



« Toulouse Capitole Publications » est l'archive institutionnelle de

**Travailler sur la télévision aujourd'hui à partir des archives de l'INA.
Constitution de corpus et méthodologies**

**Laurence Leveneur-Martel, MCF HDR en Sciences de l'information et de la
communication**

Institut du Droit de l'Espace, des Territoires, de la Culture et de la Communication

Université Toulouse Capitole/IUT de Rodez, IDETCOM

Pour toute question sur Toulouse Capitole Publications,
contacter portail-publi@ut-capitole.fr

Travailler sur la télévision aujourd'hui à partir des archives audiovisuelles de l'INA

Remerciements

Dans le spectre des images que l'on étudie en sciences humaines et sociales, je vais aujourd'hui évoquer celles qui m'occupent en tant que chercheuse en sciences de l'information et de la communication depuis plus de vingt ans maintenant, en l'occurrence **les images de télévision**. Et lorsque l'on travaille sur un tel objet en France, nous avons la chance d'avoir accès **l'INA**, qui est un service public dont les collections et services sont, à ma connaissance, quasiment uniques au monde.

Je me permets de préciser que mon intervention **n'a pas pour objectif de vous présenter un état des lieux exhaustif** des collections et outils proposés par l'INA aux chercheurs, je n'ai aucune légitimité pour le faire. En revanche, je vais plutôt vous faire part **des expériences que j'ai pu avoir avec l'INA**, notamment lorsqu'il m'a fallu constituer **différents corpus** répondant à mes objectifs de recherche, en particulier des corpus composés d'images fixes ou animées, c'est selon, et que ce soit dans le cadre de travaux personnels, collectifs, ou bien lorsque j'ai bénéficié du statut de chercheuse associée proposé par cet organisme.

Rôle et missions de l'INA

L'INA a été chargé par la loi du 20 juin 1992 de la conservation et de la communication des œuvres et documents de radio-télévision au titre du **dépôt légal**. En d'autres termes, l'INA **capte depuis le 1er janvier 1995** toutes les chaînes de production française 24h/24, 7 jours sur 7, ce qui représentent plus de 20 millions d'heures d'images et de sons.

L'INA collecte et conserve également les **sites internet** relevant du domaine de la communication audiovisuelle ([projeter liste](#)). Le décret du 19 décembre 2011, organisant le dépôt légal du web, est venu en application de la loi du 1er août 2006 (loi DADVSI sur les droits d'auteurs et les droits voisins dans la société de l'information). Lancée de manière expérimentale dès février 2009, la collecte des sites web dans le cadre du dépôt légal du web couvre **16 069 sites web médias** (soit 104 milliards de versions d'url). Il concerne également **15 637 comptes Twitter** (1,95 milliard de tweets collectés depuis 2014) et **10 454 comptes de plateformes vidéo** (28 millions de vidéos collectées depuis 2008), auxquels vont s'ajouter désormais l'archivage de certains contenus audiovisuels diffusés sur **Tik Tok**.

Ses collections sont accessibles sur l'ensemble du territoire : à la Bibliothèque nationale de France à Paris, dans 6 délégations régionales, et dans plus de 25 médiathèques et cinémathèques.

[Projeter la cartographie des fonds INA.](#)

Il faut également signaler qu'au-delà de ces fonds purement audiovisuels, l'INA met à la disposition des chercheurs, en particulier à l'Inathèque à Paris, **des sources écrites** (fonds publics ou privés), et une **photothèque** principalement accessible sur rendez-vous. [Projeter schéma des fonds écrits \(chronologie\).](#)

Ces fonds photographiques qui documentent la production audiovisuelle française sont composés de plusieurs éléments résumés ici ([projeter diapo](#)) :

Il existe d'autres fonds et collections liés au spectacle vivant que je n'aborderai pas ici, je vous invite en revanche à explorer **le catalogue en ligne de l'ina sur le site de l'inathèque de France**. Je souhaiterais compte tenu de ces remarques préalables revenir sur les **avantages et les écueils** de la constitution d'un corpus audiovisuel à partir des archives de l'INA à la lumière quelques-une de mes expériences personnelles.

Dresser une cartographie des jeux télévisés français

L'une des premières expériences de chercheuse que j'ai pu avoir avec l'INA est celle qui est relative à ma thèse de **doctorat** sur les jeux télévisés français.

Mon **questionnement** originel, pour le présenter de façon simple, s'articule autour de la double interrogation suivante : comment la télévision fait-elle jouer ses téléspectateurs depuis le début des années 1950 et de quelles façons cette relation ludique mise en place par le médium a-t-elle évolué ?

Pour répondre à ces questions de départ, il me fallait commencer par **déterminer la place réelle qu'occupe le jeu télévisé** sur les principales chaînes hertziennes généralistes françaises – auxquelles je me suis limitée –, depuis ses débuts. Pour cela, il me fallait dresser un **panorama global** des jeux télévisés français qui, sans prétendre à l'exhaustivité, me permettrait d'établir un **premier terrain d'observation** du genre, à partir duquel j'ai souhaité **construire mon corpus d'analyse**. Ce corpus pouvant dès lors être considéré comme **représentatif** de ce panorama dont il constituait un véritable **échantillon**.

L'élaboration de ce panorama supposait de partir **d'une définition ad hoc du genre** dont la construction mériterait en soi une communication, tant la notion de jeu est labile.

Cette **cartographie initiale**, qui commence en 1950 et s'achève au premier semestre 2004, et qui se veut aussi exhaustive que possible, **est le fruit du croisement de plusieurs sources**. La première d'entre elles est celle que met à la disposition de ses usagers l'Inathèque de France, centre de consultation de l'INA, à travers une interface de consultation qui est appelée **hyperbase**. Ce logiciel permet d'interroger toutes les **bases de données** disponibles, qu'elles soient écrites ou audiovisuelles, par l'intermédiaire de différents champs documentaires qui sont les éléments à partir desquels les documentalistes de cet institut construisent les notices écrites qui décrivent des émissions. Il existe, pour la télévision, **plusieurs fonds documentaires possibles** sur l'Hyperbase : un fonds qui recense les archives audiovisuelles, qui peut remonter jusqu'au début des années 1950, et un fonds issu du Dépôt Légal (mis en place depuis 1995). Montrer interface : fonds DL. Recherche documentaire informatisée : expliquer

Dans les bases de l'INA, chaque émission de télévision est accompagnée d'une notice descriptive répondant à plusieurs champs d'indexation (canal de diffusion, heure et date, titre de la collection, titre propre, genre, thématique, résumé, etc.). La lecture de ces **notices** m'a permis de ne retenir que les titres correspondant à ma définition des jeux télévisés tout du moins lorsque celles-ci étaient renseignées. Ce qui n'est pas systématiquement le cas sur la base des archives de l'INA où les notices des émissions sont parfois restreintes au titre, au canal et à la date de diffusion, ou encore aux auteurs, sans aucun renseignement sur le contenu ou le genre des émissions, notamment lorsque celles-ci sont très anciennes et ne sont pas directement accessibles.

Cette première étape de recherche sur les bases de données de l'INA a été complétée grâce aux différentes « **fiches collections** » **des archives ou du Dépôt Légal**, qui sont des notices particulièrement détaillées, présentant de façon diachronique certaines collections phares* (exemple : fiche collection Popstars), et qui propose, en plus de la description habituelle du contenu de l'émission, une description du dispositif scénique, de la mise en scène, etc. Toutefois, certaines des émissions répertoriées n'étant pas décrites dans les notices, il m'a fallu compléter ces informations par **la lecture de journaux de programmes** (principalement *Télérama* et *Télé 7 jours*), de **dictionnaires** sur la télévision, ainsi que par celle des rapports écrits après diffusion par les chefs d'antenne, notamment pour vérifier si les émissions recensées, avaient bel et bien été diffusées. Autrement dit, ce repérage de contenus audiovisuels s'est d'abord fait sur **la base de notices descriptives et de fonds issus des archives écrites que j'ai mentionnés au début**.

Le croisement des sources est d'autant plus indispensable que certaines fiches de l'INA reprennent les résumés fournis par les producteurs des émissions, ce qui nécessiterait, dans l'idéal, un visionnage systématique des émissions recensées, permettant de vérifier la forme, le contenu et les règles des jeux répertoriés. **La constitution du panorama** m'a pris un an de travail, le visionnage de l'ensemble des collections recensées n'était évidemment pas possible le temps du doctorat. De plus, si les bases de données de l'INA, et principalement les archives, permettent d'avoir connaissance de l'existence d'émissions passées, l'ancienneté de certains supports rend leur consultation impossible, et sur l'ensemble des collections recensées, **la part des programmes directement accessibles** à l'époque était très

faible jusqu'en 1986, date de la mise en place du Service d'Observation des programmes [1986-1994]. Voir graphique avec pourcentage des émissions directement accessibles

De toute évidence mon panorama souffrait donc de l'absence d'une grande partie de ces supports, principalement les plus anciens.

Enfin, l'*hyperbase*, si elle répertorie une grande partie des émissions télévisées françaises, n'offre pas une source exhaustive et laisse apparaître certains « **vides** », notamment avant la mise en place du Dépôt Légal. Ainsi, une collection aussi pérenne que *Motus* n'est pas référencée avant 1995, alors qu'elle a été diffusée en France dès 1990. Avant le Dépôt Légal, l'INA ne répertoriait que les productions originales françaises, et les jeux télévisés ne faisaient pas partie des émissions prioritairement archivées, comme par exemple les journaux d'information ou certains événements politiques médiatisés. La collection *Des Chiffres et des lettres*, dont la longévité en fait l'un des emblèmes du genre, n'apparaissait alors dans la base des archives de l'Inathèque que lorsque des tournois ou des championnats avaient été organisés. Les émissions quotidiennes d'avant 1986 n'étaient donc pas directement accessibles. De plus, les émissions issues **d'adaptations ou de coproductions étrangères** comme *La Roue de la fortune*, diffusé en France depuis 1987, n'ont pas été systématiquement archivées avant 95. Seules quelques recopies qui m'ont été fournies par l'INA sur demande m'ont permis d'accéder à certains de ces programmes. Je parle au passé parce depuis, l'INA a produit un considérable de numérisation de ses archives.

En somme, **le croisement des données et les informations fournies par l'épître éditorial** qui accompagne les émissions ne pouvaient suffire à rendre compte de l'ensemble des programmes de jeux créés et diffusés en France entre 1950 et 2004. Ce panorama ne peut donc prétendre à l'exhaustivité, d'autant moins que le **hasard** m'a parfois mise sur la piste de programmes inconnus des archives de l'INA ou des recherches précédentes, à lecture d'un journal de programme ou d'un article. Phénomène de **sérendipité** que Gérard Genette résume avec humour dans *Palimpsestes* en disant que : « L'inconvénient avec la Recherche, c'est qu'à force de chercher, il arrive que l'on trouve... ce que l'on ne cherchait pas ! »¹

Au-delà de ces premières difficultés méthodologiques, **les classifications génériques qui sont employées par l'Inathèque de France méritent d'être questionnées**. L'observation de ces nomenclatures révèle tout d'abord **l'instabilité** des critères retenus par les documentalistes de l'INA dans les typologies des notices descriptives des émissions. Ils emploient par exemple pour qualifier le genre des programmes plusieurs termes à la fois (montrer exemple de *Popstars* : plusieurs termes sont utilisés pour qualifier le genre (Feuilleton, documentaire, télé-réalité)². Ces termes sont discutables : ils sont en plus liés au choix des producteurs de définir *Popstars* comme un documentaire a par exemple suscité au CSA une concertation publique, notamment parce que ce choix induit que l'émission entre dans la catégorie des œuvres audiovisuelles, et qu'elle bénéficie dès lors des soutiens financiers qui leur sont allouées, ce qui n'est pas le cas des autres jeux télévisés, normalement exclus de la définition de l'œuvre par le CSA.). Le genre de l'émission *Popstars* est variable selon les notices : il est désigné par les mots clés « télé-réalité, jeu » sur la base DL (montrer).

La logique de l'indexation oblige l'INA à **une typologie « descriptive et cumulative³ »** qui conduit les documentalistes à préciser ces éléments, pour mieux rendre compte de la complexité et de la variété des programmes répertoriés. Ce qui les amène à **multiplier les catégories** servant à désigner une seule émission, jusqu'à en employer trois à la suite, suivant un ordre de prédominance.

¹ G. GENETTE, *Palimpsestes*, Paris, Seuil, coll. Point/ Essais, 1982, p. 8.

² L'assimilation de ces termes contradictoires suscite d'ailleurs de multiples interrogations au sein même de l'Inathèque : le documentaire, même associé au feuilleton, ne crée pas la réalité. Il est au contraire censé la restituer, même si le montage en montre une forme un peu scénarisée, tandis que la télé-réalité, dans son principe, crée ses propres règles, met en scène des candidats dans un lieu clos afin d'en suivre la vie quotidienne. Si dans le *feuilleton documentaire* la caméra se veut transparente, elle ne l'est pas dans les émissions de télé-réalité, comme l'atteste d'ailleurs la jurisprudence requalifiant en contrats de travail les contrats de participation de certains ex-candidats de l'*Ile de la tentation*.

³ Cf. P. CHARAUDEAU, "Problèmes de typologisation", *Comptes-rendus des Ateliers de recherche méthodologique de l'INA*, 25 avril 1997, p. 58.

Toutefois, **l'hétérogénéité des termes** employés par les documentalistes de l'Inathèque pour désigner une même collection, si elle se justifie par le caractère hybride d'une grande partie de ces programmes, suppose de la part du chercheur **une attention toute particulière face au choix des « mots clés »** utilisés pour répertorier les émissions de télévision. Le champ « genre » des notices de l'Inathèque doit aussi être complété par d'autres champs d'indexation comme la « thématique » par exemple qui donne des précisions sur le contenu de l'émission, voire parfois sur sa « forme » : la notice de l'émission *Popstars* affiche comme thématique le terme « variétés », que l'on pourrait également considérer comme un genre... Montrer notice.

Cette logique d'indexation propre à l'INA **ne peut se substituer à celle qui mène l'activité de la recherche**. Cette dernière doit donc faire preuve, face aux outils qui lui sont proposés par le centre de consultation de cette institution, d'une **certaine méfiance**. Ce qui suppose concrètement de la part du chercheur de **suspendre** provisoirement, le temps de la recherche sur les bases de données de l'INA, « **ses propres postulats théoriques**.⁴ » Il faut en effet **comprendre préalablement** la logique d'indexation qui oriente les typologies de l'Inathèque afin d'orienter l'étude le plus efficacement possible. Cela nécessite notamment d'avoir accès aux définitions utilisées par les documentalistes pour indexer les émissions, ce qui ne peut se faire que sur demande explicite ou par l'intermédiaire des archives écrites consacrées à cette question. Montrer document

Suivant cette méthode, ce sont donc près de **544 collections** qui figurent dans ce panorama des jeux télévisés. L'on peut estimer, pour chaque décennie, un **pourcentage d'erreur** correspondant aux collections étiquetées ou présentées comme des jeux et pour lesquelles je n'ai pas pu obtenir suffisamment d'éléments sur le contenu, ni même les visionner. (Montrer graphique : marges d'erreur). Ce panorama offre néanmoins un terrain d'observation privilégié à partir duquel j'ai pu constituer mon **corpus**, dont seule l'analyse m'a véritablement permis d'entrer au *cœur* du genre, au-delà des simples étiquettes génériques, et de réduire ces marges d'erreur. (Montrer tableau de présentation du corpus). Par ailleurs, j'ai pu intégrer dans ce panorama général les titres des collections répondant à ma définition préalable, les informations concernant les dates, fréquences et durées de diffusion de ces programmes, en préciser le concept, et y indiquer les noms de leurs principaux producteurs, auteurs, animateurs et réalisateurs. Projeter capture du panorama

Autrement dit, j'ai constitué **un panorama un peu enrichi** des jeux télévisés, qui est accessible à l'Inathèque de France (dans un format papier, donc moins valorisé que s'il était numérisé) et dont je me sers encore car à ce jour, puisqu'il est le seul qui existe sur le genre télévisuel en question. Mais sachez qu'il est possible à l'Inathèque de **constituer des corpus numérisés et annotés** différemment, notamment grâce à un outil qui s'appelle **Médiacorpus**, outil d'annotation qui permet aussi des visualisations graphiques. Montrer exemple du débat entre deux tours sur site INA.

Un travail sur les femmes animatrices et productrices de jeux et variétés télévisés

Le panorama que j'ai constitué m'a récemment servi dans le cadre d'un **travail sur les femmes animatrices et productrices de jeux et variétés à la télévision**.

Mon questionnement portait **alors sur le rôle énonciatif et la trajectoire professionnelle** des animatrices dans les jeux et variétés télévisés : ce rôle est-il similaire à celui des hommes ? Comment se construisent ces figures, dans leurs interventions à l'antenne et leur comportement ? Les femmes animatrices sont-elles affectées à certains types de jeux ou émissions de variétés dont le contenu ferait écho à des compétences dites « féminines » ou observe-t-on des similitudes avec les programmes tenus par leurs collègues masculins ? Que dire des trajectoires professionnelles des femmes animatrices dans ce genre télévisuel : observe-t-on des récurrences et surtout des atypies dans leurs parcours à l'antenne, des exceptions qui nous renseigneraient alors sur les normes en vigueur dans ces professions (Coulomb-Gully, 2010, p. 9) ? Que nous enseigne enfin la comparaison de ces trajectoires avec celles de leurs homologues, hommes et femmes, dans les variétés télévisées ?

⁴ *Idem*, p. 91.

Ces questionnements impliquaient au préalable de **choisir des définitions précises permettant de circonscrire le terrain d'étude et faciliter le repérage de programmes souvent hétérogènes ou hybrides**. Pour cela, nous disposions des **résultats du travail mené précédemment**, qui nous avait permis d'élaborer une définition stable des jeux télévisés :

Le **panorama** constitué précédemment mentionnant notamment les animateurs des émissions répertoriées, nous avons pu appuyer notre **repérage** des femmes qui occupent cette fonction sur ce travail préalable et l'actualiser par une recherche sur les bases de données de l'Institut National de l'Audiovisuel (I.N.A.) pour les programmes diffusés entre 2007 et 2019.

Concernant les programmes de variétés, nous avons choisi de retenir **la définition qu'en donne Alice Norblin**.

N'ayant pas trouvé de cartographie détaillée des variétés télévisées françaises, et ce en dépit de plusieurs travaux historiques portant sur ce genre, nous avons dû faire appel à **l'expertise des documentalistes de l'I.N.A.** En effet, l'outil de consultation hyperbase ne permet pas de procéder à une recherche dans les notices documentaires par **sexe**, car la « qualité » de l'intervenant n'est pas toujours renseignée dans les notices de l'INA :

Chaque personne rentrée dans notre lexique, lexique qui permet de renseigner les génériques, est identifiée par son sexe et, autant que possible, par une « note qualité » qui la définit. Ex : Carpentier, Maritie [note qualité] Productrice TV, France. Ce sont ces champs « note qualité », et le sexe des personnes du lexique, qui ne sont pas interrogeables par nos moteurs de recherche sans passer par du SQL .

Il nous a donc fallu récupérer des extractions issues des bases du dépôt légal et archives de l'INA, en passant par **une requête SQL⁵** que seuls certains documentalistes peuvent réaliser. Nous avons ainsi obtenu une extraction issue du dépôt légal et deux extractions issues du fond archive (**base nationale et base régionale**) qui recensent les animatrices/présentatrices/réalisatrices et productrices référencées dans les émissions de variétés. Montrer capture

Nous avons **trié ces fichiers**, à l'aune de la définition choisie, afin d'en éliminer certains programmes qui n'y correspondaient pas. Nous avons **complété ces données par la lecture des index de deux dictionnaires historiques** de la télévision (Jeanneney et Chauveau, 2001 ; Chauveau et Dehée, 2007) pour relever les noms des femmes citées et vérifier si elles avaient un rôle dans ces deux genres de programme. Ces informations ont enfin été complétées par des requêtes plus ciblées via l'hyperbase de l'INA. Nous avons complété cette cartographie partielle par **un visionnage de certaines** émissions de jeux et variétés.

Ces données ont été complétées **par la lecture de certaines coupures de presse de télévision**, le visionnage ou l'écoute **d'émissions radiophoniques ou télévisuelles** archivées par l'INA, ainsi que l'accès à certains fonds **d'archives écrites** contenant des coupures de presse ou des documents de production. En effet, travailler sur la télévision peut impliquer l'analyse de **certaines images qui ne sont pas issues des programmes télévisuels**, mais dans le cas présent de certaines **couvertures** de magazine, qui en disent long sur le rôle et la place des femmes en question dans l'univers médiatique, mais aussi des **photographies** de plateau disponibles dans certains fonds d'archives écrites (listes à présenter ici et montrer exemples d'images).

Il est un dernier point que je voudrai aborder avec vous aujourd'hui, et qui concerne cette fois-ci les logiques actuelles qui guident de récents travaux de recherche à l'INA, et qui sont relatives aux **extensions numériques** qui caractérisent les contenus télévisuels aujourd'hui.

Travailler sur les archives web de l'INA

Dans la continuité de ces travaux, j'ai pu mener en tant que **chercheuse associée** à l'Institut National de l'audiovisuel entre 2020 et 2022, une recherche qui portait sur les **publications des chaînes de télévision françaises sur Twitter et**

⁵ SQL (Structured Query Language, Langage de requêtes structuré) est un langage de définition de données (LDD, ou en anglais DDL Data Definition Language), un langage de manipulation de données (LMD, ou en anglais DML, Data Manipulation Language), et un langage de contrôle de données (LCD, ou en anglais DCL, Data Control Language), pour les bases de données relationnelles.

leurs effets sur les commentaires numériques. Twitter est devenu un terrain d'étude privilégié, ce dispositif étant plus ouvert aux chercheurs et aux professionnels, en particulier via son Application Programming Interface (API).

Dans le cadre de ce travail, j'ai eu l'occasion de **tester une nouvelle méthodologie de recherche** et d'éprouver **la constitution d'un corpus à partir des archives web de l'INA** grâce à une étude de cas autour du concours de beauté **Miss France**.

Le concours Miss France, est diffusé en direct depuis plusieurs années par la chaîne TF1, il figure parmi les programmes les plus commentés de la télévision française sur ce dispositif. Ce concours de beauté a fêté en 2020 cent ans d'existence, il réunit entre **six et huit millions de téléspectateurs** par soirée, et génère des **millions de tweets souvent polémiques**.

Miss France est une **cérémonie télévisuelle** (Dayan, Katz, 1996) qui rompt avec le quotidien des téléspectateurs et répond à un déroulement ritualisé. Cela facilite ce que l'on nomme la « mise en phase spectatorielle » (Odin, 2011) qui est l'un des aspects majeurs que la **télévision dite « sociale »** valorise depuis plusieurs années : les internautes commentateurs d'une émission s'insèrent au sein de « communautés interprétatives » *ad hoc* et les pics de commentaires postés simultanément à la diffusion de certaines émissions de télévision les matérialisent.

Sur Twitter, les community managers de TF1 et de la société productrice propriétaire de la marque « Miss France » incitent les téléspectateurs au vote, et tentent, par différentes accroches en amont, pendant et en aval de la diffusion de la soirée, de susciter leur adhésion au programme et à la chaîne. Ces incitations sont d'autant plus importantes que les **enjeux financiers** sont de taille, les **votes** rapporteraient à la production environ 600 000 euros par soirée, auxquels s'ajoutent les **recettes publicitaires** liées à la diffusion du programme qui avoisinent les cinq millions d'euros, déduits des deux millions d'euros de droits de retransmission que la société TF1 reverse à la société Banijay.

L'ancienneté du programme, sa récurrence et la stratégie numérique qui l'accompagne m'offraient donc un terrain propice à **une analyse diachronique** de l'évolution des commentaires qu'il suscite. Une perspective diachronique permettait de déterminer si les thématiques des discours qui se nouent autour de cette cérémonie annuelle évoluent au fil des ans, ou si le « rituel » télévisuel (Dayan, 2000) se double, dans le cas présent, de rituels numériques.

Compte tenu de ma problématique, j'ai **constitué plusieurs corpus. L'un est un échantillon des émissions télévisuelles, l'autre un échantillon de tweets relatifs au programme.** Pour comprendre les **choix et critères** qui ont présidé à la constitution de ces échantillons, il faut d'abord que je vous précise que pour alimenter l'archive du DL web, les équipes de l'INA font aussi des **choix**, ne pouvant tout stocker, tout documenter, elles suivent certains comptes, certains hashtags. (Montrer exemple hashtag et tweets suivis par INA) Or les critères de choix des sites, hashtags, mentions, comptes suivis laissés à **l'appréciation des équipes**, et fonction des demandes des chercheurs.

Or précisément, **Miss France ne faisait à l'époque pas partie des hashtags** suivis systématiquement par l'INA ! Seul le hashtag MissFrance2020 avait été suivi à partir de la mi-décembre 2019, sachant que j'ai démarré mon étude en janvier 2020. Le hashtag est suivi à ma demande depuis novembre 2021. Pour autant, l'archive web de l'INA contient des messages qui contiennent des références à Miss France (montrer image et période). Partant de cette observation dans la base archive, j'ai choisi de constituer mon corpus sur les 5 dernières années de diffusion de l'émission (j'ai commencé mon travail en 2020). Soit 2014-2019 (ce qui correspond à Miss France 2015- Miss France 2020). Les chaînes TV investissent Twitter à partir de 2009, mais explosion des publications se fait à partir de 2014 (accords commerciaux, débuts de publicités ciblées et de partenariats forts de mesure de l'audience sociale).

Quand j'ai commencé à vouloir travailler sur Miss France, d'abord je suis passée par les interfaces de consultation des archives de l'INA (hyperbase pour les émissions TV), et **l'interface de consultation du DL web**, en particulier celle liée à Twitter (montrer ce que l'on peut faire et ne pas faire).

Le principal inconvénient de cette interface, à l'époque, c'est **l'impossibilité d'en extraire directement les tweets**, d'être obligé de travailler avec les **outils statistiques proposés** qui ne permettent pas de faire des analyses lexicales par exemple, et c'est aussi le cas aujourd'hui sur les **nouveaux postes innovations de l'INA** qui proposent des interfaces de consultation différentes **Webmedia** (montrer vidéo) : on peut suivre côte à côte la timeline des tweets et le déroulé du programme. Mais si l'on veut faire une **analyse de contenu des messages** ; il faut passer par les **ingénieurs et**

documentalistes de l'INA, seuls habilités à proposer une **extraction de ces tweets (sous licence d'utilisation)**. Ce que j'ai pu obtenir dans le cadre de mon statut de chercheuse associée. Montrer exemple d'extraction

Compte tenu du fait que le hashtag Miss France n'était pas suivi par les équipes de l'INA, de quoi serait constitué mon export Miss France ? Les **exports étaient parcellaires** s'agissant de Miss France, mais représentatifs de ce qui a été échangé autour de ce sujet dans les publications Twitter liées à l'audiovisuel, car l'INA suit un grand nombre de hashtags liés aux émissions radio et tv, et suit de façon systématique tous les messages qui mentionnent TF1, ou qui émanent de TF1 ou des sociétés de production TV (exhaustivité). Donc nous avons en quelque sorte reconstitué un corpus, qui a comme caractéristique de contenir de façon exhaustive bp de messages liés à la chaîne TF1, au programme Miss France, à ses animateurs, à l'univers TV. **Messages qui sont donc surreprésentés dans ce corpus.**

Notre corpus concerne les **tweets récoltés le jour de la diffusion de la cérémonie Miss France**, de 05 heures le matin jusqu'à 05 heures le lendemain matin (Pic de commentaires).

Autrement dit, si ce corpus comporte bien des tweets postés par de simples téléspectateurs, n'ont été archivés que ceux qui comportent une mention @TF1, un #TF1, ou qui retweetent des publications issues de comptes médias avec le mot clé « Miss France ». Or les internautes développent des tactiques de commentaires et de partage qui les détournent parfois des hashtags officiels (Mercier, 2017). **Ce corpus n'a donc pas de caractère exhaustif ou représentatif.** En revanche, il nous offre la possibilité d'observer des **tendances** en termes de contenus, de retweets et de tonalité de discours, et surtout d'étendre l'observation sur plusieurs années. Mais la **sur représentativité** des comptes médias par rapport aux comptes des internautes a eu des conséquences dans mes analyses (voir article Réseaux). Vous pourriez me rétorquer qu'à l'époque, ce qui n'est plus possible aujourd'hui, j'aurais pu passer par un informaticien qui aurait eu la possibilité **d'aspirer les tweets** en question, mais je voulais à tout prix **tester les potentialités**, et en l'occurrence les limites de l'archive web de l'INA, c'est d'ailleurs l'une des conditions du statut de chercheur associé.

L'INA m'a donc permis d'accéder à ces tweets sous la forme d'extractions contenant **certaines métadonnées**, comme l'URL du tweet qui permet, sauf pour certains contenus ayant été supprimés par leurs auteurs, d'accéder au message dans son interface originelle.

Dans un premier temps, j'ai directement travaillé à partir des extractions fournies par l'INA. À ce stade, différents tris ont facilité le repérage des contenus mis en avant par les community managers de la chaîne TF1 ou ceux des comptes associés. J'ai étudié les discours d'escorte qui accompagnent ces publications. Certaines métadonnées, comme la date et l'heure du tweet, m'ont permis de retracer l'évolution de ces publications pendant les 24 heures qui entourent la diffusion du programme. Enfin, des tris croisés m'ont aidée à déterminer quels étaient, dans notre corpus, les utilisateurs les plus actifs.

Dans un second temps, j'ai choisi d'utiliser le logiciel IraMuteq, qui permet de produire des analyses statistiques sur des textes, de contrôler le paramétrage de leur traitement et de les représenter sous la forme de graphiques. Ces statistiques descriptives m'ont servi à repérer des régularités lexicales et relationnelles au sein du corpus. J'ai d'abord utilisé une fonctionnalité qui s'appuie sur la méthode Reinert (Reinert, 1983) : le logiciel réunit les tweets qui contiennent les mêmes segments de textes. Il construit ensuite des classes qui sont décrites à partir d'un lexique « (...) constitué des mots qui sont significativement surreprésentés dans la classe si on la compare à l'ensemble des autres classes (sur la base d'un Chi2). » (Ratinaud et al., 2019) Si cette méthode automatisée permet une analyse de corpus massifs, elle n'est toutefois pas exempte de biais ; le paramétrage du logiciel et la constitution des classes étant soumis aux choix du chercheur.

Pour chaque sous-corpus par année, j'ai supprimé les messages retweetés qui constituent un volume important des messages récoltés (annexe II). J'ai donc analysé les corpus par année, avec puis sans retweets. Les corpus avec retweets permettent de repérer les thématiques liées au programme qui ont le plus circulé sur Twitter, les corpus sans retweets permettent de rendre saillants des contenus masqués par les messages très retweetés (annexe III).

La lecture des segments de textes caractéristiques et le passage par leur URL m'ont permis d'identifier les thématiques de chaque classe lexicale par sous corpus, que j'ai ensuite regroupées sous forme de tableaux en fonction

de leurs modes de production de sens dominants, afin de pouvoir les comparer (voir à titre d'exemple un extrait de ces tableaux reproduit en annexe IV). **Ce n'est que l'ors de cette étape, le passage par l'URL, que l'on retrouve l'image audiovisuelle contenue dans le tweet.** Iramuteq m'a servi à **constituer un sous-corpus d'analyse, et c'est ce sous corpus constitué à partir des segments de texte caractéristiques que j'ai analysé dans l'interface originale de twitter pour y retrouver une matière plurisémiotique et l'environnement natif.**

Cette **méthodologie**, je l'ai **réutilisée** dans un travail sur la communication des chaînes de télévision sur certains RSN (Twitter, Instagram et Facebook). Dans ce travail, je m'intéressai notamment aux **métriques** liées aux publications des chaînes de télévision sur Twitter, **métriques que l'INA n'archive pas** et qu'il m'a donc également fallu aussi reconstituer en créant des **sous corpus que j'ai qualifiés manuellement** sur la base des extractions fournies par l'INA (montrer exemple). Je pourrai vous expliquer comment lors des échanges, mais ce que je veux souligner ici c'est que les **méthodes de collecte automatisées** auxquelles nous avons accès aujourd'hui impliquent systématiquement de retravailler les données, de les nettoyer, de les rendre homogènes ou de les compléter.

D'ailleurs, parmi les récentes possibilités ouvertes par l'INA aux chercheurs ces temps-ci existe celle d'accéder désormais aux **retranscriptions intégrales des contenus** diffusés ces 5 dernières années sur les CIC et dans les JT. Ceci grâce à l'IA. Mais ces retranscriptions textuelles sont incomplètes, voire instables, notamment dans le repérage des locuteurs.

Finalement, l'IA qui n'est ici qu'une étape vers **l'analyse qualitative du programme (à mon sens)**, qui si l'on veut, ce qui est mon cas en tant que sémiologue, **étudier de façon précise** la prise de parole des intervenants dans un programme et surtout sa mise en scène, doit passer obligatoirement par son **visionnage**. D'autres outils d'IA sont proposés à l'INA dont on parlera demain peut-être, comme des logiciels de reconnaissance faciale, etc. Mais à ce jour, rien ne remplace l'analyse qualitative des images télévisuelles, pour en saisir toute la **complexité sémiotique et communicationnelle**. Et d'ailleurs, ce sont les chercheurs qui permettent aussi d'entraîner les IA développées par les ingénieurs de l'INA (ex : trombino). Je suis à cet égard relativement dubitative face à cette tendance quantitativiste qui semble aujourd'hui s'emparer des équipes de recherche de l'INA (INALab), car on voit bien ici les **limites de ces données quantitatives**, et surtout en ce qui concerne mon expérience personnelle, **le temps passé à nettoyer et catégoriser** ou donner une forme à ces données pour en faire des corpus est du temps en moins pour leur analyse qualitative.

Conclusion

Ces exemples montrent aussi que **les logiques d'indexation pratiquées** par les documentalistes de l'INA constituent un **enjeu en soi pour** quiconque s'intéresse à la transmission des connaissances, et notamment des connaissances liées aux **images de télévision** : « (...) dans ce cas la fonction de médiation implique toujours une compréhension des règles de constitution de ladite connaissance, en relation avec l'explicitation de ce qui, dans la connaissance ou dans la façon de présenter la connaissance, est le plus propre à structurer la compétence, c'est-à-dire à la construire et à la faire évoluer : qu'est-ce qu'il faut apprendre – et comment – de la notion de « genre télévisuel », par exemple, à des documentalistes qui ont et auront à rédiger des notices d'indexation des émissions actuelles – et à venir – de radio ou de télévision ? Et en retour – c'est un retour qui est souvent oublié –, qu'est-ce qui, dans les problèmes d'indexation rencontrés, réinterroge les mises au point théoriques ?⁶ ». **Le travail de l'analyste de la télévision ne s'arrête pas aux images et aux sons**, il doit également convoquer une **multitude de documents** qui accompagnent les supports audiovisuels, et que **Geneviève Jacquinot** appelle, paraphrasant Christian Metz, « les apprêts et les après de la télévision. » Or l'usage de ces documents, qui conditionnent la constitution du corpus, suppose au préalable d'expliquer et de connaître les présupposés et les logiques qui président à leur élaboration.

Enfin, **les corpus que je vous ai présenté sont plutôt massifs**, et lorsqu'un corpus prend ce type de taille critique, cela implique de **multiplier les angles d'attaque**, pour littéralement travailler au corps le corpus, en adoptant des **approches**

⁶ . JACQUINOT, « Repenser l'enseignement de la télévision », in F. JOST et J. BOURDON (Dirs.), Penser la télévision, pp. 313-314

macroscopiques (statistiques, classes lexicales) ; et **micro-scopiques** (analyse des segments en contexte), et partant de là, proposer des **codages** qui ont l'inconvénient de réduire, catégoriser, figer le réel, mais aussi **faciliter l'appréhension et la compréhension** de ce corpus (étymologie du terme). Autrement dit, la constitution du corpus dans mon cas relève souvent d'un travail en **extension**, auquel se confronte ensuite un travail de **compréhension** : or comprendre : *cum* et *prendere*, « saisir ensemble », saisir par l'intelligence et la pensée. Travailler un corpus suppose donc selon moi les mêmes types de mouvements que ceux qui président au travail autour d'un concept, comme le souligne Canguilhem.