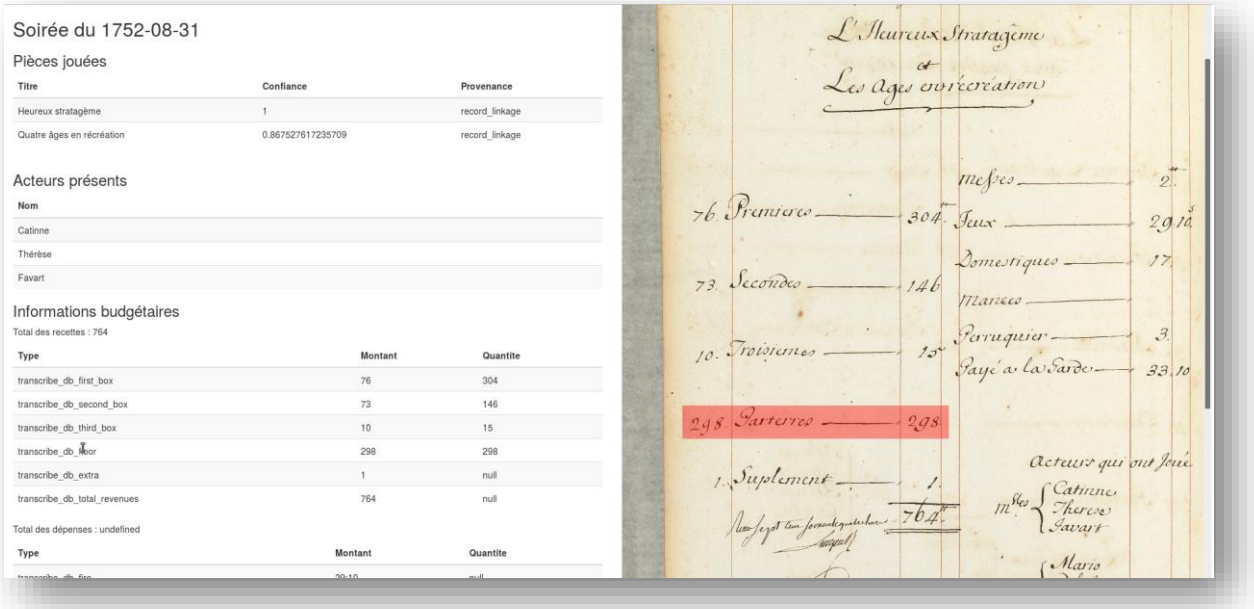
1. une interface de programmation,
2. une interface d'exploration interactive,
3. une vue calendaire,
4. des résultats d'analyses thématiques.

L'interface de programmation (API, pour *Application Programming Interface*) fournit les données « brutes » sous une forme standardisée (JSON), avec la possibilité de filtrer selon différents critères (la période, le titre de la pièce, etc.). Ce mode de diffusion est destiné à des

programmeurs/analystes qui vont être en mesure de s'approprier les données sans parti-pris et produire de nouvelles restitutions. C'est un mode qui favorise l'innovation mais qui, en contrepartie, requiert une solide compétence technique.

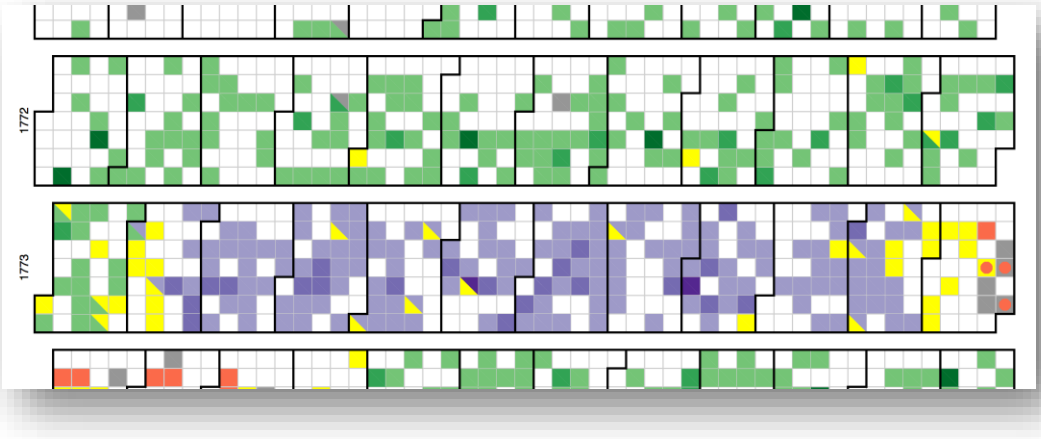


L'interface d'exploration interactive propose d'entrer dans le jeu de données par une recherche à facettes, permettant de filtrer progressivement les données sur différents critères pour isoler le plus rapidement possible les zones d'intérêt. L'outil Keshif est employé pour produire cette interface.



Au terme d'un parcours de recherche, l'utilisateur est amené à consulter un écran qui présente le détail d'une soirée théâtrale. Celui-ci incarne le lien donnée-document puisqu'on y trouve notamment le fac-similé en regard des données chiffrées de la soirée. En outre, des

liens hypertextes vers d'autres soirées (présentant des similitudes avec la soirée courante) sont proposés, matérialisant le concept de « navigation sémantique » dans le jeu de données.



Représentation calendaire des années 1772 et 1773

La vue calendaire est un choix de restitution visuelle fondé uniquement sur l'idée que la donnée historique issue des registres est ancrée dans une séquence chronologique qu'il nous a semblé pertinent de mettre en relief. Cette interface présente le siècle sous la forme d'une liste de rectangles quadrillés, chaque séquence horizontale représentant une année calendaire scindée en douze mois et en jours. Le coloriage de chaque carré-jour du quadrillage apporte une information visuelle macroscopique qui peut être paramétrée (total des dépenses, fréquentation, soirée avec Arlequin, etc.). Là encore, un lien hypertexte offre la possibilité d'ouvrir l'écran qui détaille les données d'une soirée.
Enfin, nous proposons différentes restitutions graphiques sur commande, en fonction d'un thème à l'étude et de critères très spécifiques, permettant de déployer une large palette d'analyses et de visualisation de données, au profit d'une question scientifique. Ce mode de restitution requiert l'intervention d'un expert en analyse de données, et un dialogue étroit avec l'historien.
Comme l'ensemble des dispositifs décrits ci-dessus le suggère, il a été fait le choix de proposer une restitution des données la plus agnostique qui soit. Un travail d'éditorialisation pourrait néanmoins être proposé en complément des modes déjà présentés, de manière à éveiller la curiosité chez le visiteur béotien et à favoriser la circulation des connaissances.
Enfin, la spécificité du projet CIRESFI/RECITAL reposant sur l'usage du crowdsourcing, les quatre modalités présentées se déclinent en deux variantes : les données du spectacle, et les données du crowdsourcing ! En effet, outre la base de données « métier », il nous a paru intéressant de mettre à disposition les traces d'activité de la plateforme de crowdsourcing, notamment pour favoriser l'innovation en matière de mécanismes de suivi de l'activité, de résolution de consensus, d'évaluation des bénévoles, de certification des données, etc.

Pérennité et interopérabilité
Même s'ils permettent souvent la naissance de bases de données, les financements de projets (CPER, ANR etc.) ne permettent pas d'assurer leur maintenance et leur permanence au-delà de la période du projet, d'autant que les ingénieurs d'étude et les post-doctorants ne sont eux-mêmes ni pérennes ni interopérables…

Le support technique requis pour de la maintenance corrective et évolutive de RECITAL manque cruellement, puisque la construction de la base de données a été financée dans le cadre du projet ANR CIRESFI, sans option pour la pérennisation. Néanmoins, la permanence des enseignants-chercheurs en informatique impliqués dans le projet permet un suivi minimal et la résolution des pannes et anomalies qui surviennent de façon sporadique.
De même que la maintenance de THEAVILLE repose sur le bénévolat, l'engagement et la générosité d'une personne (Florent Coubard), celle de RECITAL repose sur un enseignant-chercheur (Guillaume Raschia), qui prend alors sur son temps de recherche malgré les impératifs de productivité à la mode, liés à la gouvernance de la recherche par les statistiques. Pourtant ces machinistes des coulisses sont essentiels au fonctionnement de nos bases de données aujourd'hui.
D'où l'importance cruciale des choix de départ. En ce qui concerne RECITAL, les choix techniques pour la constitution et la diffusion de la base de données ont été opérés par les chercheurs en informatique, en concertation avec les ingénieurs-développeurs, et en tenant compte des infrastructures disponibles dans l'environnement du projet. Les langages de programmation, le système de gestion de base de données, les librairies logicielles, ainsi que le format d'encodage des données sont ouverts, standardisés et sous licence libre. En outre, l'infrastructure repose sur une plateforme de virtualisation mutualisée de l'Université de Nantes, qui assure la disponibilité et la pérennité du service.
Il n'a en effet pas été fait le choix de migrer sur l'infrastructure Huma-num, pour les raisons suivantes. Bien que l'adhésion aux principes FAIR relève d'une démarche vertueuse pour l'ouverture des données, elle est, selon nous, comprise et instanciée de façon « rigoriste » par le déploiement systématique de technologies dites du web sémantique (RDF/RDFS, OWL, SPARQL) ou, dans une version édulcorée, Dublin Core et OAI-PMH. Or il suffit de proposer une documentation, fournie en ligne avec le jeu de données, pour prétendre à la conformité aux critères *Findable*, *Accessible* et *Reusable* du FAIR, comme le pratiquent l'ensemble des plateformes *Open Data* du monde entier. Quant au principe d'Interopérabilité, sa mise en œuvre effective est à la fois complexe et illusoire, puisqu'elle requiert une stabilité absolue des silos de données que l'on souhaite interconnecter d'une part, et nécessite une très bonne connaissance du schéma (modèle, structure) des données pour construire des requêtes/filtres corrects d'autre part.

Conclusion

L'élaboration de la base de données RECITAL a permis de fédérer des chercheurs de sciences humaines et de sciences du numériques qui ne travaillaient pas ensemble. La recherche a progressé dans les deux domaines, prouvant aux littéraires et aux historiens que les informaticiens ne sont pas leurs serviteurs, mais que le numérique en tant que discipline est objet de recherche, d'impasses et de résultats inattendus. Les chercheurs en numérique ont de leur côté mieux compris les raisonnements, les méthodes et les objectifs de leurs collègues des humanités.
Des collaborations ont aussi vu le jour avec les chercheurs d'autres projets touchant au théâtre ; il est à souhaiter de pouvoir travailler davantage avec les responsables d'un projet voisin, consacré aux registres de la Comédie-Française (Projet RCF).
De même que la base THEAVILLE a permis à beaucoup de prendre conscience de l'importance de la musique (les vaudevilles) dans les pièces des théâtres de la Foire et de la Comédie-Italienne, la participation des historiens du théâtre à la construction de RECITAL a attiré leur attention sur une multitude d'informations qui enrichissent considérablement les

connaissances sur la mise en scène des pièces, sur les petits métiers des gens de théâtre, sur l'apparition du droit d'auteur etc. Nombre de chercheurs et d'amateurs de théâtre souhaitent désormais savoir ce que l'on peut trouver dans les registres… Les deux bases de données THEAVILLE et RECITAL ont indubitablement pour effet de concevoir l'histoire des spectacles d'une manière bien plus contextuelle et hypertextuelle qu'auparavant. Elles attirent aussi les praticiens, acteurs, chanteurs, metteurs en scène, qui peuvent dans l'une apprendre à chanter des airs et les entendre, dans l'autre imaginer bien plus concrètement la création ou la recréation de savoureuses comédies du siècle des Lumières.