Sken&graphie

Coulisses des arts du spectacle et des scènes émergentes

2 | Automne 2014

Paroles de danseurs et de chorégraphes

Les archives du web à la BnF

Un formidable gisement à venir exploiter

Cécile Obligi

Édition électronique

URL : <http://journals.openedition.org/skenographie/1137>

DOI : 10.4000/skenographie.1137

ISSN : 2553-1875

Éditeur

Presses universitaires de Franche-Comté

Édition imprimée

Date de publication : 1 novembre 2014

Pagination : 95-102

ISBN : 978-2-84867-495-7

ISSN : 1150-594X

Référence électronique

Cécile Obligi, « Les archives du web à la BnF », *Sken&graphie* [En ligne], 2 | Automne 2014, mis en ligne le 30 novembre 2016, consulté le 03 mai 2019. URL : <http://journals.openedition.org/skenographie/1137> ; DOI : 10.4000/skenographie.1137

Presses universitaires de Franche-Comté

LES ARCHIVES DU WEB À LA BNF : UN FORMIDABLE GISEMENT À VENIR EXPLOITER

CÉCILE OBLIGI

Qu'y a-t-il de plus éphémère qu'un spectacle ? Sans doute tout ce qui est diffusé sur le web autour de ce spectacle. La documentation produite à l'occasion d'une représentation, que ce soit pour sa production, son déroulement, sa promotion ou encore sa critique, se diffuse depuis quelques années sur un medium supplémentaire, le web. Jusqu'à présent, la recherche en arts du spectacle s'appuyait sur un certain nombre de documents qui, par recoupements, permettaient de saisir la mémoire de ce qui est par définition éphémère et unique, le spectacle. Les manuscrits de l'auteur de la pièce, les notes de mise en scène, les croquis préparatoires aux costumes et aux décors, le programme, l'affiche, les photographies de scène (ou auparavant les gravures), le dossier de presse remis aux différents publics (scolaire, journalistique...), la critique paraissant dans les journaux sur le spectacle et, depuis quelques décennies, l'éventuelle captation audiovisuelle, dessinaient peu à peu pour le chercheur la silhouette de l'objet disparu. Depuis une vingtaine d'années, il faut compter avec un nouveau medium, le web. D'abord peu utilisé (on se souvient de l'indigence et du peu de fraîcheur des informations que l'on trouvait sur les premiers sites web), il est peu à peu devenu un vecteur de communication essentiel dont plus personne ne se passe : la plus petite compagnie amatrice débutante a son site web, les grands théâtres mettent à disposition de leur public une riche documentation en ligne. Le web est devenu une source à part entière, contenant bien sûr des informations disponibles par ailleurs sur un autre support, mais aussi, et c'est en partie cela qui le rend précieux, des informations introuvables ailleurs. Ne prenons qu'un seul exemple : comment retrouver la

présentation de la saison du Théâtre de la colline si l'on n'y pas assisté ? Sur le site web du théâtre et uniquement sur son site web.

Une difficulté corse rapidement le travail du chercheur : la volatilité des informations produites. Cela tient en partie à la temporalité du web, royaume de l'immédiateté et de l'éphémère par excellence, mais aussi au fait que ce médium ne bénéficie pas, en termes d'archivage, d'une attention plus grande que ses prédécesseurs matériels (prospectus,...) de la part des producteurs : depuis les débuts du web, une grande partie de la masse produite a disparu.

DADVSI, Heritrix, la BnF : une loi, un robot, une collection

Où le législateur a instauré, par la loi dite DADVSI¹ (Droit d'auteur et droits voisins dans la société de l'information) du 1^{er} août 2006, précisée par un décret² d'application de décembre 2011, un dépôt légal du web. Prolongeant le dépôt légal auquel sont déjà soumis les imprimés, mais aussi les estampes, cartes, plans, partitions musicales, photographies, phonogrammes, affiches, vidéogrammes et documents multimédias, la loi de 2006 l'étend au web : « Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique. » La tâche est partagée entre l'Institut national de l'audiovisuel (Ina), qui collecte les sites du domaine de la communication audiovisuelle (notamment ceux qui concernent la radio et la télévision) et la Bibliothèque nationale de France (BnF), qui se charge de tous les autres sites, quantitativement largement majoritaires.

Le législateur français a fait en la matière des choix qui contribuent à donner au résultat, que l'on nomme « les archives du web », une coloration particulière. Le dépôt légal du web, contrairement à celui des autres supports, n'implique pas de démarche particulière de la part des producteurs d'un site : ce n'est pas le producteur qui dépose à la BnF, c'est la BnF qui vient archiver un site web. On demande seulement au producteur de fournir les informations techniques nécessaires en cas de difficulté à capter son site. Il s'agit donc d'une collecte davantage que d'un dépôt. Inversement, l'autorisation du producteur d'un site n'est pas nécessaire, la loi permet à la BnF de procéder à la collecte sans faire de demande

¹ Loi n° 2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information.

² Décret n° 2011-1904 du 19 décembre 2011 relatif au dépôt légal.

ou d'information préalable. Ces choix n'allaient pas de soi ; d'autres pays ont pris des orientations différentes.

Pour collecter les sites web, la BnF utilise un robot nommé *Heritrix*. Ce robot, qui est un logiciel libre développé au sein d'un organisme international de préservation du web, l'*International Internet Preservation Consortium* (IIPC), moissonne le web en partant d'une liste d'adresses web (URL) fournie entre autres par l'AFNIC³. Il extrait les liens dans les pages de code d'un site et les suit comme le ferait un internaute, mais de manière automatique. Il copie ensuite les informations ainsi collectées. Plusieurs types de collectes ont lieu chaque année. Une collecte dite large parcourt une fois par an le web en surface : elle archive ainsi un grand nombre de sites, mais peu profondément. Donc beaucoup de sites sont captés mais pas en totalité. C'est ce que l'on pourrait comparer, tant en volumétrie qu'en qualité de contenu, aux « camions » de livres qui arrivent par le dépôt légal chaque année. De même que l'on ne fait pas entrer dans les collections de la BnF un roman en fonction de sa valeur littéraire, un site web n'entre pas dans les collectes du web en raison de sa qualité. L'objectif de l'exhaustivité, déjà difficile à remplir pour les autres supports, est impossible à atteindre dans le cas du web. Plus modestement, l'ambition est de collecter au mieux le domaine français, évalué en 2012 à sept millions de sites. De même que le dépôt légal est complété pour les autres supports par des acquisitions, fruit d'un choix des bibliothécaires, la collecte large est complétée par des collectes de sites choisis. Un ensemble de 80 bibliothécaires sélectionne un certain nombre de sites qu'il leur semble important d'archiver plus profondément à une fréquence pouvant aller de une fois par jour à une fois par an. Pour le domaine des arts du spectacle, quatre agents du département des Arts du spectacle ont choisi un peu moins de 200 sites et en surveillent la collecte.

Les collectes ont commencé à la BnF en 2005, mais le web est bien sûr antérieur. Pour couvrir également la période où ces collectes n'existaient pas encore (1996-2005), la BnF a acquis – en antiquariat pourrait-on dire par analogie avec les autres supports – auprès d'une fondation américaine, Internet Archive⁴, les sites en .fr qu'elle avait archivés. La collection ainsi obtenue atteint des dimensions vertigineuses : 18.8 milliards de fichiers soit 370 téraoctets de données au début de l'année 2013 (pour la collection dans son ensemble, de 1996 à 2012 inclus). Ces archives sont consultables uniquement dans les enceintes de la BnF, par des

³ Association française pour le nommage Internet en coopération.

⁴ L'adresse du site de la fondation est la suivante : <http://archive.org/>

chercheurs accrédités. Elles ne sont pas consultables dans les salles ouvertes à tous moyennant l'acquisition d'une carte d'entrée, c'est-à-dire pas dans l'espace nommé « Haut-de-jardin », mais seulement dans les salles dites de recherche donc celles des sites François-Mitterrand (« Rez-de-jardin »), Richelieu, Louvois, à la Bibliothèque-musée de l'Opéra, mais aussi à la Maison Jean Vilar, antenne du département des Arts du spectacle à Avignon. En revanche, la loi n'autorise pas la communication des archives dans d'autres bibliothèques (exceptées les autres bibliothèques attributaires d'un exemplaire du dépôt légal, qui participent parfois à la sélection des sites pour certaines collectes) et encore moins la diffusion en ligne. C'est là encore un choix français qui n'était pas évident : le site de la fondation Internet Archive est librement accessible sur le web. La protection du droit d'auteur et du droit à la vie privée, ainsi que l'impossibilité de retirer un site en particulier des archives du web à la BnF expliquent ce choix (Internet Archive en revanche est en mesure de soustraire un site de la consultation en ligne sur simple demande).

Deux modes de sonde pour la collecte en arts du spectacle

Pour les arts du spectacle, les différents types de collectes (larges et profondes) documentent utilement les différentes formes de spectacle contemporain, autant que la production scientifique sur le sujet. Plusieurs types de sites peuvent intéresser le chercheur en arts du spectacle : les sites institutionnels de théâtres, les sites de compagnies ou d'individus (metteurs en scène, dramaturges, musiciens, etc.), les sites consacrés à l'actualité du spectacle (du site de renseignement généraliste au blog d'amateur passionné), les sites consacrés à l'étude d'un courant, d'un auteur ou d'une époque, les revues en ligne, les sites mettant en ligne des ensembles plus ou moins volumineux de sources (textes de pièces, transcriptions de manuscrits, etc...), les portails plus ou moins éphémères, plus ou moins fournis, plus ou moins scientifiques, et l'on en oublie sans doute. La masse d'information produite est telle que le modeste objectif ne peut être que d'essayer d'en collecter un pourcentage qu'on espère représentatif. Que collecte-t-on donc concernant les arts du spectacle ?

Chaque jour a lieu la collecte d'une centaine de sites dit d'actualité, c'est-à-dire des grands quotidiens nationaux et régionaux (<http://www.lemonde.fr> ou <http://www.leprogres.fr> par exemple), ainsi que des *pure players*, les media d'information qui n'ont pas de diffusion ailleurs que sur le web (<http://www.mediapart.fr> par exemple). Cette collecte permet de couvrir de

manière satisfaisante l'actualité du spectacle. En effet, depuis la page d'accueil, le robot collecte très bien les pages spécifiquement dédiées à la culture en général ou à la scène en particulier (<http://www.telerama.fr/scenes/> par exemple). Par ce biais, un chercheur repère aisément les spectacles ayant fait parler d'eux pendant des mois, tout autant que les succès éphémères et oubliés, ou encore les petites productions passées inaperçues. La collecte annuelle large, quant à elle, fournit un bon instantané de la création contemporaine : de nombreux sites disparaissent chaque année, que les compagnies ou individus qui les ont créées aient périclité ou poursuivi sous une forme ou un autre nom, l'archive est précieuse. Ces deux types de collecte, qui descendent à environ un, deux voire quatre clics de profondeur à l'intérieur du site, constituent déjà un ensemble intéressant. En effet, l'on peut constater, avec quelques années de recul, que les sites qui concernent les arts du spectacle sont d'une taille réduite, en général très inférieure à 10 000 URL par site ; ils sont donc relativement bien captés par la collecte large. On peut par conséquent considérer qu'on dispose pour le domaine des arts du spectacle d'un échantillon significatif de ce qui est produit sur le web français.

Les collectes ciblées des sites sélectionnés par des agents du département des Arts du spectacle ont pour objectif d'enrichir ces archives. L'accent a d'abord porté au début de la participation du département (2005-2008) sur des sites institutionnels de théâtres, en s'appuyant sur les outils de travail des agents du département qui travaillaient au secteur « Actualité ». Ce secteur collecte les affiches, programmes, dossiers et revues de presse en sollicitant régulièrement les théâtres français pour obtenir des documents. Il dispose pour ce faire de listes de théâtres qu'il maintient à jour et qui pouvaient être réutilisées. C'est à partir de cette base qu'a été faite la première sélection de sites en 2005, très centrée sur les sites institutionnels de lieux de spectacle, et portant principalement sur un art du spectacle en particulier, le théâtre. L'idée était à ce moment-là de réussir à capter par ce biais des informations qui n'arrivaient que peu ou pas sur support traditionnel, il s'agissait donc de documenter par le web l'actualité de théâtres pour lesquels le département était pauvre en programmes et dossiers de presse. Peu à peu, la sélection s'est élargie de manière à recouper davantage les autres spécialités du département : marionnette, cirque, mime, music-hall, théâtre de rue, danse et opéra sont maintenant représentés. Il est d'autre part apparu nécessaire de porter une attention particulière à la critique théâtrale, florissante sur le web. Une sélection de blogs de critique de spectacle influents a été ajoutée. Des sites d'information sans lien avec une structure théâtrale, des sites mettant en ligne des sources et, en général, des sites qui n'étaient

pas la transposition ou la continuité de ce qui se produisait sur un autre support, sont venus enrichir la sélection. Enfin, en 2013, un nouvel ensemble a été ajouté : il s'agit d'une douzaine de sites web en lien avec des fonds d'archives conservés au département des Arts du spectacle. Ainsi, les archives du Théâtre du Soleil sont conservées au département et le site de la compagnie, <http://www.theatre-du-soleil.fr>, est collecté en profondeur par le robot une fois par an. La sélection, même si elle tente de se fixer des lignes de conduite les plus objectives possibles, reste très partielle et très subjective au regard de la masse des sites disponibles.

Une manne considérable pour les chercheurs à la BnF

Les archives ainsi obtenues, qui couvrent maintenant une période de plus de quinze ans, offrent un beau matériau prêt à être sondé par les chercheurs. Du fait de leur ancienneté – très relative au regard de l'activité séculaire de la BnF, mais importante à l'échelle du web –, elles recèlent un certain nombre d'informations devenues introuvables par ailleurs. Pour une part, comme l'on s'y attend, le web est un simple doublon de documents produits par ailleurs sur un support matériel : pour prendre un exemple extrême, il arrive que le pdf ayant servi à imprimer un programme soit disponible tel quel sur le site du théâtre ou de la compagnie. D'autres fois cependant, le web se substitue au support matériel : la revue de la Société des auteurs compositeurs dramatiques (SACD), les Actes du théâtre, n'a plus de version papier depuis 2007, mais elle est désormais disponible sur le web. Enfin, et c'est sans doute le plus intéressant, le web est devenu un medium à part entière avec sa logique propre et une production complètement indisponible ailleurs. Les facilités d'utilisation qu'il offre permettent de mettre en ligne ce que l'on ne pouvait imaginer diffuser largement avant : vidéos, nombreux articles, biographies, compléments. Plutôt que les deux ou trois photographies qu'il est possible d'insérer dans un programme, l'on peut aisément mettre des albums entiers sur un site.

Dans tous les cas, doublon, mutation de support ou contenu totalement nouveau, le web, et donc les archives qui sont collectées par la BnF, permettent de multiples et riches utilisations. On peut retrouver des informations sur une représentation : lieu, date, acteurs, metteur en scène, équipe technique, y compris pour de petites productions n'ayant pas laissé de traces matérielles. On trouve aussi de nombreux articles n'ayant pas fait l'objet d'une impression : les sites regorgent

maintenant de pages détaillées sur tous les praticiens ayant une part dans la fabrication du spectacle. L'iconographie est plus volontiers fournie sur le web (car moins coûteuse à diffuser) et bien entendu les captations, interviews et toutes autres vidéos sont plus aisées à diffuser en ligne. La critique de spectacle prend une toute autre dimension sur le web : tel critique patenté a autant la parole qu'un simple amateur. Les archives du web permettent d'avoir un accès aussi à l'avis des non spécialistes qui n'hésitent pas à donner leur opinion, à « commenter » les spectacles aussi bien sous les billets des critiques installés officiels que sur leurs propres blogs. Enfin, le principe de l'hyperlien sur le web crée, sans qu'il y ait d'intervention humaine, une relation entre les documents et donc une plus grande richesse pour le lecteur.

L'information présente dans les archives du web est donc à la fois plus riche, plus complète, moins triée, plus polyphonique, plus contextualisée... mais infiniment plus difficile d'accès ! En effet, il est beaucoup plus compliqué de s'y retrouver dans cette masse que dans les archives traditionnelles. D'une part, la recherche plein texte dans les archives est peu concluante pour l'instant⁵ ; la recherche se fait principalement en tapant l'URL d'un site, ce qui suppose de savoir d'avance non seulement ce que l'on cherche mais encore l'adresse précise du site. D'autre part, l'information est diluée, tout est mis sur le même plan. C'est donc bien une manne, mais une manne dans laquelle il n'est pas facile de se repérer⁶.

Les archives du web, précieuses en elles-mêmes, se bonifient en vieillissant. Les informations captées il y a une dizaine d'années deviennent plus difficilement accessibles sur le web. La présentation de la collecte suscitait encore il y a quelques années des moues dubitatives ; la perte de toutes les données d'un site personnel ou la disparition d'un site régulièrement utilisé, tout autant que l'augmentation du volume et de la qualité des données, a achevé de convaincre les réticents. Le web constitue bien pour le spectacle, comme pour d'autres disciplines, une source à part entière avec laquelle il faut désormais compter et qu'il importe

⁵ Faute de ressources, il n'a pas encore été possible de réaliser une indexation plein texte. Mais le signalement des archives de du web est une priorité pour les années à venir.

⁶ Le département des Arts du spectacle peut toutefois fournir la liste des sites collectés à sa demande en profondeur (collecte dite « ciblée ») à toute personne qui le souhaiterait.

donc de collecter, conserver, enrichir et valoriser⁷ au même titre que les autres, de manière à constituer le matériau qui nourrira les recherches de demain.

⁷ Pour reprendre des verbes souvent cités issus du décret n° 94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France.