LA MISE EN BASE DE DONNÉES DE MATÉRIAUX DE RECHERCHE EN BOTANIQUE ET EN ÉCOLOGIE

Spécimens, données et métadonnées

Lorna Heaton, Florence Millerand

S.A.C. | « Revue d'anthropologie des connaissances »

2013/4 Vol. 7, n° 4 | pages 885 à 913

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-anthropologie-des-connaissances-2013-4-page-885.htm

La mise en base de données de matériaux de recherche en botanique et en écologie

Spécimens, données et métadonnées

Lorna HEATON
Florence MILLERAND

RÉSUMÉ

Le développement de grandes bases de données constitue l'une des évolutions contemporaines les plus marquantes dans les sciences des trente dernières années. Elles constituent de nouveaux supports de production, de représentation et de mise en relation des savoirs et des connaissances. Cet article vise à examiner l'enchaînement d'instruments et de dispositifs dans la configuration matérielle des bases de données et leurs conséquences sur les modalités de production des savoirs, dans deux contextes différents. Nous décrivons, dans le premier cas, le travail nécessaire à la numérisation d'un ensemble de planches d'un herbier destinées à alimenter une base transnationale de données en botanique. Dans le deuxième cas, nous décrivons le travail de documentation d'échantillons de plancton destinés à alimenter une base de données en écologie. Nous analysons les conséquences associées aux processus de mise en base de données sur le plan de la documentation des données, du caractère distribué de la production de la base de données, de l'évolution du statut des données et des enjeux de la mise en base de matériaux de recherche en lien avec l'importance du respect de la procédure et la fragilité des infrastructures. Nos analyses nous conduisent à penser les bases de données non pas comme une fin en soi, mais plutôt comme des éléments qui, de par leur

fixation matérielle et les possibilités d'ordonnancement qu'elles proposent, acquièrent une certaine agentivité qui participe à l'action. En ce sens, la base de données agit comme un opérateur d'organisation du travail scientifique.

Mots cles : base de données, matérialité, botanique, écologie, infrastructure, travail scientifique

INTRODUCTION

Le développement de grandes bases de données constitue l'une des évolutions contemporaines les plus marquantes dans les sciences des trente dernières années. Considérées comme un nouveau régime de publication (Hilgartner, 1995), la marque d'un changement de paradigme – en particulier en sciences de la vie (Gilbert, 1991), ou en tant qu'instruments de recherche susceptibles de réordonner – du moins partiellement – le travail scientifique (Hine, 2006), les bases de données constituent aussi et surtout de nouveaux supports de production, de représentation et de mise en relation des savoirs et des connaissances (Manovich, 2001 ; Bowker, 2000, 2006 ; Bowker & Star, 1999 ; Waterton, 2010).

Après les grandes bases et banques de données en génomique[1], plusieurs projets ont vu le jour en sciences de la nature. Ceux-ci s'accompagnent généralement de grandes promesses, dont celle de favoriser la collaboration scientifique à grande échelle *via* la publication et le partage de données de recherche. Le terme même de base de données est particulièrement évocateur ; centralisées dans une même « base », les « données » (ou matériaux de recherche) pourraient circuler librement entre les frontières institutionnelles et disciplinaires pour permettre de meilleures pratiques scientifiques au service d'une recherche plus collaborative et interdisciplinaire. Cette vision de la base de données « idéale » (Bietz & Lee, 2009) semble dominer autant chez les chercheurs (voir Costello, 2009), que chez les promoteurs des projets et les organismes subventionnaires (voir NSFCC, 2007 ; Atkins, 2003).

Cet article vise à interroger les bases de données du point de vue des dispositifs matériels qu'elles organisent – voire qu'elles réorganisent – pour envisager leur rôle dans le travail d'accumulation, de classement, de catégorisation et de configuration des matériaux sous forme d'informations et de données. Plus précisément, nous nous intéressons au processus que Latour et Woolgar (1979) ont appelé, dans un autre contexte, *l'effacement des modalités*, c'est-à-dire le processus par lequel des matériaux (une motte de terre, une plante ou du plancton capturé dans une bouteille) deviennent des *données* (une liste de chiffres dans un tableur, des points sur une courbe, la planche d'un herbier),

1 Voir Berman *et al.*, 2004 et Hilgartner, 1995.

puis des textes, et cela, en passant par une série d'opérations d'étiquetage, de catégorisation, de façonnage, de numérisation, de redistribution, etc. Notre propos porte, non pas sur les modalités de conception de l'infrastructure de base de données proprement dite[2], mais plutôt sur l'enchaînement d'instruments et de dispositifs dans la configuration des matériaux, et sur leurs conséquences sur les modalités de production des savoirs.

Dans son texte classique sur le Topofil Chaix dans la forêt de Boa Vista, Latour (1996) rend compte d'une mission en forêt amazonienne et y décortique « le travail de la référence scientifique ». Il montre comment, à travers un ensemble d'artefacts et de procédés établis, fondés sur des connaissances disciplinaires, les scientifiques arrivent à « transporter » le terrain et à le transformer en faisant correspondre les échantillons prélevés à des référents abstraits et standardisés. Nous reprenons des éléments de l'argument de Latour sur l'effacement des modalités pour décrire, dans les deux cas à l'étude, des opérations de *manipulation et de transformation* de données d'une part, et la mobilisation d'une série de *dispositifs matériels* (instruments scientifiques, support imprimé et informatique) d'autre part.

Mais nous allons plus loin dans notre argumentaire, en proposant de penser les bases de données non pas comme une finalité en soi (la production d'une entité), c'est-à-dire comme des accomplissements fixant une réalité scientifique donnée, servant de points de départ inertes à la « véritable » action des scientifiques, mais plutôt comme des éléments qui, de par leur fixation matérielle (Ashcraft, Kuhn, & Cooren, 2009) et les possibilités d'ordonnancement qu'elles proposent, acquièrent une certaine agentivité qui participe à l'action, et par le fait même, les insère dans le processus du travail scientifique. En même temps qu'elles disciplinent les données, les bases de données ne sont jamais terminées – leur fragilité, le défi de leur maintenance et de leur constante mise à jour, ne les rend pas si « immuables » que cela. La base de données apparaît comme une accumulation fragile et le fruit d'un travail considérable, comprenant plusieurs étapes intermédiaires, exigeant la collaboration d'acteurs diversifiés, distribués dans le temps et dans l'espace. Cet accent sur le travail de mise en base, sur l'interdépendance et les alignements entre les actions de ceux qui travaillent à assembler les bases de données et le produit matériel de leur travail conduit à adopter une perspective sur les infrastructures scientifiques mettant au premier plan la façon dont la matérialité contraint et discipline les possibilités d'action. Elle implique plus largement de considérer la façon dont les bases de données participent à la reconfiguration du travail scientifique, à travers une nouvelle distribution des activités et le passage à une échelle jamais envisagée auparavant[3].

2 Sur cet aspect, voir Heaton & Proulx, 2012 ; Millerand, 2012.

3 Alors que nos travaux précédents se sont concentrés sur la dimension locale de l'activité scientifique, nous souhaitons explorer l'idée d'une reconfiguration du travail scientifique dans le contexte d'une redistribution du travail entre scientifiques et non-scientifiques d'une part, et d'un changement d'échelle dans la collaboration d'autre part. Cette idée sera explorée dans le cadre d'un nouveau projet de recherche financé par le Conseil de recherche en sciences humaines du Canada qui a pour terrain les activités en biodiversité (Subvention Savoir 2013-2016).

Notre argumentaire est basé sur la comparaison de deux cas, l'un en botanique, l'autre en écologie. Nous décrivons, dans le premier cas, le travail nécessaire à la numérisation d'un ensemble de planches d'un herbier destinées à alimenter une base transnationale de données en botanique. Dans le deuxième cas, nous décrivons le travail de documentation d'échantillons de plancton recueillis dans le cadre d'une mission scientifique en mer destinés à alimenter une base de données en écologie. Le cas de la numérisation de planches d'herbiers, dans le contexte de la construction d'une base de données d'envergure internationale (la *Global Plants Initiative ou GPI)*, donne à voir la transposition en mode numérique de données scientifiques préexistant sous une autre forme. Le cas de la documentation d'échantillons de plancton donne à voir un processus d'« équipement » des données (Vinck, 2009) nécessaire à leur intégration dans une base de données fédérée à l'échelle d'une grande communauté scientifique américaine en écologie (le *Network-wide Information System* ou *NIS* du *Long-Term Ecological Research Network*). Dans les deux cas, la transformation matérielle restreint certaines possibilités de production des savoirs tout en ouvrant la porte à de nouvelles possibilités. Nous décrivons et nous analysons ce travail concret de transformation matérielle pour saisir à la fois le rôle de la matérialité des dispositifs (Denis & Pontille, 2011b ; Heaton, Millerand, & Proulx 2011) et la diversité des acteurs et des activités impliqués. L'attention à l'infrastructure sociomatérielle (Orlikowski 2007 ; Leonardi 2009, 2010) et aux alignements entre les actions de ceux qui travaillent à assembler les bases de données et le produit matériel de leur travail (Carlile *et al*., 2013 ; Orlikowski & Scott 2008 ; Leonardi & Barley 2010) peut permettre d'accéder à des acteurs ou activités qui n'apparaissent pas toujours clairement dans les discours officiels ou spontanés (Vinck, 2009). Les propriétés des objets et la façon dont on les « équipe » (Vinck, 2006) matérialisent des infrastructures invisibles faites de standards, de catégories, de conventions (Bowker & Star, 1999 ; Millerand & Bowker, 2008, 2009) qui, tout en permettant leur circulation, les contraignent inévitablement.

Nous commençons avec une brève présentation du travail de préparation nécessaire à la mise en base de données à chaque site. S'ensuit un travail spécifique de comparaison articulée autour de quatre points : l'effacement des modalités et le processus de standardisation à l'œuvre dans la documentation des données ; le caractère distribué de la production des données ; l'évolution du statut du spécimen et de la donnée ; et enfin les enjeux de la mise en base de matériaux de recherche. Dans chaque section, nous serons attentives aux points de similitude, mais également aux différences entre les deux cas.

MÉTHODOLOGIE

Notre recherche relève d'une approche qualitative inspirée des principes de la théorisation ancrée (Glaser & Strauss, 1967 ; Paillé, 1994). Nous nous

sommes penchées sur le travail pratique et concret de la fabrication des bases de données, un travail qui implique de réaliser des choix d'ordre technique, sur le plan informatique (en matière d'architecture, de standards, etc.), mais aussi d'ordre social, sur le plan scientifique (en termes de dénomination de données, de représentation de champs disciplinaires, etc.). En ceci, nous suivons le principe de « l'inversion infrastructurelle » préconisé par Bowker (1994) dans l'étude des infrastructures techniques. Dans le premier cas, nous avons observé le travail de l'équipe travaillant dans l'herbier à l'Institut de Botanique de l'Université Montpellier 2. Nous avons réalisé quatre entretiens semi-dirigés avec les personnes les plus directement impliquées dans le projet de numérisation, une observation non participante sur une période de deux semaines et l'analyse de la documentation mise à notre disposition, à laquelle s'est ajoutée l'étude de la plateforme *Global Plants Initiative*. Dans le deuxième cas, nous avons réalisé un travail d'observation, d'entretiens avec les acteurs, d'analyse de documents (documentation technique, rapports, etc.) et de dispositifs (base de données, prototypes, etc.) au sein d'une équipe[4] du Réseau en Écologie basée à l'Institut Scripps d'océanographie de San Diego en Californie. Le cas présenté ici représente une partie du travail ethnographique réalisé *in situ*, sur une période de deux ans (de 2004 à 2006). Chacune des coauteures a étudié en détail un des deux cas présentés ici.

Nous avons interrogé nos données dans une série de va-et-vient entre notre questionnement de recherche et nos deux corpus. Cette oscillation entre la conceptualisation et la concrétude des données a imprégné l'ensemble de notre analyse qualitative. Dans les deux cas, nous avons fait lire nos analyses préliminaires par quelques membres des équipes afin de recueillir leurs commentaires, notamment pour nous assurer de la justesse de nos propos sur le plan des descriptions. À la lumière des similarités et des différences présentées par nos deux cas, nous avons opté pour une analyse comparée dont cet article rend compte. L'analyse comparée a été réalisée conjointement. Notre comparaison a porté sur quatre éléments : 1) le statut du spécimen et de la donnée ; 2) le travail de documentation ; 3) les acteurs et dispositifs mis en relation dans ce processus distribué ; 4) les enjeux et les conséquences sur le « produit » qu'est la base de données.

LES DEUX CAS À L'ÉTUDE

Nous fournissons ici une description détaillée du processus de mise en base de données dans les contextes locaux. Il s'agit de tâches circonscrites, fortement cadrées par des protocoles et des pratiques bien établies. Les situations que nous avons observées étant bâties sur des infrastructures au sein desquelles

4 L'Institut Scripps d'océanographie de San Diego héberge deux sites du réseau LTER. Nous parlons ici de l'équipe du site *California Current Ecosystem*.

des standards, des catégories, des unités de mesure sont relativement stabilisés (Star & Ruhleder, 1996 ; Star & Lampland, 2008), nous ne mettons pas l'accent sur les problèmes (les erreurs ou le travail de construction en tant que tel)[5], mais plutôt sur le processus de numérisation ou de documentation dans son ensemble.

La numérisation des herbiers et le Global Plants Initiative

Inaugurée en 2004, financée par la Fondation Andrew W. Mellon, la *Global Plants Initiative* (GPI)[6] vise la création d'une ressource transnationale sur les plantes. Le projet rassemble, pour la première fois, des spécimens types provenant de nombreux herbiers à travers le monde. Au cœur de cette immense base de données unifiée se trouve un ensemble d'images numériques de très haute résolution de spécimens types[7]. Ces images sont complétées par des références, illustrations, notes de terrain, photos, etc. (voir Macklin, 2009). Le tout est consultable via le portail JSTOR Global Plants (http://plants.jstor.org/).

Le défi qui consiste à passer des « vraies » plantes contenues dans des centaines – voire des milliers – d'herbiers à des représentations numérisées standardisées a impliqué la collaboration d'équipes de travail réparties dans plus de 270 musées, universités et herbiers dans soixante-dix pays. Nous avons documenté le travail d'une équipe de numérisation, celle de l'herbier[8] de l'Université Montpellier 2 qui contient environ trois millions de spécimens de plantes séchées, ce qui fait de cet herbier le troisième en importance en France et le dixième au niveau mondial (Schäfer, 2005). L'herbier participe à la GPI depuis 2004. La responsable de l'équipe estime le nombre total de types à numériser d'ici la fin du projet (2013) entre 20 000 à 25 000, suivant le protocole établi par la GPI. Elle décrit le mandat comme suit :

> « C'est de leur fournir des images et des données associées aux images de planches qui sont des planches types, de spécimens types, voilà. Des planches de spécimens types. Donc on leur fournit des images et des données associées de tous les spécimens types… C'est un travail

5 Sur ce sujet, voir Dagiral & Peerbaye, 2013.

6 En 2013, la base contient près de 2,2 millions d'objets numériques, essentiellement des spécimens (2 millions). Entre 2010 et 2013, elle croît de près de 7 500 images par semaine. Voir le site web de la Fondation : http://www.mellon.org/. Le site web de la *Global Plants Initiative* : http://about.jstor.org/global-plants.

7 Au sein de disciplines systématiques comme la botanique, certains spécimens revêtent une importance particulière. Ainsi, un « type » n'est pas « typique », mais dans les mots d'une botaniste du projet : « c'est un spécimen de référence. Le ou les échantillons sur lesquels ont été décrites pour la première fois une espèce, une variété, voilà. C'est un échantillon en effet, mais l'échantillon de référence. C'est l'étalon pour les plantes. » Sur ce sujet, voir le code international de la nomenclature botanique : http://ibot.sav.sk/icbn/main.htm.

8 Précisons que les acteurs observés attribuent au moins trois significations au mot « herbier » : a) traité de botanique ; b) collection de plantes séchées ramassées le plus souvent par un particulier ; c) bâtiment (ou partie d'un bâtiment) contenant un ensemble important de collections particulières. Ainsi, l'herbier de l'Université Montpellier 2 contient 360 herbiers différents.

> faramineux… lorsque l'on fait une recherche biblio[graphique] et qu'on trouve un type de quelque chose, d'une espèce donnée, on le prend, il rentre dans une chaîne de données. »

L'équipe de numérisation se composait en 2011 de cinq personnes[9]. Une seule d'entre elles a une formation en botanique, mais toutes ont suivi des formations scientifiques en lien avec l'environnement et le monde naturel (biologie, écologie, espaces naturels ou agronomie). Toutes affirment avoir « une sensibilité botanique ». Bien que chacun soit en mesure de réaliser la plupart des activités reliées au projet, il existe une division du travail au sein de l'équipe, maintenue à travers certaines spécialisations à certaines étapes du travail. Ainsi, la « botaniste référente » fait la recherche de types dans l'herbier, et parfois la restauration des planches, mais ne les numérise jamais. En contraste, deux autres personnes consacrent la plupart de leur temps à la numérisation. Tous entrent les métadonnées[10]. La coordination au sein de l'équipe est grandement facilitée par l'utilisation d'un wiki. En plus d'y trouver en accès libre les manuels et guides de procédures pour réaliser toutes les étapes du travail, les employés signalent eux-mêmes l'avancement de leur travail dans un tableau de bord. Ceci donne un portrait actualisé en permanence de l'état d'avancement des travaux.

> « Quand je dois faire un petit compte, voir un peu où on en est, voir ensuite comment réorganiser l'équipe, le travail d'équipe, savoir s'il faut mettre l'accent sur une tâche plus qu'une autre […] comme chacun remplit au fur et à mesure de son travail, les trucs qui ont été faits, je vais me contenter de regarder les tableaux de bord, et je vais savoir ok, ben là, on a beaucoup d'images, mais il nous manque beaucoup de données, des choses comme ça. »

Le travail nécessaire à la production d'un fichier numérique à partir d'un spécimen type, qui implique aussi de rendre le fichier compatible avec la base de données de la GPI, requiert entre trois et six heures par spécimen. Il s'agit d'une série d'étapes correspondant à deux grandes catégories : 1) le repérage, la préparation et la numérisation des planches, et 2) la mise en relation des données et des métadonnées.

La première étape consiste à faire une recherche bibliographique pour repérer les spécimens types parmi les millions d'échantillons de l'herbier. On attribue à chacun des types une chemise de classement de couleur rouge de manière à pouvoir le distinguer des autres spécimens et le retrouver facilement, car le spécimen sera remis dans sa collection à la toute fin du processus. Une partie importante de cette première étape de *recherche de types* est le déchiffrage de l'étiquette sur la planche – en écriture cursive, souvent vieille de plus de cent ans, sur du papier jauni, et rédigé avec une encre ayant tendance

9 Voir Heaton & Proulx (2012) pour une description détaillée des caractéristiques et de l'organisation du travail de l'équipe.

10 Les métadonnées sont littéralement les données sur les données, c'est-à-dire les informations qui décrivent les données (ex. : le nom du chercheur, la date, etc.).

à s'effacer. Outre le spécimen, une page d'herbier contient plusieurs autres informations, qui lui confèrent toute sa valeur. On y trouvera généralement un numéro d'identification, le nom latin du type, ainsi que des informations concernant la date et le lieu de la « récolte », et parfois, une indication de l'habitat dans lequel la plante a été récoltée. Il s'agit donc de bien déchiffrer les étiquettes pour s'assurer qu'on est bien en présence d'un type d'une part, et pour pouvoir transcrire adéquatement toutes ces données dans la base de données d'autre part[11].

La planche doit ensuite être *restaurée*. Le travail consistant à manipuler les plantes séchées sans les abîmer peut s'avérer très délicat lorsque celles-ci ont été fixées (ou collées) ou encore lorsque la plante s'est enfoncée dans sa feuille réceptacle. Une fois dégagé de la planche d'origine, le spécimen est posé sur un fond distinct afin de rehausser la résolution de l'image. Avant la numérisation, les étiquettes d'origine seront recollées sur la planche. On produit ainsi une version plus « propre » de la planche, optimisée du point de vue de la résolution et du contraste.

> « Les types, on peut pas envoyer des images qui sont sur, enfin, du mauvais papier, sur du vieux papier parce que si on fait ça, on va avoir un très mauvais contraste au niveau de l'image entre la plante et la planche de fond donc on est obligé de restaurer, mettre sur une planche claire pour faire un beau contraste et tout ça. Voilà. Puis après qu'ils aient été restaurés, ils sont scannés… en très, très haute résolution puisqu'on se retrouve avec des images qui à la fin, pèsent de l'ordre de 150 à 200 mégaoctets donc pour donner une idée. Donc, c'est des images qui nous permettent lorsque l'on *zoome* dessus d'avoir un détail qui est de l'ordre de la loupe binoculaire. Donc ce sont ces images-là qu'on fournit ensuite aux Américains. »

La planche à numériser contiendra quatre éléments de contenu supplémentaires. Au moment de la numérisation, on ajoute une échelle de calibrage des couleurs et une règle métrique sur la page de fond. Ces deux éléments sont des mécanismes de compensation propres au format numérique, dans la mesure où l'échelle de couleurs vise à pallier la diversité des terminaux de consultation (l'affichage de couleurs pouvant varier d'un moniteur à l'autre) et la règle métrique offre un repère en cas de zoomage. Deux autres éléments permettront d'intégrer la planche dans la base de données : une bandelette spécifiant l'origine de l'herbier d'où provient la planche (au sein du répertoire de l'Université Montpellier 2) et un code-barres[12] qui contient l'identifiant unique

11 Notre description ne rend pas compte de toutes les hésitations, difficultés ou problèmes qui peuvent survenir en pratique. De plus, à la différence de Denis et Pontille (2013), par exemple, du fait du caractère stabilisé des infrastructures, nous avons observé moins de doutes et de questionnements *apparents*. Une discussion de la réduction ontologique qu'opère la consolidation de la base de données (Bowker & Star, 1999) dépasse le cadre de ce texte.

12 Le code-barres permet d'associer l'image numérisée à ses données spécifiques et métadonnées. Selon les conventions établies, l'identifiant unique d'un spécimen type commence avec l'acronyme de l'herbier (dans ce cas : « MPU »), suivi du code de la collection et son numéro d'identifiant unique. Le code complet comporte toujours neuf caractères.

du spécimen type au sein de la GPI. Le recours à un numériseur (*flat bed scanner*) monté à l'envers, le *Herbscan*, permet la numérisation sans devoir renverser les planches. La figure 1 montre un exemple de planche prête à être numérisée.

Figure 1. Une planche préparée pour la numérisation

L'autre composante du travail est *l'association des données* à la planche du type. Pour chaque type, on devra saisir, sur un formulaire, la description des informations contenues sur la planche. Il faut aussi enregistrer des métadonnées techniques associées au processus de numérisation pour chaque image : le code-barres contenant le numéro unique de l'identifiant, la résolution de l'image, la date de création du fichier électronique et, enfin, l'identification de l'institution et de la personne qui a réalisé le travail. Le contenu de *l'envoi à la GPI* comporte ainsi trois éléments : 1) l'image de la planche, 2) les données sur le spécimen, et 3) et un fichier de métadonnées.

> « Parallèlement à ça, comme tu vois ici, y'a un certain nombre de données qui y sont associées. Là, c'est un échantillon très réduit de ce que *nous* on saisit comme données. On saisit beaucoup plus de données parce que quitte à saisir des données, on va saisir toutes les données qu'on a à disposition par rapport à ce type-là [même si à la Fondation *Mellon* ils veulent quelque chose qui soit standardisé entre herbiers]. Ils veulent le minimum de métadonnées, enfin, de données *propres*. Donc, on envoie le minimum de données, point. Donc, ils veulent le collecteur, la date de collecte, la ville et le pays de collecte d'où provient le type, où est-ce qu'il est hébergé, dans quelle collection il est gardé. L'ensemble des données, c'est ça. »

La documentation des données écologiques et le Long-Term Ecological Research Network

Né au début des années 1980, le Réseau américain de recherche à long terme en écologie (*Long-Term Ecological Research Network*)[13] représente la plus grande communauté scientifique nord-américaine dans le domaine, avec plus de 2 000 chercheurs et étudiants. Le Réseau rassemble des équipes de recherche multidisciplinaires au sein de « sites » (ou stations de recherche) centrés sur l'étude d'un biome (ensemble d'écosystèmes propres à une région, par exemple un désert, un estuaire côtier, une forêt boréale, etc.)[14]. Un projet de développement d'une base de données commune à toutes les équipes a été développé dans le but de favoriser le partage et la réutilisation des données de recherche au sein du Réseau. Le *Network-wide Information System (NIS)* est né au début des années 2000, il se présente sous la forme d'un portail web offrant un accès public à l'ensemble des données de recherche produites au sein du Réseau[15].

Les jeux de données (*datasets*)[16] sont au cœur de cette grande base de données. Ils proviennent des différentes équipes du Réseau et portent sur des objets scientifiques variés (plancton, mammifère, plante, courant marin, etc.). La construction du NIS requiert que chacune des équipes documente adéquatement ses jeux de données de façon à ce que ceux-ci puissent être réutilisés par d'autres chercheurs – autres que ceux qui les ont produits. Ce travail de documentation implique des opérations de manipulation et de transformation des données, et la mobilisation d'une série de dispositifs matériels (instruments scientifiques, support imprimé et informatique). Nous décrivons le travail de documentation d'un jeu de données ayant été recueillies lors d'une mission scientifique en mer, par une équipe du Réseau basée à l'Institut Scripps d'océanographie de San Diego. Cette équipe regroupe au total une quarantaine de personnes, incluant des chercheurs, étudiants, techniciens ainsi que deux personnes responsables de la gestion des données : une gestionnaire de données et un programmeur. Dans le cas qui nous occupe, tous n'ont pas participé à la mission.

Le jeu de données en question est une mesure de biomasse à partir d'échantillons de plancton. La mesure de la biomasse (masse des organismes vivants dans un milieu donné) est à la base des recherches menées par l'équipe

13 Site Web du *Long-Term Ecological Research Network*: http://www.lternet.edu/

14 Le Réseau compte 26 sites répartis aux États-Unis, en Alaska, en Antarctique et dans plusieurs îles des Caraïbes et du Pacifique. Le Réseau regroupe plus de 60 champs disciplinaires, de la climatologie à la zoologie en passant par la biologie végétale, la géologie ou l'océanographie.

15 Le Réseau a adopté le principe de la publication et de l'accès public des données scientifiques. Cela signifie qu'à partir du moment où ils ont recueilli leurs données de recherche, les chercheurs du Réseau disposent d'un délai maximum de trois ans pour les mettre en ligne.

16 La notion de jeu de données fait référence à un ensemble de données recueillies dans le cadre d'un projet de recherche ou générées par un instrument. En pratique, l'expression peut recouvrir des réalités très différentes : la collecte ponctuelle de données dans le cadre d'une mission scientifique ou le recueil continu de données sur plusieurs dizaines d'années (ex. : la mesure d'un taux de dioxyde de carbone).

d'écologues, formés principalement en océanographie et intéressés par les effets des changements climatiques sur les écosystèmes marins. Les prélèvements se font en mer, dans le cadre de missions ponctuelles, sur plusieurs jours ou semaines, à bord d'un navire scientifique. Avant chaque mission, la gestionnaire de l'information prépare les formulaires de documentation de données sur lesquels, une fois en mer, les membres de l'équipe consigneront les informations relatives aux prélèvements et mesures effectuées. Les formulaires papier, fixés sur une tablette, seront remplis par les chercheurs, étudiants et techniciens qui se relaient à tour de rôle sur le pont pour manipuler les instruments et collecter les échantillons, et dans le mini laboratoire en cabine pour réaliser les premières analyses. La gestionnaire des données explique les détails de la collecte de données servant à mesurer la biomasse :

> « En mer, les bouteilles d'échantillonnage (*sampling bottle*) sont descendues sur le côté, de façon à ce que l'eau soit capturée à différentes profondeurs, l'eau est alors remontée dans de grandes bouteilles, de 15 litres, et puis on utilise des petites bouteilles de 200 millilitres pour prélever des sous-échantillons. Après, on prend une bouteille de chaque et on l'apporte au labo, là on a un dispositif de filtrage dans lequel on insère un filtre d'une certaine taille et on verse notre échantillon dans le filtre. Tout le plancton qui contient la chlorophylle reste pris dans le filtre. Ensuite on prend ce filtre, on le met au freezer 24 heures de façon à ce que les cellules se séparent, ensuite on prend ces échantillons gelés et on les met dans de l'acétone pour extraire la chlorophylle des cellules […] on obtient un liquide qu'on va mettre dans le fluorimètre […] ça va nous permettre de dire que la fluorescence va être proportionnelle à la chlorophylle. »

La figure 2 montre un membre de l'équipe de recherche en train de manipuler les échantillons dans le mini-laboratoire sur le navire. La tablette et l'ordinateur portable sur le bureau permettront de consigner les données.

Figure 2. Analyse et documentation des échantillons de plancton en cabine

Les données recueillies (la mesure du taux de chlorophylle) sont ensuite *enregistrées* et *documentées* de façon précise. Leur documentation commence en réalité en mer, au moment où elles sont collectées.

> « D'abord quand tu recueilles tes échantillons, tu commences à utiliser la feuille [le formulaire], et tu écris la profondeur, la station, latitude, longitude, etc. Donc les échantillons ont désormais des labels qui indiquent leur contexte de recueil. Après […] pour chaque lecture, tu écris sur la ligne appropriée sur la feuille, la fluorescence, disons la lecture se fait en unités de fluorescence. […] On écrit aussi les caractéristiques de l'instrument de mesure […], combien d'eau a été filtrée, etc. »

Documenter les données vise d'abord à enregistrer les informations essentielles qui permettront de les identifier et de les analyser. Ainsi, les premières informations consignées dans le formulaire décrivent le contexte de collecte (ex. : niveau de profondeur, station, latitude, longitude (pour la localisation géographique), etc.). Sont également inclus : la date, le nom du chercheur, le titre du projet, etc. Ces informations, appelées « métadonnées de premier niveau », servent essentiellement à identifier le jeu de données. Les informations détaillées sur les données elles-mêmes (ex. : les volumes des échantillons analysés, les caractéristiques de l'instrument de mesure, les valeurs mesurées, les techniques d'analyse, etc.) servent quant à elles à l'interprétation et à l'analyse, elles constituent les « métadonnées de deuxième niveau ». Sans ces informations détaillées, les données recueillies s'avèrent peu utiles.

Il arrive souvent que les métadonnées soient incomplètes. La somme des informations à consigner est en effet substantielle (les métadonnées représentent généralement une quantité d'informations supérieure aux données elles-mêmes) et la tâche est généralement perçue comme fastidieuse. Cependant, les membres des différentes équipes parviennent généralement à pallier les oublis et erreurs, dans la mesure où ils connaissent bien les protocoles et peuvent les reconstruire ultérieurement.

Une fois remplis, les formulaires papier sont transférés sur support informatique dans un tableur (de type Excel) et passent des mains de l'équipe de recherche à celles de la gestionnaire d'information. Ils sont envoyés par courrier électronique ou transférés via un dispositif de stockage physique (disque externe). Le fichier en main, le travail de la gestionnaire d'information consiste alors à en réviser le contenu, le corriger, le compléter et, surtout, standardiser les documentations produites.

> « Une fois qu'on a le fichier, on doit le mettre dans un formulaire « modèle », un gabarit. Parce que des fois [en mer] ils changent l'ordre des colonnes parce que c'est plus facile pour eux de les remplir, des fois ils ajoutent des colonnes, donc on doit les vérifier. Aussi, par exemple, ils vont avoir enregistré les latitudes en degrés, minutes, secondes, c'est généralement comme ça qu'on fait parce que c'est ce qui est le plus simple à comprendre. Donc nous, on doit convertir ces trois colonnes en degrés

> décimaux[17], parce que c'est comme ça que les données peuvent être *interopérables*. […] Aussi, par exemple, s'ils nous ont envoyé les données chlorophylle avec 3 points de décimales, nous on sait que c'est en fait 1 point décimal, les 3 points c'est à cause de l'instrument parce que c'est numérique, donc on vérifie quand même avec eux, mais on corrige. […]. »

La standardisation des documentations (métadonnées) commence d'abord par un travail de correction : on remet les colonnes en ordre, on ajuste les valeurs, on convient d'un nombre maximum de décimales, etc. Mais il implique aussi et surtout un travail de traduction : on convertit au format standard (ici la latitude exprimée dans le format « degrés, minutes, secondes » au format numérique) afin de rendre les données « interopérables », autrement dit aptes à circuler dans d'autres environnements.

> « Après, une fois qu'on a complété les métadonnées, on fait passer les données de notre *environnement de production* [la base de données du site à accès privé] à la sphère publique, on change les privilèges et donc là tout est public et consultable. »

La dernière étape du travail de la gestionnaire de données consiste à s'assurer que les métadonnées sont complètes, car si l'équipe qui les a collectées en mer peut pallier les éventuels oublis ou erreurs de documentation, une autre équipe du Réseau qui voudrait les utiliser, par exemple à des fins de comparaison, aura absolument besoin de connaître l'ensemble des informations relatives au jeu de données. Par exemple, elle devra pouvoir identifier le contexte de la cueillette (pour quel projet de recherche, à quelle date, par quelle équipe, etc.), la technique d'échantillonnage (des bouteilles de quelle contenance, à quels niveaux de profondeur, etc.) et d'extraction (filtrage, de quel type, taille, etc.), la méthode d'analyse (fluorescence, pigments, etc.), etc. Sans ces précieuses informations, le jeu de données ne pourra être réutilisé adéquatement. Ce n'est qu'une fois documenté en bonne et due forme que le jeu de données pourra être publié en accès public.

LES TRANSFORMATIONS ASSOCIÉES À LA MISE EN BASE DE DONNÉES DE MATÉRIAUX DE RECHERCHE

Quatre idées nous semblent pertinentes à explorer pour tenter de cerner les conséquences associées aux processus de mise en base de données évoqués

17 La latitude s'exprime plus facilement dans le format « *degrés, minutes, secondes* » *(ex. :* 45° 30'N pour Montréal) alors que le format numérique décimale permet les calculs (ex. : 45.500 pour Montréal).

ici : 1) la documentation des données en tant que processus d'effacement des modalités et de standardisation, ainsi que les enjeux de la persistance de la matérialité et de la dématérialisation, 2) le caractère distribué de la production des données, 3) l'évolution du statut des données et 4) les enjeux de la mise en base de matériaux de recherche.

Documenter les données

L'effacement des modalités et le processus de standardisation

Afin de permettre aux données de circuler au-delà de leur contexte local de production, il faut les doter d'informations supplémentaires. Vinck (2009, p. 66) définit le « travail d'équipement » comme

> l'activité collective qui consiste à s'accorder sur les éléments qu'il convient d'ajouter aux objets intermédiaires afin qu'ils s'inscrivent dans un espace d'échange entre acteurs plus ou moins hétérogènes.

Ces ensembles d'informations permettront de situer les données par rapport à une utilisation potentielle. Tant au sein de la GPI qu'au sein du NIS, documenter les données vise d'abord à enregistrer les informations essentielles qui permettront de les identifier et de les analyser. Il s'agit de décrire une production locale (d'un laboratoire ou d'un herbier) dans les termes d'un format standardisé afin que d'autres puissent les réutiliser.

On travaille à effacer les traces des contextes particuliers dans lesquels les données ont été produites pour laisser place à des données épurées, dénuées des idiosyncrasies propres aux personnes et aux cultures des laboratoires ou des herbiers. Ce travail vise à rendre les données mobiles, c'est-à-dire indépendantes des environnements dans lesquels elles ont été produites, et pourtant pleines des précieuses informations documentaires qui permettront de les resituer dans leurs contextes d'origine. Dans les mots de Latour (1996), il y aura eu un processus d'effacement des modalités.

Dans le cas de la GPI, la source est la plante. Quand elle est collée sur une planche d'herbier, elle est déjà « équipée » de certaines informations, dont l'étiquette d'origine qui permet de déterminer qu'il s'agit d'un spécimen type. Dans le cas du NIS, la source est le plancton, contenu dans une bouteille d'échantillonnage équipée d'une étiquette et d'un certain nombre d'informations sur le contexte de sa collecte. Dans les deux cas, ces informations doivent être transposées dans des formulaires. Ces formulaires jouent un rôle primordial dans le processus de standardisation. Ils contraignent et « disciplinent » la production de la base de données en formalisant les formats, en contraignant la prise de notes, en forçant les chercheurs à remplir des champs prédéfinis, etc. En ce sens, ils constituent des instruments de coordination (Heaton, Millerand, & Proulx, 2011) tout en imposant des contraintes fortes sur l'action (Timmermans & Berg, 2003).

Dans le cas de la GPI, une étape n'a pas d'équivalent dans le NIS, en l'occurrence l'étape qui consiste à documenter la production même de la version numérique de la planche. Concrètement, ces métadonnées sont enregistrées dans un formulaire qui sera associé à l'image numérisée grâce à un code-barres. Cette association est particulièrement significative étant donnée l'importance que revêt le spécimen type.

Le spécimen : les enjeux de la persistance de la matérialité et de la dématérialisation

Le spécimen et le plancton sont traités de manière très différente au sein de la GPI et du NIS. Selon Latour (1996, p. 180-182), l'échantillon présente deux caractéristiques en tant que « référence » : il sert de raccourci en considérant un brin d'herbe (ou un tube de plancton) comme le représentant de milliers d'autres et il agit comme élément de preuve. Les échantillons prélevés pendant la mission LTER devraient être représentatifs de la situation au moment de leur collecte. Dans le cas de la GPI, non seulement les planches d'herbier sont-elles représentatives de leur espèce (typiques), mais elles sont également des types, c'est-à-dire des spécimens ayant permis d'identifier l'espèce. Leur *singularité* est la raison même de leur inclusion dans la base de données. Plutôt que de chercher à effacer cette singularité, on doit la documenter et standardiser les documentations produites.

Dans les deux cas, les matériaux sont conservés, mais pas de la même façon. Les données (et métadonnées) produites par le codage, l'enregistrement et la documentation, sur support informatique, des observations écologiques sont utilisées en lieu et place des matériaux d'origine, dont une partie seulement est conservée[18]. Comme dans le cas étudié par Latour dans la forêt de Boa Vista (1996, pp. 201-202), les schémas et données calculées remplacent ici le plancton prélevé. Par contraste, les spécimens types conservent quant à eux leur statut de collections naturalistes. La numérisation des planches d'herbiers ne les retire pas des herbiers physiques, elle ne fait qu'en créer une représentation numérique, avec ses propres caractéristiques[19]. Le spécimen type n'est pas converti en signes et en abstractions (comme la motte de terre, qui par mouvement de substitution devient une couleur discrète dans un cube géométrique), il est numérisé et cette numérisation lui confère une nouvelle mobilité.

La mobilité des spécimens numérisés peut soulever de nouveaux défis. Au-delà des planches individuelles, l'ordonnancement des spécimens dans un herbier informe sur les relations qui les organisent. L'organisation nomenclaturale ou taxinomique est reflétée dans la disposition physique des spécimens dans

18 Généralement, un seul sous-échantillon est conservé par prélèvement.

19 De plus, la restauration des planches vieillies les transforme. Il ne s'agit pas d'une restauration (qui cherche à ramener à l'état d'origine), mais d'une nouvelle version de la planche, grâce à l'ajout de nouveaux éléments.

un herbier et, jusqu'à un certain point, dans la disposition des collections de plantes à l'intérieur de l'herbier (bâtiment). Ainsi, un botaniste pourra faire des déductions sur les plantes conservées dans un herbier sans même regarder les planches, simplement en se promenant dans les rayons. Quand les spécimens sont séparés de leur environnement physique, l'accès intuitif à certaines de ces relations semble perdu.

La mise en base de données en tant que processus distribué

La botanique et l'écologie partagent une tradition de gestion locale de données et de collections, par des chercheurs individuels ou des équipes restreintes (Andelman *et al*., 2004 ; Zimmerman, 2007). Or produire une base de données à l'échelle de la GPI ou du NIS n'est pas une affaire locale. Ces bases de données sont le fruit de l'implication de multiples acteurs, qui sont distribués géographiquement et qui interviennent ponctuellement pour alimenter ou modifier ces énormes collections hétérogènes.

Une question importante est celle de la coordination de la production de la base de données. Schmidt et Wagner (2004, p. 367) constatent un éventail de pratiques de coordination « génériques » dans tous les cas où de grandes collections sont bâties ou gérées de manière distribuée. Ils en identifient en particulier trois types : l'adoption de formats standardisés, de conventions de dénomination (pour faciliter l'identification) et de processus de validation. Ces trois pratiques se retrouvent dans nos deux cas[20].

Tous les sites participants à la GPI reçoivent un procédurier de 75 pages et le même modèle de numériseur, le *HerbsScan.* Le procédurier détaille le protocole de préparation et de numérisation afin que chaque image numérisée corresponde au format établi et remplisse certains critères de qualité. En outre, les équipes participantes reçoivent une formation d'une durée de cinq jours au début de leur participation au projet. Les membres du réseau LTER doivent utiliser un langage standardisé de description de métadonnées, le *Ecological Metadata Language* (EML). Ce langage standard décrit l'ensemble des informations à fournir, ainsi que la façon dont elles doivent être structurées.

Les partenaires de la GPI sont invités à développer leurs propres façons de faire, en autant qu'ils respectent le format d'exportation, la nomenclature des fichiers et la syntaxe des données et métadonnées à fournir (une vingtaine de pages du procédurier y est consacrée). De la même façon, tous les partenaires du NIS utilisent la terminologie fournie par le standard EML. Les formulaires utilisés pour documenter les données qui formeront le NIS utilisent *in fine* les noms de champs, formats, unités de mesure, etc., propres au standard. Pour

20 Conformément avec notre orientation sur le travail des acteurs, nous décrivons la distribution et la coordination du travail de mise en base de données *du point de vue local.* L'adoption du point de vue des instances centrales ou des partenaires aurait sans doute révélé des perspectives différentes sur la question.

ce faire, un guide des « bonnes pratiques » a été développé visant à faciliter la production de formulaires normalisés.

En termes de validation, avant tout envoi de « batch » (1 000 à 1 200 images et métadonnées) à la GPI, il est de la responsabilité de l'herbier de vérifier la conformité du fichier des métadonnées, notamment sur le plan du format. La GPI ne s'occupe pas du tout de la validation au niveau scientifique.

> « Ils nous font entièrement confiance. Ils peuvent pas... ils ont pas du tout de... les compétences pour ça et c'est pas leur rôle. Et puis eux, ils ont à gérer tellement un grand nombre d'images que ce serait impossible. » (Entretien avec la responsable de l'équipe)

Le portail JSTOR, grâce auquel on peut consulter la GPI, fait un contrôle de qualité sur dix pour-cent des images choisies aléatoirement. Le résultat du contrôle de qualité est publié sur une interface Web interne, et rendu accessible aux partenaires pour qu'ils puissent y corriger leurs métadonnées. Le contenu validé sera ensuite publié en ligne.

Les données du NIS sont également validées localement par les gestionnaires d'information de l'équipe productrice, qui utilisent des logiciels de validation basés sur des mécanismes d'assurance et de contrôle de qualité (*Quality Assurance Quality Control*). Ceux-ci valident ensuite les données et documentations produites avec les chercheurs de l'équipe via des interfaces Web à accès restreint, avant de les envoyer au NIS.

Dans les deux cas, la cohérence interne des bases de données dépend du respect de procédés standardisés et des conventions de nomenclature. La procédure (et l'instrumentation dans le cas de la GPI) permet de garantir les résultats. Au-delà de son utilité pour bâtir et gérer des collections de documents divers, l'importance de l'autorité procédurale en lien avec les technologies de l'information a été bien documentée dans des domaines aussi variés que la médecine (Timmermans & Berg, 1997, 2003), le fonctionnement des communautés en ligne (Levrel & Cardon, 2009), et la gestion des organisations (Pollock, Williams, & D'Adderio 2007). En termes de validation, si l'on suit bien une démarche de contrôle de la qualité dans les deux cas, il existe des différences dans l'attribution des responsabilités (qui peut se prononcer sur la qualité des données). En effet, dans le cas du NIS, tous les chercheurs membres du Réseau ont leur mot à dire sur la qualité des données produites par les différentes équipes. La validation n'est donc pas du seul ressort des sites locaux.

L'évolution du statut des données

Les données de recherche produites par les chercheurs renvoient à des manières de faire et de penser propres aux « cultures épistémiques » (Knorr-Cetina, 1999) qui caractérisent les communautés scientifiques.

La botanique a une longue histoire de botanistes indépendants, associés en réseaux lâches, chacun avec ses particularités. La disposition des éléments sur

les planches des herbiers et, jusqu'à un certain point, celle des informations elles-mêmes, peut ainsi varier énormément (Keeney, 1992). Mais le partage et la réutilisation des données y sont une pratique ancienne, marquée par une longue tradition d'échange de spécimens et de correspondance entre botanistes (Hine, 2008 ; Keeney, 1992). En fait, il était pratique courante de récolter plusieurs échantillons d'une même plante lors d'une sortie terrain, dans le but de les envoyer aux collègues pour compléter leurs herbiers individuels[21]. Par ailleurs, les herbiers constituent à la fois une archive scientifique et une matière brute pour la production de connaissances. Les plantes qu'ils contiennent sont l'aboutissement de voyages et de missions, d'échanges divers et d'apprentissages botaniques. À ce titre, ils présentent un intérêt historique de collections.

La valeur des spécimens se trouve dans leur capacité à représenter les plantes vivantes dans leur habitat naturel. Les classifications établies à partir des types se veulent non seulement un index de la collection, mais aussi une source précieuse montrant la diversité des organismes dans leur état vivant. Le spécimen est préservé dans le but de conserver autant de caractéristiques que possible pour les études futures. Les spécimens étant conservés pour leur utilité potentielle plutôt qu'en fonction d'une utilisation spécifique, prévue à l'avance, il est très important d'en produire une représentation qui soit la plus fidèle possible. Cette exigence comporte en elle-même une tension : il s'agit de produire une représentation des caractéristiques les plus importantes des spécimens tout en évitant de s'attarder sur celles qui pourraient induire en erreur (parce que pas suffisamment représentatives). Linnaeus soutenait qu'une illustration devait refléter la plante du point de vue des caractéristiques les plus importantes pour sa classification (Daston, 2004). Ainsi, il est mal vu dans certains cercles botanistes d'identifier des spécimens à partir de photographies seulement, car celles-ci incluent nécessairement les particularités du spécimen photographié, contrairement au dessin manuel qui permet de représenter un spécimen type idéal (Hine, 2008, pp. 115-116)[22]. En ce sens, la fidélité d'une image numérique seule ne suffit pas, il faut pouvoir s'appuyer sur un ensemble adéquat de métadonnées qui permettront d'interpréter adéquatement le type en question.

Comme la botanique, l'écologie est perçue comme une « petite » science ou « science artisanale » (Borgman *et al.*, 2007). À la différence d'autres disciplines habituées à collaborer au sein de grandes équipes et à partager des équipements et des données de recherche à grande échelle (la physique ou l'astronomie par exemple), les données sont collectées sur de petites surfaces et gérées localement par des chercheurs individuels ou en équipes restreintes (Andelman *et al.*, 2004 ; Zimmerman, 2007). La formation au travail de terrain constitue d'ailleurs une étape indispensable dans la formation des jeunes

21 Entretien avec un membre de l'équipe de numérisation de l'herbier de l'Université Montpellier 2.

22 Hine (2008) discute des priorités qui ont dominé la conception et l'implantation des bases de données de collections de spécimens, ainsi que des enjeux relatifs au statut des illustrations, les critères de choix de métadonnées, etc.

écologues (Roth & Bowen, 2001). En outre, la façon dont les données sont classées, étiquetées et stockées renvoie à des manières de faire propres à un chercheur, un laboratoire, voire une institution, et cela même en présence de protocoles ou taxonomies acceptés de tous.

Il reste que le partage des données a toujours été au cœur des pratiques de recherche en écologie (Zimmerman, 2008). Il est encore plus prégnant chez les écologues qui travaillent dans le domaine de l'écologie sur le long terme, où les échelles de temps des recherches vont de la décade au siècle et impliquent de travailler sur des données longitudinales, traversant plusieurs carrières de chercheurs[23]. Mais les écologues du Réseau LTER ont été habitués jusque-là à échanger leurs données essentiellement avec des pairs, généralement connus d'eux, et souvent de la main à la main. Le contact interindividuel entre chercheurs est extrêmement important, le partage de données étant aussi l'occasion d'opportunités de communication et de collaboration (Zimmerman, 2008 ; Edwards *et al*., 2011).

La production de documentations standardisées conduit à la création de représentations codifiées des données écologiques (sous la forme de métadonnées). Du point de vue du chercheur, équiper ses données de documentations détaillées et standardisées et les publier dans une grande base de données anonyme implique un changement de perspective : matériau de recherche local, entaché des manières de faire propres à son laboratoire et à son équipe, la donnée est détachée de son contexte d'origine, *nettoyée*, publiée en ligne et accessible publiquement. Il ne s'agit plus seulement d'un jeu de données utile à un projet de recherche (voire au projet de recherche d'une collègue), mais d'ensembles d'informations qu'il s'agit de décrire et de documenter dans les termes d'un format standardisé afin qu'ils puissent être réutilisés dans d'autres projets, par d'autres équipes. La documentation et la publication d'un jeu de données deviennent une *production* de recherche en tant que telle, une fin en soi.

Il serait erroné de penser que ce changement de statut se fait naturellement et sans difficulté. En pratique, la production des documentations reste une tâche laborieuse, très coûteuse en temps et pour laquelle les équipes du Réseau LTER disposent de peu de ressources[24]. Faute de pouvoir consacrer le temps et les ressources nécessaires à la production de documentation de qualité, certains préféreront ne pas publier leurs données plutôt que de mettre en ligne des fichiers incomplets, difficilement réutilisables ou qui prêtent le flanc à des interprétations erronées[25], et cela, malgré la présence d'une mention d'exonération de responsabilité sur le portail du NIS en cas de « mauvaise interprétation » des données et métadonnées.

23 Voir Hobbie *et al*., 2003.

24 Ces tâches sont considérées de seconde importance et souvent confiées aux étudiants ou personnels de recherche. Sur ce sujet, voir Millerand, 2012.

25 Entretien avec un gestionnaire d'information du Réseau LTER. Sur ce point, voir aussi Costello, 2009.

Les enjeux de la mise en base des matériaux de recherche : promesses et problèmes

La mise en relation de matériaux de recherche

En plus de rendre accessibles spécimens et données (en principe) à tous les chercheurs dans le monde, des infrastructures comme la GPI et le NIS ouvrent la voie à de nouvelles manières d'explorer les matériaux de recherche, fondées sur une *mise en relation* complexe, davantage que dans la description monotypique d'un seul spécimen ou d'un seul jeu de données n'appelant que des comparaisons simples. La mise en relation d'entités qui, jusque-là, n'avaient pas été « juxtaposées » constitue l'une des conséquences les plus importantes des processus de mise en base de données observés dans les deux cas à l'étude.

La forme numérique du spécimen type permet non plus seulement son archivage, mais également sa combinaison ou juxtaposition avec des données de toutes sortes. Outre la possibilité de comparer des plantes présentées l'une à côté de l'autre sur un moniteur, il devient possible de les examiner en relation avec d'autres matériaux (des carnets de notes par exemple) et d'autres types de données (données climatologiques, entomologiques, zoologiques, etc.). De fait, les bases de données sont souvent perçues comme un moyen de réaliser pleinement le potentiel des collections elles-mêmes (Berendsohn, 2003 cité dans Hine, 2008)[26].

En écologie, les bases de données ouvrent également la voie à la production de nouvelles connaissances en permettant la réalisation d'analyses croisées difficilement réalisables autrement. Dans les deux cas, la juxtaposition de données jusque-là jamais rapprochées offre de nouvelles possibilités en termes de manipulation et de comparaison qui peuvent aboutir à la formulation de nouveaux projets de recherche, notamment interdisciplinaires. Par exemple, les matériaux contenus dans la base de données de la GPI sont mobilisés non seulement en botanique, en biologie et en sciences de l'environnement, mais aussi en sciences de la santé, en anthropologie et en histoire de l'art. Dans le cas du NIS, un jeu de données longitudinales recueillies dans le cadre d'un projet sur la qualité de l'eau de lac est réutilisé dans le cadre d'un projet sur l'impact du changement climatique sur les écosystèmes aquatiques.

Les dimensions de comparaison et de transdisciplinarité sont au cœur des promesses des projets de grandes bases de données[27]. Mais ces nouvelles possibilités de traitement et d'analyse offrent de nouvelles manières d'explorer les données qui peuvent contribuer à leur redéfinition.

D'*entités de collections*, les spécimens deviennent des *objets de recherche* sur lesquels il est possible de faire des traitements et de conduire des analyses.

26 Ajoutons que la numérisation des spécimens rend également possible la mise en relation de botanistes, en permettant la consultation d'une même image par plusieurs botanistes éloignés, donnant ainsi à voir de nouvelles formes de travail distribué (Hine, 2013).

27 C'est l'enjeu majeur des initiatives en faveur du développement de cyberinfrastructures (de type bases de données partagées) en sciences de la nature. Voir à ce sujet NSFCC, 2007.

Auparavant associées à une équipe de recherche et à un, voire deux projets, les jeux de données écologiques désormais équipées de métadonnées standardisées, peuvent devenir des données expérimentales utilisables par d'autres, dans d'autres contextes institutionnels et disciplinaires.

Il y a de forts parallèles ici avec l'étude de Beaulieu (2004), qui retrace le passage des collections physiques de cerveaux (prélevés *post mortem* sur les dépouilles) aux bases de données d'images magnétiques (les « atlas du cerveau »). Alors que les premières sont caractérisées par leur rareté et leur matérialité physique (caractéristiques d'une collection), les secondes sont marquées par leur grand nombre et leur mobilité, permettant ainsi des traitements et des analyses comparés impossibles auparavant[28]. Enfin, le processus de recherche avec les bases de données peut s'apparenter à une logique de prospection et de découverte, davantage qu'à un processus bien ordonné de tests d'hypothèses. Il devient alors possible de réaliser des analyses dirigées par les données (*data-driven research*), sans que des hypothèses aient été définies au préalable (Beaulieu, 2004). Tout comme les « atlas du cerveau » de Beaulieu, les bases de données de la GPI et du NIS deviennent à la fois des objets épistémiques et des données de recherche expérimentales. C'est le cœur du « tournant informationnel » envisagé à l'ère des bases de données

Les défis de la maintenance et de la mise à jour

Le NIS fait l'objet de mises à jour constantes, les différentes équipes alimentant la base de données au fur et à mesure. À la différence de la GPI qui a pour vocation première d'être un portail, le NIS est d'abord la base de données du Réseau LTER, certes accessible publiquement et destiné à l'ensemble de la communauté en écologie, mais dont les premiers usagers restent les membres du Réseau. Ces derniers, autant que les responsables du NIS, en assurent les mises à jour. Le standard utilisé pour normaliser les métadonnées (standard EML) fait lui-même l'objet de nouvelles versions publiées régulièrement.

Du côté de la GPI en revanche, si elle a investi d'énormes ressources dans la création de la base de données, nous ne trouvons aucune référence à une quelconque stratégie de maintenance ou de mise à jour. Une fois terminée et mise en ligne, la base de données serait une entité stable et durable, un mobile immuable. Pourtant, dans la mesure où elle renferme des informations qui doivent refléter l'évolution constante des connaissances, les enjeux de sa mise à jour et de sa maintenance sont cruciaux (Bowker, 2006).

L'organisation d'un herbier est en évolution constante, il doit intégrer les changements dans la systématisation des connaissances botaniques :

> « Même si la plante séchée ne bouge plus, la botanique continue à faire des progrès et amène des changements de nomenclatures et /ou taxinomiques : il faut donc renommer l'échantillon (récolté et déjà nommé du nom valable

28 Les bases de données d'« atlas du cerveau » sont d'ailleurs devenues des références en neuro-informatique (comme les atlas géographiques en géographie) (Beaulieu, 2004).

> à l'époque) selon les connaissances actuelles. Une fois renommé, il faut le déplacer matériellement dans la collection à l'endroit correspondant au nouveau nom, tout en gardant une trace (*ou fantôme*) à l'ancienne place. Ce travail de vérification du nom et de reclassement est de loin la première servitude matérielle des herbiers bien avant les tâches, pourtant très nécessaires, d'entretien. » (Schäfer, 2005).

À première vue, la correction des métadonnées des spécimens types numérisés, pour rendre compte des changements de nomenclatures ou de taxonomie par exemple, pourrait se faire relativement facilement, le travail de déplacement étant facilité par le format numérique. Il s'agirait simplement de changer les valeurs ou de reclasser les spécimens (possiblement en ajoutant un champ « autrement/préalablement connu sur le nom de ») ou encore d'ajouter un lien entre l'ancienne localisation et la nouvelle, et de fournir un indice d'historisation[29]. Cependant, ce travail implique un certain nombre de ressources en termes de planification, temps et financement.

Pontille (2010) montre à quel point le travail de mise à jour d'une base de données n'est pas un processus automatique, technique et routinier, mais exige un travail d'interprétation, de vérification et de discernement. Il ne s'agit pas simplement de transférer de l'information d'un support à un autre, mais de produire des informations et/ou des liens robustes et durables pour assurer la stabilité et la fiabilité des bases de données. Ce travail complexe est souvent et paradoxalement un travail invisible (Dagiral & Peerbaye, 2012).

Si on adopte une vision des bases de données comme des entités stables et durables, il s'agit alors de les « réparer » ou de les rebâtir périodiquement, en suivant ce que Denis et Pontille (2011a) ont appelé une « logique de cure ». Il nous semble que les bases de données de nos deux cas pourraient correspondre également à une « logique de care » (Mol, 2008)[30]. Dans cette perspective, les bases de données sont aussi le résultat fragile d'un travail constant de fabrication et de mise en œuvre.

CONCLUSION

Cet article a examiné l'enchaînement d'instruments et de dispositifs dans la configuration matérielle des bases de données et leurs conséquences sur les modalités de production des savoirs, dans deux contextes différents. Nous nous sommes attardées plus particulièrement sur le travail de mise en base des données, en esquissant rapidement le traitement, en amont, de la matière brute (herbier ou échantillon de plancton) et en évoquant l'utilisation des données par d'autres, en aval, comme un potentiel. Il reste qu'il n'est pas aisé

29 Voir à ce sujet Favre, Bentayeb & Bouassaid, 2007.

30 Cet aspect fait l'objet d'un article actuellement en préparation.

de saisir précisément l'importance d'un « maillon » de la chaîne de production des connaissances scientifiques indépendamment des autres, d'autant plus que la linéarité du processus est de plus en plus mise en cause par l'évolution des pratiques scientifiques (Wouters & Schroeder, 2003 ; Wake, 2008).

Nous avons décrit et analysé le travail concret de transformation matérielle des *spécimens* en *données* d'une part, et d'équipement des *données* avec des *méta*données d'autre part. En amont de la mise en base des données, l'accumulation des matériaux de recherche (spécimens ou données) demeure une activité éminemment importante, mais la valeur de l'échantillon physique change avec le support numérique. En effet, il est fort probable que les botanistes visitent moins souvent les herbiers physiques, dans la mesure où ils ont accès aux planches numérisées. La collection des spécimens types numérisés présente par ailleurs l'avantage de ne pas se dégrader aussi rapidement que les planches physiques (même s'il existe d'autres dangers qui guettent les supports numériques). Dans le cas des données en écologie, seul un échantillon « de référence » est conservé. L'échantillon d'origine est rapidement remplacé par une mesure. Bref, le matériau physique semble revêtir moins d'importance. Son caractère « unique » est capturé une fois pour toutes, et ensuite infiniment copiable et « multipliable » (Benjamin, 1939 [2012]). Le format numérique lui confère une mobilité accrue. La valeur de l'échantillon numérique semble résider essentiellement dans ce qui l'entoure : les métadonnées qui permettent sa recontextualisation ou les autres matériaux auxquels il est juxtaposé. Cela dit, si le support numérique apparaît *a priori* « meilleur » en facilitant la conservation et la mise à jour des collections et des données, le besoin de coordonner ces mises à jour, non seulement dans la base de données, mais aussi dans les autres supports (herbiers ou bases de données locales), constitue un défi de taille.

La mise en bases de données, dans les deux cas étudiés, est un processus distribué entre différentes personnes et dans de nombreux sites, qui requiert nécessairement des activités de coordination dont il importe de conserver les traces. En plus des métadonnées sur les spécimens, il est nécessaire d'intégrer des informations sur le processus de mise en base lui-même (par exemple, le nom de la personne qui a scanné la planche). Il s'agit donc d'« équiper » la base de données elle-même pour permettre sa mobilisation dans le travail qui sera fait ultérieurement (Heaton, Millerand, & Proulx, 2011).

Nous avons voulu insister sur les qualités matérielles des bases de données tout en mettant en lumière les activités d'ordonnancement qui les constituent. Le processus de mise en bases de données peut s'apparenter à ce qu'évoque Latour (1996) sur l'utilisation du Topofil Chaix par les scientifiques spécialisés dans l'analyse des sols. En suivant l'interprétation qu'en font Denis et Pontille (2011b, p. 187) dans leur étude sur l'aménagement des espaces circulables :

> « En arpentant les lieux et en prenant des notes, on pourrait en effet considérer que […] l'expédition scientifique vise à extraire puis à transporter un morceau du monde d'un lieu à un autre par une chaîne d'inscriptions servant de garants représentatifs faciles à manipuler et à analyser. ».

Cet effacement des modalités sert à rendre les données « mobiles » (Latour & Woolgar, 1979), dénuées de leurs spécificités locales et indépendantes des environnements dans lesquels elles ont été produites. Les métadonnées agissent en tant que « mobiles immuables » (Latour, 1989) qui permettent une circulation contrôlée des objets.

Nous avons proposé de penser les bases de données non pas comme une finalité en soi (le développement un dispositif technique), mais plutôt comme des éléments qui sont partie intégrante de l'action. De par leur fixation matérielle (Ashcraft, Kuhn, & Cooren, 2009) et les possibilités d'ordonnancement qu'elles proposent, les bases de données acquièrent une certaine agentivité et participent à la reconfiguration du travail scientifique, à travers une nouvelle distribution des activités et le passage à une échelle jamais envisagée auparavant.

Dans des contextes où des données hétérogènes sont mises en relation par différents acteurs visant différentes finalités, il devient difficile de concevoir la base de données comme une entité stabilisée et technologiquement close. Dès lors que les bases de données servent de point de départ à d'autres explorations scientifiques, par exemple par d'autres équipes de chercheurs dans une logique de prospection et de découverte (*data-driven research*), nous suggérons qu'elles relèvent moins du régime de la représentation – contrairement aux inscriptions relevées par Latour – que du régime d'action : la base de données vise non seulement la représentation des spécimens et des données, mais elle sert aussi de matériau pour l'action. En ce sens, la base de données agit comme un opérateur d'organisation du travail scientifique.

Remerciements

Nous remercions nos collègues, les participants à la recherche et le Conseil de recherches en sciences humaines du Canada qui a subventionné les recherches à la base de cet article. Nous remercions aussi les évaluateurs, évaluatrices anonymes dont les commentaires critiques ont contribué à l'amélioration de ce texte.

RÉFÉRENCES

Atkins, D. E. *et al.* (2003). Revolutionizing science and engineering through cyberinfrastructure. Report of the NSF blue-ribbon advisory panel on cyberinfrastructure. National Science Fondation. http://www.communitytechnology.org/nsf_ci_report.

Andelman, S., Bowles, C., Willig, M., & Waide, R. (2004). Understanding environmental complexity through a distributed knowledge network. *Bioscience, 54*, 240-246.

Ashcraft, K. L., Kuhn, T. R., & Cooren, F. (2009). Constitutional Amendments: "Materializing" Organizational Communication. *The Academy of management annals*, *3*(1), 1-64.

Beaulieu, A. (2004). From brainbank to database: the informational turn in the study of the brain. *Studies in History and Philosophy of Science. Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 35*(2), 367-390.

Benjamin, W. (2012) *L'œuvre d'art à l'époque de sa reproductibilité technique*, nouvelle trad. Lionel Duvoy de la 4e version de l'essai (1939), Paris : Allia.

Berendsohn W. G. (2003). ENHSIN in the context of evolving global biological collections information system. In M. J. Scoble (ed.), *ENHSIN: The European Natural History Specimen Information Network* (pp. 21-32), London: Natural History Museum.

Berman, H. M., Bourne, P. E., & Westbrook, J. (2004). The Protein Data Bank: A Case Study in Management of Community Data. *Current Proteomics* (1), 49-57.

Bietz, M. J. & Lee, C. P. (2009). Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work. In E. Balka, L. Ciolfi, C. Simone, H. Tellioglu & I. Wagner (eds.), *ECSCW 2009: Proceedings of the 11th European Conference on Computer Supported Cooperative Work, 7-11 September 2009, Vienna, Austria* (pp. 243-262). London: Springer-Verlag.

Birnholtz, J. P. & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. *Paper presented at the ACM GROUP'03*, November 9-12, Sanibel Island, FL.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries. *International Journal on Digital Libraries, 7*, 17-30.

Bowker, G. C. (2006). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Bowker, G. C. (2000). Biodiversity Datadiversity, *Social Studies of Science*, 30(5) : 643-683.

Bowker, G. C. (1994). Information Mythology: The World As/Of Information. In Bud-Frierman (ed), *Information Acumen: The Understanding and Use of Knowledge in Modern Business* (pp. 21-32), London: Routledge.

Bowker, G. & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Cardon, D., & Levrel, J. (2009). La vigilance participative. Une interprétation de la gouvernance de Wikipédia. *Réseaux*, n° 154, 51-89. DOI: 10.3917/res.154.0051.

Carlile, P. R., Nicolini, D., Langley, A., & Tsoukas, H. (eds.) (2013). *How Matter Matters: Objects, Artifacts, and Materiality in Organization Studies.* Oxford : Oxford University Press.

Costello, M. J. (2009). Motivating Online Publication of Data, *BioScience*, 59(5): 418-427.

Dagiral, É. & Peerbaye, A. (2013). Voir pour savoir. Concevoir et partager des « vues » à travers une base de données biomédicales. *Réseaux*, 2(178-179), 163-196. DOI : 110.3917/res.3178-3179.0163.

Dagiral, É. & Peerbaye, A. (2012). Les mains dans les bases de données : connaître et faire reconnaître le travail invisible. *Revue d'anthropologie des connaissances*, 6(1), 191-216. DOI : 110.3917/rac.3015.0229.

Daston, L. (2004). Type specimens and scientific memory. *Critical Inquiry 31*(1): 153-182.

Denis, J. & Pontille, D. (2011a). Organization, Information, Maintenance. Proc. Conference of the International Communication Association (Organizational communication), Boston, May 26-30, 2011. consulté le 15 septembre 2013 sur http://hal.archives-ouvertes.fr/docs/00/68/50/26/PDF/JDDP_Information_Organization_Maintenance.pdf.

Denis, J. & Pontille, D. (2011b). Aménager des espaces circulables. La dynamique des déictiques. *Sciences de la Société*, n° 80, 177-192.

Denis, J. & Pontille, D. (2013). « Une infrastructure élusive » Aménagements cyclables et troubles de la description dans OpenStreetMap, *Réseaux*, 2013/2, n° 178-179, 91-125. DOI : 10.3917/res.178-179.0091.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667-690. DOI: 610.1177/0306312711413314.

Favre, C., Bentayeb, F., & Boussaid, O. (2007). Évolution de modèle dans les entrepôts de données: existant et perspectives. In *EDA* (pp. 21-36). Consulté le 14septembre 2013 sur http://eric.univ-lyon2.fr/~bentayeb/documents/version_complete/EDA07_Favre.pdf.

Gilbert W. (1991). Towards a Paradigm Shift in Biology, *Nature*, 349(6305): 99.

Glaser, B. G. & Strauss, A. (1967). *The Discovery of Grounded Theory : Strategies for qualitative research.* Chicago: Aldine Publishing Co.

Heaton, L., Millerand, F., & Proulx, S. (2011) *The Role of Collaborative Tools in Making Coordination Sustainable:* The Case of *TelaBotanica* Proc. Conference of the International Communication Association (Organizational communication), Boston, May 26-30, 2011.

Heaton, L. & Proulx, S. (2012). La construction locale d'une base transnationale de données en botanique : une mise en lumière du travail invisible des « petites mains ». *Revue de l'Anthropologie des connaissances, 6*(1), 141-162.

Hilgartner, S. (1995). Biomolecular Databases: New Communication Regimes for Biology? *Science Communication*, *17*(2), 240-263.

Hine, C. (2013). The emergent qualities of digital specimen images in biology, *Information, Communication & Society*, *16*(7), 1157-1175, DOI: 10.1080/1369118X.2012.696124

Hine, C. (2008). *Systematics as Cyberscience*. Cambridge, MA : MIT Press.

Hine, C. (2006). Databases as Scientific Instruments and their Role in the Ordering of Scientific Work, *Social Studies of Science*, 36(2), 269-298.

Hobbie, J. E., Carpenter, S. R., Grimm, N. B., Gosz, J. R., & Seastedt, T. R. (2003). The US Long Term Ecological Research Program. *BioScience*, 53(1), 21-32.

JSTOR (2012). *Explore plants.jstor.org* consulté le 15 septembre 2013 sur http://cognizant06.ithaka.org/content/jstor-plant-science.

JSTOR Plants Handbook consulté le 6 août 2013 sur *www.snsb.info/.../attach/.../JSTOR-Plants-Handbook*.

Keeney, E., (1992). *The Botanizers: Amateur Scientists in Nineteenth-Century America*. Chapel Hill: University of North Carolina Press.

Knorr-Cetina, K. (1999). *Epistemic Cultures. How the Sciences Make Knowledge*. Cambridge: Harvard University Press.

Latour, B. (1996). « Le "pédofil" de Boa Vista – montage photo-philosophique ». In B. Latour, *Petites leçons de sociologie des sciences* (pp. 171-225). Paris : La Découverte.

Latour, B. (1989). *La Science en action*. Paris : La Découverte.

Latour, B. & Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Los Angeles, CA: Sage.

Leonardi, P. M. (2009). Crossing the implementation line: The mutual constitution of technology and organizing across development and use activities. *Communication Theory*, *19,* 277-309.

Leonardi, P. M. (2010). Digital materiality? How artifacts without matter, matter *First Monday* 15 (6-7). Consulté le 5 novembre 2013 sur http://firstmonday.org/ojs/index.php/fm/article/view/3036/2567. DOI:10.5210/fm.v15i6.3036<http://opurl.bib.umontreal.ca:9003/sfx_local?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&__char_set=utf8&rft_id=info:doi/10.5210/fm.v15i6.3036&rfr_id=info:sid/libx&rft.genre=article>.

Leonardi, P. M. & Barley, S. R. (2008). Materiality and change: Challenges to building better theory about technology and organizing. *Information and Organization,18,* 159-176.

Macklin, J. (2009). *Global Plants Initiative*. Présentation à la première rencontre du réseau Canadensys, Montréal Janvier 2009. *docs.canadensys.net/2009_01_GPI_other.pdf* (consulté le 18 novembre 2010).

Manovich, L. (2001). *The language of new media*. Cambridge, MA: The MIT Press.

Millerand, F. (2012). La science en ligne : les « techniciens invisibles » dans la production des cyberinfrastructures et des savoirs scientifiques. *Revue d'Anthropologie des Connaissances, 6*(1), 163-190.

Millerand, F. & Bowker. G. C. (2009). Metadata Standards. Trajectories and Enactment in the Life of an Ontology, in M. Lampland & S. L. Star (eds.), *Standards and Their Stories* (pp. 149-165). New York: Cornell University Press.

Millerand, F., & Bowker, G. C. (2008). Metadata, trajectoires et « enaction ». *La cognition au prisme des sciences sociales*, 277-303.

Mol, A. (2008). The Logic of Care. Health and the Problem of Patient Choice. New York: Routledge.

NSFCC - National Science Foundation Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. Washington, DC: National Science Foundation.

Orlikowski, W. J. (2007). Sociomaterial practices: Exploring technology at work. *Organization Studies, 28*(9), 1435-1448.

Orlikowski, W. J., & Scott, S. V. (2008). Sociomateriality: Challenging the separation of technology, work and organization. *The Academy of Management Annals, 2*(1), 433-474.

Paillé. P. (1994). L'analyse par théorisation ancrée. *Cahiers de recherche sociologique*, 23 : 147-181.

Pollock, N., Williams, R., & D'Adderio, L. (2007). Global Software and its Provenance: Generification Work in the Production of Organizational Software Packages *Social Studies of Science* 37: 254-280.

Pontille D., (2010). Updating a Biomedical Database: Writing, Reading and Invisible Contribution, in D. Barton & U. Papen (eds.), *Anthropology of Writing: Understanding Textually-Mediated Worlds* (pp. 47-66) London: Continuum.

Roth, W. M. & Bowen, G. M. (2001). Of disciplined minds and disciplined bodies: On becoming an ecologist. *Qualitative Sociology, 24*(4), 459-481.

Schäfer. P. A. (2005). *Les Herbiers à l'Institut de Botanique* (Université Montpellier 2) http://www.tela-botanica.org/page:herbiers_institut_bota_montpellier (consulté le 21 novembre 2012).

Schmidt, K. & Wagner, I. (2004). Ordering Systems: Coordinative Practices and Artifacts in Architectural Design and Planning. *Computer Supported Cooperative Work 13*: 349-408.

Timmermans, S. & Berg, M. (2003). *The Gold Standard. The Challenges of Evidence-Based Medicine and Standardization in Health Care*. Philadelphia, Temple University Press.

Timmermans, S. & Berg, M. (1997). Standardization in Action: Achieving Local Universality Through Medical Protocols, *Social Studies of Science* 27, 273-305.

Vinck. D. (2011). Taking intermediary objects and equipping work into account in the study of engineering practices, *Engineering Studies*, 3(1), 25-44.

Vinck. D. (2009). De l'objet intermédiaire à l'objet-frontière. Vers la prise en compte du travail d'équipement. *Revue d'anthropologie des connaissances*, 3(1), 51-72.

Vinck, D. (2006). L'équipement du chercheur : comme si la technique était déterminante. *ethnographiques.org* [en ligne], 9, http://www.ethnographiques.org/documents/article/ArVinck.html.

Wake, M. (2008). Integrative biology: Science for the 21st century. *BioScience. 58*, 349-353.

Waterton, C. (2010). Experimenting with the archive: Sts-ers as analysts and co-constructors of databases and other archival forms. *Science, Technology & Human Value*s, 35(5), 645-676.

Wouters, P. & Schroder, P. (2003). Promise and practice in data sharing: Networked research and digital information. In P. Wouters & P. Schroder (eds.), *The Public Domain of Digital Research Data*. Amsterdam: Nerdi, NIWJ-KNAW.

Zimmerman, A. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data *Science, Technology, & Human Values 33*(5), 631-652.

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal of Digital Libraries, 7*(1/2), 5-16.

Lorna HEATON est professeure agrégée au Département de Communication de l'Université de Montréal (Canada). Codirectrice du Laboratoire des usages et du design des technologies d'information et de communication (LUDTIC), et membre du Centre interuniversitaire de recherche sur la science et la technologie (CIRST). Ses recherches portent sur le travail collaboratif en contexte d'utilisation des technologies, l'innovation communautaire et les relations entre concepteurs et usagers, particulièrement dans les environnements Web 2.0. Ses travaux de recherche se situent particulièrement dans le domaine scientifique, le design de l'environnement et la santé.

Affiliation :	Département de communication Université de Montréal, Pavillon Marie-Victorin CP 6129, Succ. Centre-Ville Montréal, Québec, H3C 3J7 Canada
Courriel :	lorna.heaton@umontreal.ca

Florence MILLERAND est professeure au Département de communication sociale et publique à l'Université du Québec à Montréal (UQAM). Elle codirige le Laboratoire de communication médiatisée par ordinateur (LabCMO) et elle est membre du Centre interuniversitaire de recherche sur la science et la technologie (CIRST). Ses recherches se situent au croisement des études en communication et des études en science, technologie, société (STS). Elle étudie plus particulièrement la manière dont les technologies de communication contribuent à façonner le monde contemporain, par la façon dont elles sont conçues et utilisées. Les objets auxquels elle s'intéresse vont des infrastructures d'information dans les sciences (initiatives de type e-science) aux médias sociaux en passant par les forums (en santé) et les plateformes participatives.

Affiliation : Département de communication sociale et publique
Université du Québec à Montréal (UQAM)
Case postale 8888, succursale Centre-ville
Montréal (Québec) H3C 3P8
Canada
Courriel : millerand.florence@uqam.ca

Abstract: Databasing Research Materials in Botany and Ecology: Specimens, Data and Metadata

One of the most remarkable developments in science over the past thirty years has been the development of large databases that act as new supports for the production, representation and juxtaposition of knowledge and information. This article examines the chains of instruments and devices involved in the material configuration of databases and their consequences for modes of knowledge production in two different contexts. In the first case, we describe the work required to digitize sets of plant specimens from herbaria in order to construct a transnational plant science database. In the other, we describe the work of documenting plankton specimens for an ecological database. We analyze the consequences of databasing in terms of what it means to document data, the distributed nature of database production, the evolution of the status of data itself and issues that databasing raises in terms of the importance of respecting procedure and the fragility of infrastructures. Our analyses lead us to consider databases not as an end in and of themselves, but as elements that, through their material form and the ordering possibilities they suggest, are endowed with a certain agency. In this sense, the database acts helps organize scientific work.

Keywords: database, materiality, botany, ecology, infrastructure, scientific work

Resumen: La elaboración de base de datos de materiales de investigación en botánica y en ecología : muestras, datos y metadatos

El desarrollo de las grandes bases de datos es uno de los adelantos contemporáneos más importantes en las ciencias de los últimos treinta años. Ellas consituyen nuevos soportes de producción, de representación y de la puesta en relación entre el saber y el conocimiento. Este artículo tiene por objetivo examinar la concatenación entre los instrumentos y los dispositivos en la conformación material de las bases de datos, y sus consecuencias sobre los modos de producción del saber en dos contextos diferentes. Describimos, en un primer caso, el trabajo necesario para la digitalización de un conjunto de planchas de un herbario destinadas a formar parte de una base de datos transnacional en botánica. En el segundo caso, describimos el trabajo de documentación de las muestras de plancton destinadas a alimentar una base de datos en ecología. Analizamos las consecuencias asociadas a los procesos de elaboración de las bases de datos en relación a la documentación de los datos, al carácter distribuido de la producción de base de datos, a la evolución del status de los datos y a los desafíos que pueden surgir de la conformación de bases de materiales de investigación en relación con la importancia del respeto a los procedimientos y a la fragilidad de las infraestructuras. Nuestros análisis nos conducen a considerar las bases de datos no como un fin en sí mismas, sino como elementos que, por su fijación material y por las posibilidades de ordenamiento que ellas proponen, adquieren una cierta forma de agencia que participa a la acción. En ese sentido, la base de datos actúa como un operador de la organización del trabajo científico.

Palabras claves: base de datos, materialidad, botánica, ecología, infraestructura, trabajo científico