Rapport de recherche
Contrat post-doctoral du LabEx HASTEC (PRES HESAM)
2012-2013

Nouvelles technologies et nouvelles sources du savoir : exploiter les archives numériques de conversations de réseau en corpus. L'exemple des communautés françaises de l'informatique (1983-1993)

Laboratoire de rattachement : DICEN-IDF (CNAM)

Programmes collaboratifs :

- 6. « Cultures de science et technologies des savoirs »
- 7. « Cultures savantes numériques »

Sommaire

Introduction

Comment les collectifs informaticiens (des chercheurs aux techniciens en passant par les ingénieurs) réunis en France autour du développement de l'informatique communicante et des réseaux numériques ont-ils communiqué à travers ces mêmes réseaux qu'ils ont contribué à développer ? Comment cette communication en réseau nous apprend quelque chose sur l'histoire des réseaux informatiques et numériques en France et leur place dans un dialogue Sciences, Technologies et Société ? Comment évaluer ces nouvelles sources et les mettre en corpus pour mieux les analyser, et enfin inscrire ces corpus dans une logique d'ouverture et d'interopérabilité des archives des Sciences humaines et sociales ? Voici le bouquet de questions posées par ce travail post-doctoral à un corpus de communications de type “conférence électronique” (listes et groupes de discussion médiée par Internet) témoignant des échanges des collectifs impliqués dans la mise en place de réseaux informatiques de transfert de données numériques en France et en Europe.

Nous nous sommes penchés sur des listes et groupes de discussion mises en place, maintenues et utilisées par ces collectifs au moment où les réseaux informatiques de type Internet font leur entrée en France, c'est-à-dire à partir du tout début des années 1980 et jusqu'au moment de basculement de l'accès et des usages d'Internet vers le grand public, une décennie plus tard. La délimitation temporelle du corpus s'est affinée au cours des premiers mois du travail : ayant prévu de commencer au milieu des années 1980 pour terminer au milieu des années 1990, nous avons précisé des bornes temporelles grâce à une mise au jour de repères historiques importants.

En **1983**, le Conservatoire National des Arts et Métiers, où nous avons effectué notre post-doctorat, réussit après des mois de collaborations avec des réseaux de sociabilité informelle et associatifs d'ingénieurs entre les Etats-Unis et l'Europe à créer une première connexion transnationale sur le réseau Usenet (l'infrastructure technique et administrative européenne étant EUnet, la branche française, Fnet). Ce n'est pas encore un accès « Internet » au sens strict ; le terme apparaît d'ailleurs la même année pour marquer la nouvelle possibilité du réseau pionnier Arpanet (développé depuis 1969 grâce à l'agence de recherche avancée Arpa financée par le département de la défense américain) de se connecter à d'autres réseaux informatiques (Inter-Net) grâce à un protocole dédié, le TCP-IP. Il faudra attendre **1988** pour que la première véritable liaison au TCP-IP soit établie par l'INRIA, institution proche du CNAM, mais la connexion de 1983, qui donne aux communautés d'ingénieurs informatiques français un accès aux courriel (individuel et collectif) et aux groupes de discussion en ligne, est en soi une première « *expérience d'Internet* »[1]. Si **1990** est l'année où les technologies du Web apparaissent, **1993** voit la mise à disposition des utilisateurs un logiciel de navigation dé réseau adapté aux besoins des utilisateurs les plus néophytes[2], la création d'une organisation pour les standards Web (W3C) et la reconnaissance, en termes d'économie et de politique, aux Etats-Unis, puis rapidement en Europe et en France[3] de l'importance des réseaux informatiques.

C'est donc entre ces deux dates que nous avons choisi de réunir les documents qui ont constitué notre corpus. Ce corpus est composé de deux types de matériaux principaux : les listes de discussions (en anglais *mailing-lists*, reposant sur les technologies du courrier électronique) et les groupes de

1 Selon, entre autres, Laurent Bloch, ingénieur informatique qui a travaillé au CNAM, et l'une de nos sources orales pendant cette recherche.

2 Il s'agit du navigateur Mosaïc, ancêtre de Netscape puis du Mozilla d'aujourd'hui

3 Avec respectivement le discours de lancement du programme « Information Highways » par le secrétaire d'Etat américain de l'époque Al Gore, le plan « Société de l'information » lancé par la Commission européenne et le rapport Théry « Autoroutes de l'information ».

discussion (en anglais newsgroups, relatifs au réseau Usenet), qui sont les premières formes de sociabilité médiées par les réseaux informatiques, ancêtres de nos réseaux sociaux actuels mais aussi persistants sous leur ancienne forme dans les usages actuels. Nous avons cherché, parmi ces systèmes de communication asynchrones, les discussions qui prenaient pour thèmes Internet et les technologies informatiques de réseau et impliquaient des acteurs issus de la recherche et/ou de l'ingénierie dans ces domaines, en privilégiant les acteurs issus du monde académique (laboratoires de recherche ou services de support informatique) tout en constatant que l'accès à des listes et groupes, rarement fermés, pouvait inclure des professionnels de l'informatique issus d'entreprises privées.

Les supports matériels et logiciels de l'informatique ayant considérablement évolué depuis les années 1980, qui sont la décennie où les ordinateurs personnels (ou « micro ordinateurs ») commencent à prendre le pas sur les gros systèmes que seuls les laboratoires d'université ou d'entreprise pouvaient accueillir, les archives de ces groupes et listes ont été quelque peu difficile à trouver. La plus ancienne liste que nous avons retrouvée, hébergée sur les archives de listes du site du CNRS, est celle du groupe GERET, un collectif s'intéressant aux techniques des protocoles d'Internet, initiée en 1989. Nous en avons trouvé une deuxième similaire, commençant en 1993, impliquant un groupe de travail sur les noms de domaine Internet (Domain Name System), hébergée sur les archives de listes de Renater. En ce qui concerne les groupes, leur existence remonte à 1979 et des archives importantes ont été constituées et mises en ligne par Google. Cependant, les groupes nous intéressant de prime abord n'ont été créés qu'en 1993, quand la hiérarchie de Usenet s'enrichit d'une branche francophone (notée fr.*)[4]. Ce matériau est celui sur lequel nous avons travaillé prioritairement, même si un certain nombre d'autres listes et groupes, écartées pour des raisons de cohérence ou découvertes trop tard pour les faire entrer formellement dans notre corpus, ont également servi de terrain à notre travail. Ces communications sont spécifiques à une sociabilité de réseau à une époque où ceux-ci étaient encore difficiles d'accès, pour des raisons sociales mais surtout techniques.

On se demande ainsi en quoi ces communications entre experts présentent une réflexivité (des discussions prenant pour thème l'informatique de réseau, ses techniques et usages, ayant lieu sur un support informatique de réseau) voire une récursivité (ces discussions permettant aux experts d'avancer sur le sujet, et de développer plus loin et plus efficacement ces réseaux informatiques). Comment les savoirs intellectuels et pratiques, ainsi que les croyances qui les accompagnent, ont pris forme au cœur des technologies auxquelles ils sont liés ? Comment les compétences techniques des experts se traduisent-elles dans des compétences discursives liées à l'interlocution de réseau, et vice-versa, dans l'organisation, le consensus, la décision et l'action en groupe ?

Il s'agit de prendre connaissance et d'analyser ces témoignages aussi bien au niveau du contenu qu'ils offrent sur le développement socio-technique du réseau en France, mais aussi au niveau de leur forme et de leur médiation même par la technique. Cette médiation technique a lieu en deux temps : l'échange communicationnel asynchrone par réseau (une correspondance électronique), mais aussi l'archivage de ces échanges dans des documents parvenus jusqu'à nous. Ici, la mémoire n'est pas tant celle des acteurs eux-mêmes (puisqu'ils parlent au présent dans ces échanges électroniques) mais plutôt celle de l'archive numérique native, un document produit, préservé et transmis dans le cadre des environnements informatiques et en réseau. Cette mémoire documentaire est conditionnée aussi bien par les couches basses, la plupart du temps invisibles, de la matérialité numérique (machines, protocoles, formats) que par les couches hautes, à l'interface de l'utilisation et des logiciels (applications, écrans).

Qu'est-ce que ces documents gardent, non seulement des contenus, mais aussi des traces de

4 Nous reviendrons plus en détail sur la collecte des sources.

l'usage technique impliqué dans cette communication médiée par les réseaux ? C'est la grande question réflexive qui sera posée à notre corpus. En effet, étudier l'histoire d'Internet, parmi tant d'autres sujets qui ont pu prendre forme depuis l'émergence des technologies numériques, nécessite de se pencher sur les nouvelles sources qui permettent cette étude, en sus des sources plus traditionnelles comme les archives institutionnelles et les témoignages oraux ou écrits des acteurs. Ici, ce sont des artefacts qui parlent : les contenus de ces échanges collectifs médiés par les réseaux dont il nous reste, sinon la totalité, en tout cas des morceaux, des archives partielles, et beaucoup de traces. Comment faire, à partir de dispositifs et contenus de communication symboliquement importants (des échanges collectifs des acteurs des sciences et techniques d'informatique de réseau médiés par la technologie même qu'ils contribuent à développer) des proto-documents (Pédauque, 2006) pour l'étude des technologies de réseau numériques ? Quels éléments peuvent faire source ?

On le voit, la dimension témoignage dépasse de beaucoup la question de la mémoire individuelle et/ou collective des locuteurs, et implique la mémoire de la machine elle-même et de ses usages, inscrite au cœur des conversations archivées numériquement grâce aux logiciels. Si l'on s'autorise un rapprochement sémantique, les sources historiques rencontrent les sources du code informatiques[5]. Cette double dimension de source est rendue elle aussi complexe par l'attention accordée à la matérialité de son support et cadre, d'autant plus que malgré la pérennité des données et métadonnées à travers la forme texte brut, le stockage numérique et ses évolutions rendent quasiment impossible la traçabilité d'une forme originelle et authentique.

Ce post-doctorat, s'il pose à un sujet historique une question communicationnelle (étudier les collectifs qui ont *fait* Internet en France par le biais des dispositifs de communication par lesquels ils ont pu échanger, s'organiser, débattre des problèmes techniques, sociaux, et autres relatifs aux technologies de réseau), a donc une ambition parallèle, d'ordre méthodologique et historiographique : explorer les nouvelles sources produites par la communication numérique et les méthodes pour les récupérer, les exploiter, et les partager avec le reste de la communauté des chercheurs SHS.

Notre point de départ est l'analyse des communications médiées par réseaux (ou CoMeRe)[6]. Si notre travail se situe à la croisée des Sciences de l'information et de la communication et de l'Histoire et de l'Anthropologie des techniques, la nécessité de prendre en charge un objet spécifique, communicationnel, qui est aussi un objet de langage, nous a fait nous tourner vers les Sciences du langage. Cette multiple orientation disciplinaire est caractéristique de l'approche Digital Humanities (Humanités numériques), qui déploie un questionnement SHS avec l'aide des outils de l'informatique ; dans notre cas, ce recours à l'informatique et au numérique concerne le traitement analytique des données des messages (le traitement automatique du langage pour l'analyse du contenu et du discours : aspect « linguistique » et « discursif ») et le traitement documentaire des archives des CoMeRe étudiées (leur conversion dans un format manipulable, la prise en charge des métadonnées : aspect « informationnel »). Il s'agit d'étudier ces CoMeRe comme des témoignages, voire des sources pour nourrir l'histoire d'Internet. Les documents et traces laissés par la communication médiée en réseau peuvent aider à mieux étudier cela. Celle-ci devient le prétexte d'une réflexion historiographique sur les formes de la communication en réseau comme possible supports et contenus d'analyse des rapports

5 On appelle « code source » le texte d'un programme informatique sous sa forme langage de programmation.

6 L'expression originale, de l'anglais *computer-mediated communication*, ne se traduit par bien dans le français « communication médiatisée par ordinateur », parce que le terme « médiatisation » réfère en français aux sujets qui rencontrent une grande couverture médiatique dans une logique de médias de masse. Le terme « CoMeRe », proposé par les linguistes du consortium CORPUS ECRITS au sein d'HumNum, un de nos collaborateurs pendant le post-doctorat, réinscrit l'acte communicationnel dans une logique de médiation de l'interlocution par les ordinateurs en réseau.

entre technique et société ; mais aussi d'un travail pratique, consistant à prendre en main ces documents, en faire des corpus et archives, tester l'outillage méthodologique et analytique permettant d'en faire des sources pour les SHS.

Ainsi, notre but général est de travailler, à travers des cas d'étude sur l'histoire de l'Internet en France, à légitimer des nouvelles sources issues de documents numériques natifs, dans une triple perspective.

1) Tout d'abord, il s'agit de montrer leur pertinence comme sources pour l'analyse scientifique des lieux de constitution des savoirs informatiques par la communication sur et par les réseaux informatiques. C'est le propos de notre première partie, qui s'attache à présenter les cadre théoriques orientant le regard que nous avons porté sur l'histoire d'Internet en France vue à travers les communications électroniques de leurs acteurs. Nous y présentons à titre d'exemples intéressants pour comprendre cet éclairage théorique certaines des avancées que nous avons pu mener, notamment sur le rôle de la communauté des Unixiens dans le déploiement d'infrastructures de communication via les réseaux informatiques, une histoire encore inédite.

2) Ensuite, nous souhaitons réfléchir au plan épistémologique sous-tendant le choix d'étudier ces communications et aux méthodologies nécessaires à leur appréhension. Cette deuxième partie éclaire comment ces communications deviennent des sources, à travers un parcours qui va de la médiation de la discours en réseau à sa mise en archive en passant par différentes étapes de mise en document. Nous interrogeons de manière critique le rapport de la mémoire technique générée et conditionnée par les technologies numériques à la mémoire humaine de la parole passée individuelle ou collective, ainsi que les méthodes et bonnes pratiques relevant de l'analyse de documents numériques natives tels que ceux de notre corpus

3) Enfin, nous proposons de revenir sur les travaux pratiques de mise en corpus et de préparation des documents pour l'analyse instrumentée (accompagnée par un logiciel) qui ont accompagné notre travail. En effet, les sources de la recherche sont aujourd'hui, et plus que jamais, à l'épreuve de la mise en commun des ressources et instruments de la recherche, notamment motivée par les nouvelles possibilités offertes par les technologies numériques – c'est l'approche que défendent les Digital Humanities (Humanités numériques) pour les ressources numérisées. Qu'en est-il pour des sources numériquement natives ? Nous détaillerons les étapes qui nous ont mené à travailler à la standardisation de ces documents pour prévoir leur sauvegarde, leur partage et leur exploitation analytique par la communauté des chercheurs.

En ceci, nous inscrivons ce travail dans une démarche doublement patrimoniale : travailler à légitimer et pérenniser les sources de la recherche scientifique, aussi récentes et inédites soient-elles ; mais aussi attirer l'attention sur le patrimoine des nouvelles cultures savantes de l'informatique, mémoire bien vivante, bien documentée, mais soumise à l'instabilité et à l'obsolescence de la mémoire d'Internet.

1. Communiquer le savoir sur les réseaux en faisant les réseaux : les lieux du savoir technique d'Internet

La communication à travers les réseaux numériques est considérée comme l'une des plus grandes innovations des technologies informatiques. Elle est devenue un vaste champ d'étude transdisciplinaire en particulier depuis l'avènement des médias dits sociaux. Nous nous intéressons plus particulièrement au rôle qu'elle a joué dans le déploiement même des réseaux informatiques avant qu'Internet ne devienne une technologie accessible au grand public au milieu des années 1990. En effet, la possibilité de communiquer à travers les réseaux informatiques a été une des premières applications mises en place sur ces réseaux, dès le début des années 1970. Support de la communication entre scientifiques et ingénieurs participant à la construction de ces réseaux, elle est dès le départ un moteur de développement récursif pour ce projet : c'est pour mieux communiquer entre pairs que l'on développe les réseaux, c'est pour mieux développer le réseau que l'on communique entre pairs par son moyen.

Comment le savoir circule-t-il à travers ces nouveaux canaux de communication ? On doit immédiatement ajouter à cette question la dimension du savoir-faire, puisque les experts en informatique de réseau apprennent à communiquer par la technologie en même temps qu'ils acquièrent et testent des connaissances techniques. On peut donc reformuler la question ainsi : comment les experts de l'informatique en réseau communiquent-ils un savoir-faire en construisant collaborativement le savoir lié à ce domaine ?

Nous présenterons dans cette partie une introduction générale à l'histoire de la communication médiée par les réseaux, puis les apports que les activités scientifiques du Labex HASTEC et de leurs partenaires nous ont permis de préciser théoriquement l'appréhension des différents niveaux récursifs de cette problématique.

1.1. Généalogie de la communication médiée par les réseaux : pouvoirs du courrier électronique

La communication médiée par les ordinateurs en réseaux favorisant l'échange social et humain ne va pas de soi. Les premiers travaux de mise en réseau des ordinateurs sont en effet d'abord dédiés au partage de ressources informatiques, en particulier pour le calcul partagé et l'interrogation de bases de données à distance. L'informatique communicante, à ses débuts, est donc limitée à une communication entre machines.

C'est avec l'apparition d'application dédiées sur le premier réseau de données envoyées par paquets, et non plus par circuits électroniques (les réseaux téléphoniques), que la communication humaine va pouvoir trouver sa place dans l'informatique communicante. Le projet de réseau Arpanet, mis en place par les équipes de l'agence américaine pour la « recherche en projets avancés » (ARPA, Advanced Research Project Agency) financé par la défense américaine et réunissant un réseau de laboratoires d'universités américaines en science informatique, est effectif dès 1969. En 1972, la première application de courrier électronique (ou courriel) voit le jour associé à un protocole dédié (le STMP, Simple Mail Transfer Protocol), et deux ans plus tard son utilisation constitue déjà les deux tiers du trafic du réseau Arpanet.

Si le courriel est la première histoire à succès de ce qui va devenir l'Internet[7], elle est marquée

7 Les équipes de l'ARPA mettent au point dans les années 1970 la suite protocolaire TCP-IP qui permet de connecter

par plusieurs formes de reconnaissance dans le champ de la science et de l'ingénierie informatique, selon une gradation qui va du plus pratique au plus idéologique.

Le premier degré de valorisation relève du domaine pratique. Pour travailler à améliorer un réseau encore expérimental, les différents laboratoires associés au projet, dispersés sur les côtes est et ouest des États-Unis trouvent dans le courriel un bon outil pour communiquer sur leurs travaux, à l'aide même de ce réseau qu'ils sont en train de développer. L'email est donc dès le départ un outil pour communiquer *et* pour travailler sur la communication en réseau. Les premières listes de diffusion dans les années 1970 (*mailing lists*) ouvrent la possibilité d'échanger non plus de manière interpersonnelle mais collective ; ils sont aussi un moyen pour tester l'effectivité de la propagation de l'information sur les réseaux, d'en évaluer les contours topologiques.

Il est donc assez évident qu'il est un support d'échange et de coopération apprécié, et potentiellement un support de coopération pour le développement d'Internet lui-même. Cette dimension réflexive le place dans l'héritage des modèles de la communication scientifique, dans la mesure où un certain nombre de productions de la science (publications, conférences) sont aussi des véhicules de communication des analyses et de leurs résultats mais aussi des débats encore non formalisés en théorie ou méthode. Ainsi, le système de téléconférence de Usenet (un de nos cas d'étude principaux) mis au point par la communauté d'ingénieurs et scientifiques utilisant les systèmes Unix à la fin des années 1970, est au départ un support de recherche d'information et d'échange à propos de leurs travaux sur ces systèmes (« Unix User Networks ») qui vient se greffer sur un réseau socio-académique de collaboration international (Paloque-Berges, 2013d). Sous sa forme collective (listes et groupes de discussion) qui intéresse notre travail post-doctoral au premier chef, la discussion collective asynchrone médiée par les réseaux informatiques a longtemps été appelé « conférence électronique »[8]. Il est possible que les modèles scientifiques (conférence, mais aussi publication) aient subi une évolution, voire une altération, par leur biais : ils accélèrent et multiplient en effet la diffusion et la discussion de la littérature blanche académique (version traditionnelle ou numérique) et de la littérature grise en ingénierie informatique de réseau[9] (Le Crosnier, 1995). Cette question de l'évolution des modèles scientifiques à l'aune des outils et environnements numériques est d'ailleurs d'une saisissante actualité avec les débats contemporains sur l'Open Access.

Le deuxième degré de cette valorisation est utopique. Si les réseaux informatiques sont d'abord dédiés à la communication entre machines, ils sont nourris par les visions prospectives des penseurs de l'ère cybernétique qui théorisent la communication homme-machine et homme-homme à travers des réseaux d'ordinateurs. Avant même son invention, l'un des visionnaires des réseaux informatiques, Joseph Licklider, dès la fin des années 1960, loue l'interaction homme-machine car elle développe l'intelligence humaine dans l'alliance de deux modes d'organisation cognitive (Licklider utilise la métaphore de la symbiose). Autre visionnaire dont l'informatique communicante reconnaît la paternité, Douglas Engelbart parle d'augmentation de la cognition grâce à une utilisation plus souple, plus intuitive des techniques de la communication médiée par ordinateur et de leurs interfaces. L'utopie de l'intelligence collective, louée par la première génération de chercheurs en SHS s'intéressant à Internet, repose ainsi sur une croyance dans le pouvoir de la communication électronique de faire plus et de faire mieux avec l'informatique en réseau (Lévy, 1999).

différents réseaux informatiques entre eux. L'Arpanet adopte le protocole en 1983 et le terme Internet émerge à ce moment là.

8 cf. en particulier l'ouvrage somme, presque un annuaire des réseaux informatiques dans les années 1980, de John Quarterman, *The Matrix*, sorti en 1989, qui détaille les capacités techniques et sociales du « *online conferencing* » ou « *electronic conferencing* » (Quarterman, 1989).

9 Un exemple : la collaboration internationale sur les protocoles et standards de l'Internet à travers les RFC, « Request For Comments », qui se fait depuis 30 ans largement sur la base des échanges courriels collectifs.

L'utopie de la communication « augmentée » n'a cependant pas été toujours valorisée. Son histoire est accompagnée d'une forme de suspicion envers ce nouveau mode de communication qui facilite l'échange, mais l'inscrit dans une dimension informelle qui échappe au contrôle. Son invention est d'ailleurs accompagnée par l'idée que ce moyen d'échanger du langage humain par le bais des réseaux informatiques pourrait faire perdre le temps précieux du partage des ressources informatiques qui sont l'objet des première expérience de l'informatique communicante, à une époque où le temps de travail passé sur les gros systèmes informatiques (avant l'arrivée de la micro-informatique et des ordinateurs personnels) est réduit et coûteux. Les anecdotes entourant sa naissance font ainsi état d'une technologie inventée clandestinement, en marge des programmes de recherche officiels de l'Arpa, certes rapidement adoptée par les directeurs de programmes pour son côté pratique, mais toujours marqué par une suspicion à l'égard de la perte de temps et la trivialité possible de son utilisation (Paloque-Berges, 2011a). Son invention est symbolique de l'apport des nouveaux inventeurs de l'ère numérique, les « hackers », qui, avant de devenir des figures sociales célébrées de la créativité informatiqus, ont été l'objet de toutes les inquiétudes, notamment en raison de leurs méthodes informelles de conception et d'utilisation des technologies et de leur culture anti-bureaucratique défiant l'autorité. Ils sont parmi les premiers experts de l'informatique à avoir promu la communication médiée par les réseaux, autant pour des usages professionnels que pour des usages plus personnels et ludiques[10].

La communication électronique a été perçue depuis comme portant les plus grands espoirs de prise de pouvoir par la foule, dans sa formulation utopique, ou au contraire, dans sa version dystopique, comme le lieu où la foule montre son aspect « vulgaire » (*vulgus*), média de désinhibition des tabous culturels et sociaux.

Le troisième degré de cette valorisation est idéologique, l'innovation technique portant en elle la promesse d'un progrès social et politique, car fournissant des modèles de gouvernance égalitariens et libertariens (Paloque-Berges, 2013d, Flichy, 2001). Dans la généalogie des dispositifs de communication en réseau, la forme email (« message électronique » ou « courriel ») est un point de départ, une forme prototypique pour la quasi intégralité des échanges numériques asynchrones : elle est le « premier exemplaire » d'un modèle de la communication inter-individuelle puis collective sur les réseaux informatique, exemplaire répété, varié et complexifié au fur et à mesure du développement d'applications dédiées au facteur humain dans la transmission numérique (du travail coopératif à la sociabilité de l'échange interpersonnel de courriels aux médias sociaux du Web 2.0.).

Cette dimension exemplaire s'illustre dans la place privilégiée qu'a la communication médiée par réseau dans l'histoire d'Internet, et plus exactement dans sa pré-historiographie, c'est-à-dire les premières histoires racontées et mises par écrit par des acteurs, témoins directs du développement des réseaux numériques[11]. Symbole d'une appropriation « humaine » de la communication informatique pour des besoins organisationnels[12], l'invention de l'email figure dans toutes les chronologies et histoires d'Internet. Cette exemplarité concerne les filtres idéologiques appliqués par les acteurs et témoins directs à l'usage et l'observation de ces réseaux informatiques devenus « humains ». En France,

10 Selon l'un des promoteurs de la philosophie hacker, Eric S. Raymond, les listes électroniques ont progressivement évolué, notamment sous l'impulsion des collectifs hackers, d'un usage professionnel (d'abord tourné vers la coopération sur l'informatique de réseau) à un usage personnel : « *The facilities for electronic mailing lists that had been used to foster cooperation among continent-wide special-interest groups were increasingly also used for more social and recreational purposes.* » (Raymond, 2001).

11 Alexandre Serres parle pour cette première phase historiographique d'une histoire-chronique, événementielle, fondée sur le récit élogieux des hauts-faits des pères de l'Internet et de leur confrères, et manquant de réflexion méthodologique critique sur les sources (Serres, 2000).

12 Cf. chapitre 1, partie 1, « Histoires du Net : écrire l'histoire des origines du réseau des réseaux », pp.32-81, in (Paloque-Berges, 2011a).

puisque c'est le terrain qui nous intéresse ici, Christian Huitema dédie le premier chapitre de son témoignage *Et Dieu créa Internet* (1995) aux emails et aux groupes de discussion, lieux les plus populaires de l'Internet bien avant que le Web n'arrive au début des années 1990 ; il lie ce mode de communication à l'avènement d'une « démocratie électronique ». La messagerie est ainsi reconnue comme un vecteur de développement d'Internet sur le plan de l'évolution des modes d'organisation qui ont permis à Internet de devenir le modèle d'une gouvernance technique réflexive et récursive. Elle est un des motifs centraux d'un « imaginaire » idéologique, un discours célébrant l'ouverture et la liberté des moyens de communication et de circulation de l'information inscrit dans le développement et l'usage de ces moyens mêmes (Flichy, 2001).

1.2. Une anthropologie historique de la communication technique sur Internet

Les séminaires des membres du LabEx HASTEC et de leurs partenaires nous ont accompagné dans notre recherche de cadres théoriques et méthodologiques à même de mieux dessiner les contours de notre objet de recherche en tant que « technologie intellectuelle et matérielle ». En effet, selon notre hypothèse de départ, la communication électronique au sein des communautés techniques de l'Internet n'est pas seulement un support d'interaction et d'organisation : elle est l'nscription d'une activité discursive et pratique qui produit des savoirs, et donc des formes de pouvoir.

L'histoire de la communication technique sur et par les réseaux informatiques commence à un moment où émerge un nouveau régime du savoir dans « *une multiplication des lieux où du savoir et/ou de l'innovation sont produits, une diversification des modes d'appropriation de ces savoirs* » (Pestre, 2010). Les technosciences, comme Dominique Pestre les appelle, se situent à la croisée de trois univers : la science académique et ses savants, l'Etat fournisseur de mesures et de normes, et le développement industriel fondé sur une recherche économique, qui s'allient dans une croyance rationaliste dans le progrès social permis par les sciences et techniques et autour de la figure de l'expert-ingénieur.

Comment est construit et inscrit le savoir que l'on cherche ? Comment le délimiter non pas seulement selon des discours, mais grâce aux objets et aux processus par lequel il est médié ? La communication médiée par les réseaux devient un objet intellectuel que l'on peut comprendre à travers ses manipulations et opérations matérielles.

1.2.1. Topiques et croyances dans l'imaginaire des techniques

L'un des séminaires que nous avons suivi dans le cadre des programmes collaboratifs d'HASTEC est celui dirigé par Nathalie Luca, « Techniques du faire-croire » (PC 3). La technique y est interrogée en tant qu'incarnation d'un croire qui peut prendre des formes scientifiques, en tout cas perçues comme telles (des images, des mises en discours, des objets de démonstration...). L'attention accordée à la croyance dans ce programme de recherche concernait d'abord le domaine du religieux et de l'eschatologique, mais s'est aussi portée sur d'autres domaines de croyance, notamment idéologiques. La question générale « à quoi croit-on ? », qui relève d'une anthropologie des idées, est ainsi reformulée à l'aune des contextes et opérations techniques qui permettent de comprendre « comment l'on croit », et donc les techniques du faire-croire qui remontent à la rhétorique antique. Notre sujet de recherche est concerné au premier chef par cette réflexion, les technologies numériques étant devenues l'objet d'espoirs et de peurs contemporaines.

Cette question de la croyance, au cœur des utopies d'Internet, est liée à la construction d'un imaginaire techno-idéologique (Flichy, 2001) ; nous nous sommes donc penchés sur les rapports entre

technique et imaginaire. Nous avons pour cela assisté au séminaire « Penser la technique en société », organisé par le CNAM en collaboration avec les laboratoires d'histoire des sciences et des techniques du PRES HESAM, et avons rencontré dans les propositions de Anne-Françoise Garçon (qui organisait les séances de 2013 sous l'intitulé « Régimes de la pensée opératoires ») de quoi réfléchir aux mises en discours de l'activité technique rencontrées dans les communications électroniques des scientifiques et ingénieurs des réseaux informatiques.

Garçon promeut un regard anthropologique sur l'histoire des techniques, selon l'approche anthropo-historique récente déployée depuis les années 1990. Si la méthode anthropologique peut paraître contradictoire pour l'analyse historique, même récente, elle engage en fait à se tourner vers des sources différentes que celles des archives institutionnelles ou des objets techniques pris comme œuvres matérielles autonomes. La technologie y est présentée comme une science humaine qui s'appuie sur la circulation des mots qui introduisent aux objets techniques avant même qu'ils ne soient appréhendés matériellement. Nous sommes relativement proches en ceci de l'imaginaire tel que défini par Patrice Flichy : l'imaginaire d'Internet est formé de discours d'anticipation qui accompagnent et déterminent le développement et l'innovation technologique des réseaux informatiques, entraînant même des phénomènes de prophétie auto-réalisatrice (Flichy, 2001). Les techniques, ainsi, se constitueraient dans une mise en récit de narrations et de représentations, un imaginaire qui va parfois plus vite que le développement de la technique.

Garçon rappelle que le régime anthropologique des techniques (à partir de l'époque moderne) est lié à l'écrit. En effet, l'écrit permet le développement de moyens d'inscrire les procédés techniques dans la mémoire collective grâce à une énonciation qui en viendra à caractériser le langage de l'ingénieur ; par ailleurs, il accompagne la division du travail scientifique dans la production expérimentale puis industrielle d'objets normalisés par l'usage, qui en viendra à caractériser le langage du gestionnaire. Parmi ces mises en récits écrites, rechercher les topiques (lieux communs de pensée) permet de situer des points d'accroche dans les imaginaires des techniques par rapport à leur perspective sociétale. Ces topiques sont associées à des normes qui, évoluant, ouvrent le champ à différents régimes de la technique compris à travers ses « *pensées opératoires* ». La description des procédures liées à une technique, ainsi, permet d'étudier les normes opératoires de la pensée technique sans la réduire à des positionnements sociaux ou aux gestes techniques. L'intérêt accordé à une « *épistémologie vernaculaire* », permet d'éviter d'étudier le développement des techniques selon un progrès linéaire, mais d'en envisager les réappropriations artisanales face aux stratégies industrielles. Cela nous intéresse tout particulièrement puisque le développement des technologies informatiques, loin de ne procéder que par une logique cumulative de la conception à l'usage, montre des allers-retours entre invention et innovation qui passent notamment par des formes de bricolages, de l' « exploit » hacker à la personnalisation des artefacts par les amateurs et les néophytes.

La notion de topique nous intéresse car elle est utile pour faire émerger la complexité de l'objet technique :

- dans le temps, selon différentes vitesses d'évolution des techniques et de leur imaginaire et selon les logiques de transmission du savoir à l'oeuvre ;
- dans des réseaux de sens qui articulent les manières de faire aux manières de dire, la pensée à l'activité.

Elle constitue ainsi une chaîne opératoire qui met en discours la constitution de l'objet technique. Plus spécifiquement, la topique s'intéresse à l'origine à la mise en récit à travers des lieux communs et à la réception.

Cependant, ce qui nous manque dans cette perspective, c'est une vision claire des moyens d'intégration dans la pratique par les objets techniques eux-mêmes. Les techniques intellectuelles de l'écrit, comme les listes par exemple (et cet exemple nous intéresse puisque nous parlons aussi d'une

forme de liste, les listes et groupes de discussion) opèrent la banalisation de nouveaux procédés, des processus d'adaptation (Garçon, 2012 : 189). Mais qu'en est-il du rôle joué par les pratiques et techniques matérielles (dont la logique informatique relève par son lien à l'action et à l'objet ordinateur) ?

1.2.2. Lieux de savoir et géographie de l'Internet

Nous avons suivi le séminaire de Christian Jacob donné à l'EHESS sur les « Lieux de savoirs ». Jacob donne une définition fonctionnelle de la notion de savoir, qui permet d'inclure dans l'étude des pratiques savantes des dimensions autres que celles des sciences légitimes – ce qui est pratique dans le cas de l'informatique de réseau, dont le statut de « technoscience » relève d'un hybride. Le savoir est ainsi défini à travers deux tournants dans les méthodes de la recherche en SHS qui complètent le tournant « linguistique » présenté plus haut (sémantisation des techniques de production du savoir) :

- un tournant pratique pratique, qui s'intéresse aux manipulations techniques,
- un tournant spatial qui prend en compte des lieux où sont déployés les dispositifs de production du savoir).

Ces deux démarches nous semblent importantes pour comprendre la production du savoir sur les réseaux informatiques à travers les savoir-faire construits dans l'utilisation de ces réseaux : une matérialisation du savoir analysée dans les pratiques logicielles de la communication en ligne.

Le savoir peut ainsi être compris comme une compétence objectivée dans un énoncé ou un artefact par sa matérialisation, son inscription dans des objets et des lieux. Cela nous semble très important pour comprendre le lien fort entre les messages communiquant des sujets relatifs aux techniques de réseau et les applications logicielles qui les accueillent, les conditionnent, les transportent. Le savoir, en tant qu'il est communiqué, n'est pas seulement transmis : il est produit à la rencontre de l'activité (le développement et la maîtrise des logiciels de communication en réseau), et de la sémantisation (la mise en discours accompagnée par les logiciels) dans des lieux qu'on peut décrire d'un point de vue intellectuel, mais aussi pratique, en intégrant dans l'analyse de la matérialité des échanges des implications sociales, culturelles, politiques, économiques... Dans cette perspective, les communications des collectifs savants que nous étudions peuvent être étudiées sous l'angle d'un « *lieu de labeur et d'élaboration* » qui donne une identité au collectif en tant que producteurs en relation plus moins indépendante des institutions auxquelles ils sont affiliés (Mandressi, 2007).

Cette approche nécessite une réflexivité puisque l'observateur lui-même est aux prises avec des objets qu'il sélectionne et exploite. La problématique des sources, des instruments et plus généralement des méthodes devient alors primordiale et passe au premier plan.

Pour répondre à cela, la méthode d'analyse des Lieux de savoir porte moins sur les idées et la reconstruction des systèmes de pensée ou des systèmes techniques en eux-mêmes (une approche internaliste) et davantage sur les mises en forme et les modes de production qui sous-tendent leur compréhension et leur manipulation ainsi que les contextes sociaux (une approche externaliste). L'approche génétique, adoptée en études littéraires et en épistémologie depuis quelques décennies, donne une idée des objets sur lesquels on peut se pencher dans cette perspective : des brouillons, des carnets de notes, des correspondances, des instruments permettant de comprendre la construction d'une œuvre de l'esprit à travers son évolution matérielle[13]. Ce savoir « en train de se faire » n'est pas soumis à une variabilité informe. Le travail de savoir est codifié dans le savoir-faire pratique, l'artefact technique et le contexte parce que ces derniers sont un terrain sur lesquels il se bâtit selon des normes discutées et parfois disputées.

Dans le cas de l'informatique de réseau, les articles ou la documentation techniques (littératures

13 Qui ont pu être appelées des « *écritures ordinaires de la recherche* » (Lefevbre, 2013).

blanche et grise) qui diffusent un concept technologique au sein d'une communauté d'ingénieurs ou de scientifiques ne donnent accès qu'à sa forme à un instant donné, si l'on met à part les notes bibliographiques que l'auteur y inclut afin de montrer ses références. Selon une approche orientée sur la pratique et la matérialité, au contraire, l'étude des situations d'utilisation d'une technique ou des artefacts qui sous-tendent ou résultent de cette situation permet de saisir le collectif en train de construire ce savoir, indissociable d'un savoir-faire. En ceci, l'informatique de réseau n'est pas un cas de science appliquée au sens strict mais un savoir technique en construction dans l'usage. Le contexte d'usage est précisément dans notre cas ce qui confronte l'utilisation d'un outil de communication et la production d'une écriture à visée communicationnelle à des normes techniques (l'environnement informatique et de réseau), social (les règles de sociabilité dans la prise de parole en ligne) voire politique (la gouvernance et la régulation des réseaux). Ces normes incluent un contexte plus large que celui de l'environnement informatique, qui dans notre cas comprend les modes d'organisation des milieux de la recherche et du support en technologies d'information et de la communication dans les institutions universitaires et affiliées, la politique d'équipement au niveau national comme au niveau local du laboratoire de recherche dans le public ou le privé.

Dans ce cadre de réflexion, l'hypothèse des Lieux de savoir est celle d'une continuité entre pratiques quotidiennes et pratiques d'experts des technologies qui si elles ne sont pas équivalentes, présentent cependant une similarité en tant qu'elles sont l'opération de déplacements, de (re)formulations, de manipulation des normes techniques et sociales. Les lieux de communication numérique qui ont fait Internet et que nous étudions sont tout entier empreints du quotidien de la communication informelle, d'une part, même si des formalisations peuvent avoir lieu (au niveau social, par exemple la civilité de l'échange en ligne, comme au niveau technique, à travers les contraintes et affordances des protocoles et logiciels de communication). D'autre part, ils sont des lieux d'accueil des expérimentations des scientifiques et ingénieurs informatiques en train de développer, tester et échanger sur les technologies de réseau. En ceci, les listes et groupes de discussion que nous étudions préfigurent les espaces applicatifs du Web, en particulier les lieux d'expression et de communication que sont les blogs, le « *bric à brac du chercheur* », des textes fourre-tout, sans structure formelle mais au potentiel relationnel important pour la communication des sciences et des techniques (Dacos et Mounier, 2011).

Enfin, considérer ces modes d'interactions en réseau comme des Lieux de savoir, c'est creuser la métaphore spatiale attachée à l'imaginaire d'Internet en tant que cyberespace. Si cette dernière, ainsi que la métaphore associée de la frontière électronique, a été critiquée et relativisée après avoir rencontré un grand succès dans les années 1990 (Paloque-Berges, 2011a), elle constitue un point d'arrivée important pour notre étude. L'idée que cet espace qui repousse sans cesse ses limites grâce au progrès technologique puisse produire une « démocratie technique » repose sur le fantasme qu'il ouvre virtuellement un forum, un agora électronique, où la mise en commun et en discussion des savoirs peut produire des structures de pouvoir plus égalitaires, une des croyances fortes liées à Internet. En effet, la construction et le déploiement des réseaux informatiques que nous avons étudiés jusqu'à leur ouverture au grand public au milieu des années 1990 ont abouti à la formulation d'une utopie géopolitique autonome et indépendante des structures traditionnelles – utopie ambivalente qui produit de nouvelles articulations entre savoir et pouvoir technoscientifique fondé notamment sur des structures organisationnelles et communicationnelles relativement informelles comme les discussions électroniques, mais offre à la vue un décalage entre idéalisme et réalisation pratique (Paloque-Berges, 2013d). Si la version idéaliste de l'histoire d'Internet avant l'ouverture au grand public pourrait faire penser que l'on assiste à la prise de pouvoir d'« *associations hybrides* » (impliquant la participation d'experts et d'amateurs et plus généralement d'acteurs issus de différents milieux socio-professionnels – Callon, Lascoumes et Barthe, 2001), une analyse des communautés de pratique montre en fait, sous la

surface d'une organisation informelle associative et médiée par les réseaux la prééminence d'une vision d'ingénieur expert. Ainsi, le regard spatial des Lieux de savoir est important pour comprendre la géographie de l'Internet, c'est-à-dire la manière dont les réseaux informatiques ont été cartographiés à l'échelle 1:1 par les catégories socio-professionnelles des informaticiens dont « *l'écriture en réseau définit le territoire* » (Guichard, 2007).

1.3. Une problématique communicationnelle orientée Science, Technologie et Société (STS)

Aborder la question de la communication électronique dans l'histoire des techniques implique une réflexion croisée entre sciences de l'information et la communication et étude des sciences et des techniques, en particulier le courant « Science, Technologie et Société » (STS), qui pose notamment la question du rôle des infrastructures sociotechniques, notamment à travers la question de l'équipement technique sous-tendant la communication dans le domaine techno-scientifique.

1.3.1. Une communication scientifique, technique ou ordinaire ?

L'intérêt que nous défendons ici pour les communications électroniques entre scientifiques et ingénieurs de réseau relève du nouvel intérêt de l'histoire des sciences pour les documents longtemps considérés comme secondaires. S'intéresser à ces documents relève d'une réflexion sur *« l'extension du terrain de l'enquête* [...] *le renouvellement le plus important qu'a connu l'histoire des sciences* » depuis vingt ans et qui comprend « *d'autres sources que l'imprimé: les manuscrits, les correspondances, les documents administratifs, la littérature grise et les carnets de laboratoires, les machines et les instruments* » (Brian, 2001) ; outre les aspects épistémologiques révélés par cette nouvelle réflexion, ce sont des aspects institutionnels (les désaccords entre groupes de scientifiques sur des « styles scientifiques ») et politiques (les rapports à la gouvernance des sciences) qui peuvent être éclairés.

Cependant, ces communications ont-elles un « style » scientifique ? (Mourlhon-Dallies et Colin, 2004) se sont intéressés aux « *rituels énonciatifs des réseaux informatiques entre scientifiques* » dans le cadre des communications sur les groupes de discussion Usenet, qui nourrissent notre corpus. Ils ont observé que les codifications généralement impliquées dans les discours spécialisés sont, dans ce contexte, surdéterminées par un autre type de codification, celui de « *l'esprit du réseau* » tel qu'il s'incarne dans des formalismes divers de l'écriture dans la communication médiée par les réseaux. Les groupes de Usenet « *revisitent les codes communicatifs ordinaires* » et les détournent pour mieux les adapter aux usages communicationnels en réseau numérique (le formalisme épistolaire revu dans le cadre de l'échange courriel, la prise en compte de l'effet d'immédiateté de la communication numérique, des codes formels concernant la typographie ou le lexique). Cela remettrait ainsi en question « *la notion de spécialisation en termes d'opposition lexicale entre 'langue technique' et 'langue courante'* ». La spécificité des échanges entre scientifiques se trouverait ainsi non pas tant dans la démarcation entre un savoir savant et un savoir profane sur les objets de recherche, mais entre un savoir expérimenté et une ignorance de néophyte quant aux codes mêmes de la communication sur réseaux informatiques. Les chercheurs proposent ainsi une méthode d'analyse portant sur l'énonciation fondée sur l'idée qu' « *en matière de communication spécialisée, la connivence dans la formulation est donc un élément tout aussi important que le contenu du message lui-même* ».

Nous pouvons souligner que l'étude de Mourlhon-Dallies et Colin, bien que fort intéressante et pertinente dans sa mise au jour des codes de communication sur Usenet et Internet, est grevée par un *a priori problématique* : ils considèrent que la communication qui s'y déroule se passe entre scientifiques, ce qui n'est pas historiquement exact. En effet, Usenet est un moyen de communication très ouvert, et ce dès ses premières années dans les années 1980, et s'il montre, au moins pendant une décennie, une

concentration d'utilisateurs plutôt compétents en informatique, c'est davantage un lieu d'échange d'ingénieurs que de scientifiques (bien que la distinction soit difficile à faire, car nous n'avons pas de données socio-professionnelles précises sur l'usage de Usenet et d'Internet avant son ouverture au grand public dans les années 1990). Si un style existe, au-delà des codifications d'usage du discours électronique, il se situerait davantage à la croisée du discours de l'ingénieur et du hacker, à la fois techno-centré, positiviste, défiant les autorités des institutions traditionnelle et porté sur la satire et l'auto-dérision (Paloque-Berges, 2011a, 2013d).

1.3.2. Réseaux des sciences : la communication comme médiation technique et sociale

La sociologie des sciences, et en particulier le mouvement de l'Ecole des Mines, s'intéresse de prime abord aux réseaux de la science en tant qu'ils permettent de saisir la « *genèse et circulation des faits scientifiques* », comme l'indique le sous-titre de l'ouvrage dirigé par Michel Callon (1989) dans leurs conditions contextuelles et temporelles. A partir de la théorie de l'acteur-réseau (Actor Network Theory, ou ANT), les objets techniques relevant de l'équipement des scientifiques acquièrent une voix au même titre que les humains qui les conçoivent ou les utilisent (selon le principe de « *symétrie radicale* » proposé par Bruno Latour). A travers une série de prescriptions, ils sont les porte-parole de leurs concepteurs et ouvrent un espace d'usage (qui pourra être réapproprié et réinventé par les utilisateurs) :

> l'équipement apporte la parole de ceux qui l'ont conçu, élaboré, perfectionné, fabriqué. Il l'apporte écrite dans le hard (telle touche a telle fonctionnalités, telles opérations sont impossibles ; il l'apporte sous la forme de modes d'emploi plus ou moins ésotériques et ambigus qui l'accompagnent, ils l'apporte en imposant d'opérer certains branchements standardisés sur des matériels ou des équipements existants ; il l'apporte également sous la forme des démonstrateurs, qui interviennent avant qu'on ne le fasse fonctionner. (Callon, 1989 : 19)

L'adoption de certains équipements par certains acteurs plutôt que d'autres ainsi que leur manière de les utiliser rend sensible des choix qui crée des « *systèmes d'alliance* ». En ceci, les objets techniques sont des maillons dans une chaîne de médiateurs, selon une vision de la science comme « *processionnaire* », c'est-à-dire un réseau hétérogène lié et redéfini constamment par la circulation de ces objets et de leurs savoir associés (*ibid.* : 22).

Cependant, cette perspective ne met permet de saisir la formation des réseaux qu'au moment du choix de l'équipement technique. Or ce choix est guidé en amont par des processus de reconnaissance socio-professionnelles : si les réseaux renforcent les communautés, un sens du lieu commun (ce qui permet à la communauté de se reconnaître) préexiste à la formation des réseaux communautaires. Nous avons ainsi étudié la manière dont la communauté Unix s'est constituée dans la reconnaissance d'une identité de l'ingénieur informatique travaillant dans les institutions académiques à des fonctions supports non valorisées, voire ignorées des chercheurs et de la hiérarchie, et fondant son projet de réseaux informatiques sur la volonté de créer une communauté internationale de pairs, autonome, indépendante, et avec une culture technique fortement revendiquée doublée par une sociabilisation accrue autour de rencontres de professionnels (Paloque-Berges, 2013d). Le développement des réseaux techniques (et donc le choix de certains de s'équiper en matériel et logiciel d'équipement pour supporter la communication électronique) n'aurait pu se faire sans un réseau humain qui s'étend par une forme de bouche-à-oreille d'initiés. Ici, la communication est motivée socialement avant même de 'êêtre par l'équipement, comme le décrit Bourdieu dans *Science de la science et réflexivité* :

> D'abord, on a un « groupe paradigme » qui s'intéresse au même problème de recherche et constitue un réservoir de contacts potentiels. Puis, des relations réelles s'instaurent à travers

> un « réseau de communication » qui s'accroit par cooptation successives. Puis on voit se créer peu à peu un véritable cluster. [...] La reconnaissance en tant que groupe est fondée sur l'existence d'un style intellectuel commun (dogme central) et d'une vie sociale [...] et aussi, évidemment, sur les premières inventions. (Bourdieu, 2001 : 135-136).

Dans le cas des collectifs informaticiens qui nous intéressent, si le « style intellectuel » est difficile à mettre en évidence, l' « *esprit de réseau* » caractéristique des échanges électroniques en constitue un avatar. Comment cet esprit de réseau en vient-il à sous-tendre l'idée qu'une communauté de collaborateurs autour du projet Internet existe, et surtout, sur quels modes de reconnaissance et donc d'exclusion fonctionne-il ?

1.3.3. Les contradictions d'une utopie de la démocratie technique

Les études sur l'histoire d'Internet ont révélé l'existence, à la genèse des réseaux informatiques, d'une « *république des informaticiens* » (Flichy, 1996). Sur le modèle de la notion de « communauté scientifique » héritée de Merton, les premières années du développement d'Internet sont expliquées par la formation d'une utopie bâtie sur l'éthique de la science ouverte (« *ethos of open science* », Merton, 1942). Perçue comme l'« *organisation sociale de référence d'Internet* » (Hert, 1997), elle revendique les valeurs d'échange et de coopération et d'égalitarisme dans la participation au projet. Cette communauté est marquée à la fois par une volonté d'autonomisation et rattrapée par l'intérêt croissant d'utilisateurs de plus en plus extérieurs aux collectifs initiaux de goûter à ces technologies. À la genèse d'Arpanet existe ainsi une « Netville » qui, malgré le caractère ouvert des technologies de réseau (la décision ayant été prise de ne pas faire des applications et protocoles de réseau des logiciels propriétaires), a vécu pendant ses deux premières décennies dans un équilibre instable, hésitant entre la volonté de créer une expertise propre et d'imposer des standards à tout ceux qui participent au projet d'Internet, et celle de maintenir une flexibilité et une logique d'inclusion des compétences afin de laisser au projet son caractère expérimental et innovant (King, Grinter et Pickering, 1997). Plus tard, alors que le grand public arrive à Internet, les tenants de la culture technique pionnière d'Internet ont tenté de transformer cette volonté d'autonomisation en programme politique, avec notamment le « *manifeste d'indépendance du cyberespace* » (Barlow, 1996), texte symbolique des utopies de la cyberculture.

Cependant, la mesure d'évaluation des succès et échecs du développement des réseaux n'est pas proprement autonome, puisque une fonction support a été assignée à l'informatique de réseau dès le départ (au contraire des « sciences fondamentales » de l'informatique, appartenant au domaine des mathématiques) : son but pratique était de renforcer la production du travail scientifique dans un « *raccourci entre usage et recherche* » (Hert, 1997 : 85). Son évolution jusqu'à aujourd'hui dans le domaine des équipements en information et en communication des organisations et des particuliers a confirmé cette orientation préliminaire. L'importance de l'usage, et donc de l'applicabilité des technologies de réseaux à des logiques d'utilisation selon des normes non réductibles aux seules dimensions scientifiques et techniques (mais devant répondre à un contexte social et économique), est ce qui a cassé la logique d'autonomisation des savoirs de l'informatique de réseau, le lien fort entre technologie informatique et société l'empêchant de se constituer en champ au sens strict du terme.

Il est intéressant de constater que cette instabilité relève non pas seulement d'une opposition externe (les experts sollicités par le monde extérieur) mais d'une contradiction externe : dans le projet Internet initial est déjà formulé le projet d'un réseau de communication global devenu bien commun – ferments de ce qui a été nommé l'utopie d'Internet. Nous avons étudié cette contradiction à propos des collectifs Unixiens, qui tentent de créer une communauté inclusive, défiante des autorités et des élites, en développant les équipements de réseau aux Etats-Unis et en Europe par une série de mobilisations

informelles mais qui à leur tour buttent contre la difficulté de créer une médiation efficace et idéologiquement conforme à leur projet de départ auprès des grands publics (Paloque-Berges, 2013c et 2013d).

En ceci, cette médiation difficile indique que « réseaux de la science » ne fonctionnent pas seulement par des alliances librement choisies mais par des contraintes sociales très fortes sur la possibilité d'adopter un dispositif technique comme l'équipement des réseaux informatiques. Le succédané de la théorie de l'acteur-réseau, à savoir la possibilité pour le domaine scientifique de créer des mobilisations de plein droit autour d'une « démocratie technique ». Si l'on suit le discours de justification des acteurs, on pourrait avancer que l'inter-objet « communication technique en réseau » tend vers la réalisation pratique de « *forums hybrides* » et de la « *démocratie technique* » (Callon, Lascoumes et Barthe, 2001). En effet, les groupes de discussion Usenet (ancêtres des forums Web) sont assimilés dans l'utopie Internet à une forme démocratique fondée sur l'apprentissage collectif et le dialogue et qui se distingue des modes traditionnels de médiation des sciences et des techniques. Ces derniers, selon le modèle des « forums constituants » sont marqués par une médiation descendante ; les forums hybrides, *a contrario*, se fondent sur des relations verticales, travaillées par la collaboration, la participation d'acteurs variés, en particulier sur le plan de l'expertise (les « profanes » pouvant participer de plein droit à la discussion). Les modes de communication médiées par les réseaux sont promus dans l'imaginaire Internet comme l'avènement des formes de discussion idéales que représenteraient les forums hybrides, et Usenet en serait un des symboles historiques dans l'histoire des médias de réseau numérique. Cependant, le modèle de la démocratie technique peut être critiqué pour son aveuglement relatif aux inégalités persistantes dans l'accès et l'usage des dispositifs des forums hybrides ; postulant une égalité et une transparente de fait de ces dispositifs, il oublierait les asymétries subsistant entre les acteurs. C'est une critique qui a été dressée également à l'encontre des descriptions idéalistes (dites cyberculturelles) des communications Internet marquant les premières analyses de SHS sur les réseaux informatiques (Wellman, 2011).

Nous avons présenté ce travail à l'école d'été de la Cité des Télécoms[14] et nous sommes rendus compte à cette occasion, lors d'un débat sur la difficulté pour l'historiographie des techniques de s'attacher aux publics d'usage, que la contradiction des utopies pionnières d'Internet résidait dans ambiguïté du terme « utilisateur » (*user*). L'utilisateur, en informatique, est d'abord un expert, premier utilisateur des outils d'équipement qu'il développe pour mieux travailler ; mais il est aussi l'utilisateur dit « final », qui guide le développement d'applications plus faciles d'utilisation (grâce aux « interfaces orientées utilisateur »). Entre les deux : la foule des profils d'utilisateurs variés aussi bien sur le plan des compétences techniques, des fonctions professionnelles que des intentions et réalisations d'usage. Les informaticiens de réseau, qui chérissent le terme « communauté », ont créé « *artificiellement une position médiane qui ne traduit pas réellement communauté réelle* » (Hert, 1999), en investissant symboliquement les espaces du virtuels comme promesse de donner du sens au collectif communiquant (c'est un mythe véhiculé par les premières théories des médias, en particulier celle du « *village global* » de McLuhan). Les dispositifs de communication médiée par les réseaux sont particulièrement propices à croire dans la possibilité de trouver une nouvelle agora où le « peuple » pourrait s'exprimer sans entrave, grâce au progrès technique pour tous. Mais c'est bien évidemment sans compter que l'utilisation de cette technique est déterminée socialement.

Pour finir, nous évoquerons la notion d'interobjectivation (Voirol, 2013) qui nous nous intéresse car elle est fondée sur une approche croisée et critique de la théorie de la reconnaissance de Honneth et

14 Organisée par Pascal Griset et Léonard Laborie en septembre 2013, dans le cadre du projet ANR Resendem « Les grands réseaux techniques en démocratie : innovation, usages et groupes impliqués dans la longue durée (fin du 19e - début du 21e s.) » (2010-2014, Irice/Paris 4, CEMMC/Bordeaux 3, Triangle/Ens Lyon, Laboratoire Communication et Politique/CNRS).

de celle de l'Acteur-Réseau de Latour[15]. A partir du processus de reconnaissance subjective de la médiation de l'objet, il y ajoute la dimension pratique du programme d'action (*praxis*). Dans notre cas, l'objet médiateur partagé dans les relations collectives est celui de la communication numérique en réseau au sein des collectifs participant au développement des réseaux informatiques. Il définit des lignes de partage et d'interaction entre la reconnaissance sociale des individus et le groupe nourrie par l'éthique hacker, entre responsabilité individuelle et désir de consensus. L'exemple des Unixiens que nous avons analysé montre comment la communication médiée par les réseaux informatiques est l'objet d'une médiation importante pour ces collectifs, mais opère plutôt le renforcement de l'autorité de l'ingénieur plutôt que sa redistribution dans une chaîne plus souple d'utilisateurs-acteurs.

Nous avons donc fait le tour des perspectives théoriques générales que l'on peut convoquer pour appréhender l'objet de la communication médiée en réseau. C'est dans ce cadre que nous avons pu avancer sur l'histoire de l'équipement français en matière de réseaux informatiques de communication de type Internet, dont nous avons présenté les résultats dans (Paloque-Berges 2013c et 2013d), certains étant donnés ici à titre d'exemple pour vérifier la pertinence des théories convoquées.

Communication entre techniciens appartenant à la sphère socio-professionnelle et idéologique des technosciences, elle traduit des logiques de champ et révèlent des enjeux de pouvoir persistant au-delà de l'apparente informalité des pratiques communicationnelles et des choix *ad hoc* qui leur sont liés. Si elles s'appuient aussi sur des logiques de réseau, au sens politique comme technologique, c'est à travers des formes d'interobjectivation qui restent relativement exclusives aux utilisateurs non experts de l'informatique de réseau, bien que leurs discours d'accompagnement, mais aussi un certain nombre de réalisations techniques qui ne concernent pas la période de notre corpus, tissent un lien entre innovation et usage à l'échelle de la société. Dans le cadre de ce rapport, nous ne nous attarderons pas sur nos résultats en terme d'anthropologie historique d'Internet en France jusqu'aux années 1990, car nous souhaitons creuser la question des méthodologies à convoquer pour faire de telles analyses, ainsi que la réalisation pratique de mise en corpus des archives des communications électroniques pour ce faire ; nous invitons le lecteur à consulter nos publications récentes qui racontent cette histoire à partir de ces sources.

15 Cette notion guide un travail d'analyse effectué dans le cadre d'un cas d'étude sur le corpus de post-doctrat avec deux collaboratrices du laboratoire DICEN, Claire Scopsi et Haud Gueguen, pour une communication au colloque « History and Philosophy of Computing » 2013 intitulée « What network computing does to communication. A retrospective analysis of early debates confronting and inventing online communication ethics ».

2. Propositions méthodologiques pour l'analyse des communications en réseau documentarisées en archives numériques natives

Internet devenu objet de recherche scientifique en sciences humaines et sociales présente de nombreux angles d'attaque : étude des techniques d'informations et de communication, des usages, des contenus. S'il serait naïf de placer ces études dans la perspective d'une nouveauté radicale et absolue, on peut cependant noter l'originalité de ce média dans la mesure où il est un terrain où chercher, et avec lequel chercher. Ainsi, les communications électroniques sont envisagées comme des sources documentaires pour l'étude d'une histoire des communications médiées par les réseaux.

Si la numérisation des documents ajoute de la valeur à l'information, elle permet aussi de la mettre en circulation. Mais comment appréhender les documents électroniques dits natifs (Chabin, 2003), c'est-à-dire ayant fonction de support d'une activité qui utilise Internet, produit originellement en contexte numérique pour être transmis sur le réseau ?

Nous souhaitons ici resserrer notre propos aux problématiques documentaires engagées dans l'étude des communications électroniques comme source pour l'histoire des télécommunications numérique en particulier, mais aussi plus généralement dans une perspective de SHS. Notre travail post-doctoral engageant un travail pratique sur ces sources (qui sera décrit en troisième partie), il faut d'ores et déjà se pencher sur les problèmes méthodologiques propres aux documents numériques qui portent jusqu'à nous les contenus et traces de ces communications, afin de spécifier à partir du cadre théorique général décrit en première partie les enjeux de nos sources dans l'optique de les étudier en tant qu'archives. En ceci, nous nous nourrissons de la démarche de l'archivistique, qui nécessite de « *connaître les documents de la mémoire du passé, identifier les documents de mémoire du présent, et de décrire le lien logique entre les deux ensembles* » (Chabin, 2000 : 35).

2.1. Savoir évaluer les tensions entre mémoire et archives numériques

Les liens entre technologie numérique et mémoire remontent à l'aube de l'histoire de l'informatique, comme en témoigne l'importation dans le vocabulaire des ordinateurs du champ sémantique de la matière grise (« cerveau électronique », « mémoire vive »...)[16]. L'ordinateur a ainsi alimenté beaucoup de fantasmes sur la possibilité de transporter et stocker des données comme jamais on n'avait pu le faire auparavant, et les réseaux informatiques les ont relayé et amplifié. Cependant, une réflexion sur la mémoire technique du numérique doit prendre en charge la faillibilité de la mémoire numérique pour dégager les méthodes spécifiques d'une archéologie du numérique. En effet, un document numérique n'est pas stable : il ne l'est que temporairement, selon des circonstances matérielles et logicielles à un moment donné. Ensuite, le propre de la mémoire numérique est de réécrire les données infiniment (efface et réécrit).

Le terme « archive », utilisé en anglais, est distinct du terme « enregistrement » (*record*), distinction qui n'apparaît pas dans le mot français. L'archive en français a été jusqu'à récemment plutôt liée au sens « d'enregistrement sur des documents d'archives au sein d'un dépôt pour référence » utilisé en priorité pour les archives administratives. Mais il s'enrichit de plus en plus de la flexibilité du terme anglais, en particulier de la notion d'archivage électronique, appropriée par les informaticiens et appliquée aux techniques de sauvegarde et de conservation de documents électroniques sur support disque, une opération physique : l'acte d'archivage (écriture, copie et organisation dans des fichiers

16 C'est John Von Neumann qui introduit le terme « mémoire » dans sa conception de l'architecture logique de l'ordinateur qui a fait date (Campbell-Kelly et Aspray 1997).

gravés sur un support) préexiste aux archives, et non pas le contraire (Chabin, 2000 : 54). Si l'informatique est productrice d'archives, les réseaux numériques ont mis un peu plus de temps à accueillir une réflexion pratique sur l'archivage de leurs documents, comme le déplore Marie-Anne Chabin en 2000 : « *Internet ne favorise pas l'archivage des sites et leurs archives car tous les regards sont encore tournés vers la nouveauté, instrument de toutes les séductions* » (*ibid.* : 57) – une situation qui changé aujourd'hui alors que fleurissent un peu partout, dans les universités, les institutions, mais aussi dans les communautés, des projets d'archivage du Web.

Dans cette partie nous tentons une réflexion archéologique sur la mémoire technique des communications électroniques perçues comme une archive au sens foucaldien, à savoir des systèmes d'énoncés qui témoignent de l'expérience épistémique, sociale, culturelle et technique d'Internet.

2.1.1. Une mémoire technique : de la communication au document à l'archive numérique

La jeune historiographie des réseaux numériques, on l'a vu, très imprégnée du discours des acteurs, accorde une importance particulière aux moyens de communication en réseau pour travailler sur ces réseaux. Si la communication électronique est investie de la mémoire des acteurs, c'est aussi parce qu'elle porte dans le temps et dans l'espace cette mémoire et l'inscrit sur des supports documentaires. Parmi les visionnaires des réseaux numériques déjà rencontrés, Joseph C. R. Licklider louait dès les années 1960 les possibilités du « cerveau électronique » (l'ordinateur conçu comme tel par la cybernétique) en tant que mémoire exosomatique[17] qui n'est pas seulement à l'extérieur de l'humain mais en « symbiose » avec lui (selon la métaphore proposée par Licklider pour décrire la communication homme-machine). La suite de l'histoire de l'informatique (notamment en réseau) sera consacrée à donner forme à ces données d'un point de vue informationnel, ergonomique, et sémiotique (avec l'arrivée des interfaces graphiques plus lisibles et « usables » par les utilisateurs profanes).

L'appropriation de la communication machinique par l'humain, d'abord par le courrier électronique, a introduit la possibilité de télécommuniquer l'échange interlocutoire à travers la convergence des moyens antérieurs (épistolaire, télégraphique, téléphonique) dans des dispositifs techniques nouveaux. Comme pour ses ancêtres télécommunicants, l'échange électronique, s'il est sauvegardé, produit des documents, à la différence près que l'inscription des communications se fait moins sur un support pérenne de stock (le papier) que sur un support instable de flux (les données numériques inscrites et constamment réinscrites sur les disques informatiques). L'instabilité et la labilité du numérique font qu'il a été difficile de considérer les échanges en réseau comme des discours écrits, et a fortiori comme des documents, alors que, et c'est ce que les chercheurs français en sciences de la documentation appellent le « paradoxe de Roger », « *le développement des échanges spontanés (la conversation) et leur fixation sur un support public pérenne et documenté. Autrement dit,* [Internet] *transforme automatiquement ce qui relevait de l'intime et de l'éphémère en document ou proto-document.* » (Salaün, in Pédauque, 2006 : 17-23). Le document, selon la définition d'autorité de Jean-Michel Salaün, est à la fois un vu (une forme associée à un support), un lu (un contenu, avec son encodage linguistique, ses données et métadonnées) et un su (qui s'insère dans une logique économique et juridique). Les documents d'Internet, produits des activités humaines sur le réseau, focalisent des communautés, permettent la rencontre, l'identification et l'échange entre les usagers, selon des modalités sociales et techniques qui ont été décrites par les sciences de l'information, de la communication et de la documentation à l'ère du Web, notamment à travers la notion de « *docu-*

17 L'autre grandes figure tutélaire de l'ordinateur-mémoire est Vannevar Bush, en particulier dans son travail expérimental sur les supports machiniques de stockage et de consultation de l'information, le MemEx (Memory Extender) (Bush, 1945).

médiatisation » (Zacklad, 2007).

Les dispositifs de communication asynchrone ont bien une qualité de document numérique, et cela même à l'aune de la définition classique du document, ce qui « *rassemble de l'information en un objet unique et fini. Si l'on s'en tient au critère du support et du produit final, les forums de discussion usenet sont clairement des documents numériques.* » (Marcoccia, 2001). Cependant, selon d'autres aspects, ils mettent en tension cette définition : par la mise en perspective de l'aspect conversationnel (temps réel) avec la stabilité de l'écrit (persistance des traces) et leur hiérarchisation dans l'espace de communication (message / sujet / fil de discussion / forum...), par la multiplicité des sources de la production de contenu dans un système d'édition automatisé (et non pas édité par une intentionnalité), et par la distribution des échanges sur le mode interactif, ce qui implique de trouver des effets de cohérence et d'intention ailleurs que dans les structures classiques du document (rôle des participants dans l'organisation, l'animation et la régulation des échanges).

> Un forum de discussion est alors un objet hybride pour l'analyse de discours, à la fois une archive identifiable par un support et un contenu et une conversation qui n'est jamais totalement finie. Il faut alors considérer les forums comme des documents numériques dynamiques, des documents comme objets de processus, ou, pour prendre les termes de l'analyse de discours, comme une archive en train de se constituer (en l'occurrence, une « archive conversationnelle »).
>
> [...]
>
> On peut [...] soit considérer qu'il s'agit de documents composites (des documents de documents), soit qu'ils sont produits par un processus d'écriture collective. La participation à un forum peut ainsi être vue comme la participation à l'écriture collective d'un document, ou d'un projet (Lewkowicz & Zacklad, 1999). Un forum de discussion peut donc être défini comme un document numérique dynamique et collectif.
>
> […] un forum de discussion instaure un cadre participatif complexe (Marcoccia, 2002) beaucoup plus proche de celui de la conversation que de celui de la relation éditeur-lecteur. [...] Un forum de discussion est de ce point de vue un document interactif (ou interactionnel) au sens strict, c'est-à-dire ne reposant pas sur l'opposition émetteur-récepteur. (*ibid.*)

La prise en charge de ces aspects documentaires dans l'écosystème des réseaux numériques (le Web principalement), à savoir non seulement leur fixation, mais leur description, standardisation, et insertion dans un jeu de métadonnées et de normes permettant de mieux chercher et trouver l'information en ligne, s'appelle la documentarisation (Pédauque, 2006). La documentarisation des documents numériques natifs (c'est-à-dire produits contextuellement avec des outils numériques), des échanges communicationnels aux contenus textuels et audio-visuels des pages Web, est devenu un enjeu de la recherche sur les environnements en réseau. Par le biais de services commerciaux (les moteurs d'indexation et de recherche de Google en tête) ou d'initiatives d'archivage privées (comme Internetarchive.org) ou publiques (les services nationaux d'archivage du Web comme, en France, à l'INA et à la BNF), on s'attache à présent aux « *nouvelles extensions temporelles* » (Froissart, 2005) ayant émergé de la conservation et de la remédiation des documents et données numériques pour mieux comprendre la manière dont les documents numériques traversent le temps comme support d'une mémoire technique et à travers des dispositifs d'archives.

L'extension temporelle représentée par les outils numériques supportant Internet hérite d'une cinquantaine d'année de conceptualisation de la technique informatique comme non seulement support, mais aussi amplification des capacités mémorielles, dans la lignée des penseurs de la cybernétique (Licklider, Engelbart). Mémoire exosomatique, son développement est lié à celui des technologies de stockage et de transmission de l'information, et son lien à la mémoire humaine serait rétroactif, selon

un mouvement de rétroaction propre à la technique[18]. Quel effet, alors, sur cette mémoire humaine, directement et indirectement, dans la perception de cette mémoire ? Pourquoi les penseurs de la mémoire technique parlent-ils de transformation des rapports du présent au passé ?

Nous avons avancé dans ce questionnement en nous rappelant le tournant qu'avait constitué les techniques d'enregistrement sonore dans la constitution d'archives orales pour l'ethnographie au début du 20ème siècle. Depuis, une réflexion s'est engagée non plus seulement sur les moyens techniques de recueillir les témoignages, mais sur ce que fait ce recueil technique aux témoignages eux-mêmes. La théorie interactionniste de Palo Alto, sous l'égide de l'anthropologue de la communication Gregory Bateson, a bien montré que l'observation et l'enregistrement modifiait, sous l'effet d'une rétroaction, les témoignages eux-mêmes.

Exportée à l'étude des communications médiées par le numérique, cette réflexion nous intéresse puisque nous observons les interactions en ligne – après qu'elles aient eu lieu et se soient donc inscrites dans des documents numériques fruit de l'enregistrement informatique des communications, et en ceci notre travail ne relève donc pas d'une méthodologie proprement ethnographique, davantage d'une anthropologie de l'écriture numérique. A partir de là on peut se demander comment intervient la mémoire technique dans la réflexivité possible des utilisateurs sur leurs pratiques et leurs outils de communication en réseau. Bien que ce champ d'interrogation ne soit que très peu défriché, des débuts de réponse ont été proposés selon différentes perspectives.

- La relation du collectif à la production et le maintien de son savoir à travers la mémoire technique ; dans une communauté réunie autour d'un outil de publication et de communication de contenus en ligne par exemple (les weblogs Wordpress), les archives des documents numériques produits par la communauté « *sont considérées comme la mémoire vivante de la communauté, de ses choix et de ses erreurs, et les nouveaux venus peuvent être invités à parcourir le passé d'une communauté pour apprendre, ce qui aurait une influence sur les modes d'interactions et les contenus ajoutés* » (Ruzé, 2011 : 6).
- La relation du collectif à lui-même face à l'enregistrement documentaire de ses échanges : la documentarisation des échanges d'un dispositif de communication synchrone de type « salon de conversation » ou *chatroom* (archivés au sein du dispositif et consultables par ses utilisateurs) permettrait l'observation *a posteriori* des échanges, fournirait une menace probatoire en cas de conflits, et permettrait à la communauté de se réguler dans le consensus symbolique (Pastinelli, 2009).
- La relation du collectif à un regard ultérieur et extérieur : face à une opération de ré-archivage par un tiers, comme Google l'a fait avec les communications Usenet, la perception des documents pose le risque de la décontextualisation des échanges, d'une confusion entre mémoire et histoire, et de leur réinterprétation par de nouveaux dispositifs de traitement et d'affichage de l'information (Paloque-Berges, 2013b).

L'intérêt croissant pour les archives du numérique doit ainsi s'interroger sur l'inflation des techniques d'enregistrement et des volumes d'information, et réfléchir sur ce que la mémoire numérique dit non seulement sur le passé, mais aussi sur le présent :

> Les historiens et philosophes de l'histoire, distinguant histoire et mémoire, s'interrogent d'ailleurs depuis plusieurs années devant le développement de ce que François Hartog (2003) a appelé le « présentisme », ce régime d'historicité dans lequel l'horizon d'attente perd du terrain au profit du présent, alors que le présent qui advient apparaît désormais de

18 « *La technique engendre sa propre évolution par le jeu de rencontres aléatoires entre des savoir-faire acquis – ce que nous avons appelé rétroaction combinatoire – et par la mise en mémoire de ces acquis sur des supports techniques – ce que nous avons appelé la rétroaction informationnelle* » (Lebeau, 2005 : 169).

> plus en plus comme le passé d'un futur qui est déjà en train (ou sur le point) d'advenir. Ce singulier rapport au temps se traduit par la multiplication des efforts visant à conserver l'empreinte de ce présent devenu histoire, ce qui se produit, paradoxalement, au moment même où on n'hésite plus à substituer le témoin à l'historien (Hartog, 2000), avec l'espoir de trouver dans la mémoire et l'expérience subjective de l'événement un surcroît d'authenticité qui semble faire défaut autant à la version objective du passé produite par l'historien qu'à celle qui est mise à plat dans les archives (voir aussi Todorov 1995). En somme, alors même que le présent est pensé et traité comme relevant de l'histoire et qu'on s'applique donc à l'enregistrer pour en conserver les traces, celles-ci se trouvent simultanément mises en concurrence avec la mémoire. (Pastinelli, 2009)

2.1.2. Des archives nouvelles : sources et documents numériques natifs

En ceci, l'idée de mémoire vivante comme source légitime des études en histoire des sciences et des techniques a fait son chemin depuis un certain temps, avec l'intégration de documents autrefois peu considérés dans les sources à côté des documents plus classiques relevant de l' « oeuvre savante » (Brian, 2001) : à l'impératif de chronologie érudite s'ajoute une volonté d'aller fouiller à la fois en profondeur dans les strates épigénétiques des écrits savants, mais aussi en extension, en allant consulter des documents produits dans les sphères dites profanes. En ceci, l'histoire des sciences a accepté dans ses sources l'analyse des archives de correspondances comme témoignage sur certains processus scientifique en train de se faire.

> Il reste encore à approfondir de nombreuses questions, notamment en encourageant un programme visant à enquêter systématiquement sur la genèse des dispositifs d'archives, terrain sur lequel les archivistes et les historiens ont beaucoup à apprendre les uns des autres, comme cela a été le cas dans le domaine particulier de l'histoire des sciences. En fait, c'est à chaque fois comme dans les exemples précédents, et à un niveau très micro-empirique et micro-analytique, la division du travail entre archivistes et historiens qui parait en profonde transformation (Brian, 2001).

Si les nouvelles technologies informatiques sont particulièrement prisées dans les SHS en général et en histoire en particulier en tant qu'outil de numérisation de sources (permettant leur préservation ainsi que de nouvelles méthodes d'analyse des données), peu d'attention a été accordée aux sources numériques natives, en partie parce qu'elles relèvent d'une histoire si récente qu'elle en est presque immédiate. Ces « *archives nouvelles* […] *excitant la curiosité* » posent des problèmes méthodologiques et techniques relativement nouveaux sur lesquels doit se pencher la réflexion historiographique, mais aussi d'autres disciplines en SHS quand elles ont besoin, même ponctuellement, d'une analyse historique de leur terrain, comme le propose (Ruzé, 2009).

Avant même d'être articulés aux méthodes d'autres SHS (dans notre cas les sciences de l'information et de la communication ou l'anthropologie), les problématiques historiques doivent s'interroger sur la nature des archives nouvelles, le repérage des sources les plus pertinentes dans une abondance de documents numériques natifs, ainsi que la réflexion méthodologique sur le traitement qualitatif non seulement des contenus, mais aussi des traces qu'ils laissent. Dans le contexte d'une étude sur les pratiques d'un collectif autour d'un outil en ligne, la question de la collaboration se pose comme centrale au niveau même de la matérialité même des documents qui supporte et conditionne la collaboration et l'organisation socio-technique du groupe. Dans le contexte de l'activité numérique, ces sources sont souvent « cachées », dans la mesure où il existe un dédoublement entre ce qui apparaît, même temporairement, à la surface des interfaces publiques (*front end*), qui relève d'une temporalité du résultat, et ce qui s'est passé dans les coulisses, au sein des interfaces d'administration, à l'accès

restreint à ceux qui participent à la production et l'édition documentaire (*back end*), et qui relève de la temporalité du travail ou du processus (avec les versions, modifications, etc.). Dans le cas des listes et groupes de discussion, ce dédoublement est différent puisqu'il ne s'agit pas de documents édités sur des pages Web. Cependant, l'accès aux archives des messages (en dépôt, pour les listes et groupes fermés, ou actualisées, s'ils sont en cours) révèle une vision d'ensemble des activités communicationnelles. Retrouver ces sources documentaires et les organiser pour l'analyse suppose ainsi un travail d'ingénierie historique qui rapproche le chercheur de l'archiviste : la récupération des documents, l'extraction de différents types d'informations (du contenu aux métadonnées) par dépouillement manuel ou instrumenté, et leur réorganisation pour l'analyse.

2.1.3. Des processus d'archivage semi-formalisés : une illusion d'archive ?

La première difficulté pour notre préparation du travail pratique a été de localiser des gisements de listes et groupes de discussion. En terme de matière documentaire, le problème s'est situé à mi-chemin d'une surabondance et d'une rareté des documents sources à portée de mains. En cause, les défauts des systèmes d'archivage présents sur le Web nous permettant l'accès aux documents pour mieux définir et analyser notre corpus.

En effet, les listes et groupes relevant de documents numériques natifs et étant gérés par des logiciels d'administration dédiés (SYMPA pour les listes, News software pour Usenet), il est relativement facile pour leurs administrateurs de générer des archives directement sur des interfaces Web. Un grand nombre d'organisations et d'institutions, en plus de services associés aux logiciels d'administration, ont ainsi mis à disposition des archives de listes en ligne. C'est le cas par exemple de l'organisation RIPE NCC (Réseaux IP Européens - Network Coordination Centre), un registre régional d'adresses IP pour l'Europe, qui offre l'accès aux archives des nombreuses listes qu'elle administre pour ses membres et maintient les archives de listes inactives pour des « *raisons historiques* »[19].

Dans le cas de listes françaises dédiées aux réseaux informatiques et Internet, nous avons eu davantage de mal à trouver des gisements variés, bien qu'il existe deux lieux d'archives de listes importants administrés par des organes de l'enseignement supérieur :

- le site du CNRS, sur une page intitulée « Service de listes de messagerie par CNRS/DSI »[20] ;
- celui de Renater (réseau informatique de la recherche en France), sur une page intitulée « Universalistes »[21].

Ces services, qui maintiennent un grand nombre de listes académiques, souffrent cependant d'un déficit de maintenance (avec de nombreuses erreurs de serveurs) ainsi que d'un moteur de recherche très limité. Ainsi, nous avons du mal à trouver les listes les plus anciennes correspond aux limites projetées de notre corpus. La liste « DNS » (dédiée aux discussions sur l'attribution des noms de domaine Internet, Domain Name Systems, et qui commence en 1989) a été trouvée sur « Universalistes », mais c'est en discutant avec Stéphane Bortzmeyer[22], un acteur de l'Internet français que nous avions beaucoup rencontré dans les groupes discussions Usenet et qui continue aujourd'hui à administrer la liste « DNS », que nous avons pu repérer la plus ancienne : il s'agit de la liste « IP », initiée par le groupe universitaire GERET (Groupe de travail Exploitation de Réseau Ethernet TCP/IP) en 1989.

Dans le cas des groupes de discussion français, c'est grâce au service Google Groups que nous avons pu accéder à une multitude d'archives. Ce service ouvert en 2001 est d'abord dédié à la création et à la gestion par les utilisateurs de groupes natifs Google. Il abrite aussi une très large archive des

19 [https://www.ripe.net/ripe/mail/inactive-lists]
20 [https://listes.services.cnrs.fr]
21 [https://groupes.renater.fr/sympa]
22 Entretien avec Stéphane Bortzmeyer le 25 octobre 2012.

groupes Usenet récupérée auprès du département de Zoologie de l'Université de Toronto dont les administrateurs réseaux avait constitué une archive remontant à 1981[23], et auprès de la compagnie Deja News après le rachat de son service de forum, Deja communities, par Google. Deja News fournissait un accès Usenet payant depuis 1996, et offrait à ses abonnés un accès aux archives, première initiative de grande ampleur pour conserver la mémoire de Usenet. Google a donc redocumentarisé ces archives pour les mettre à disposition sur le Web comme une sous-partie de son service de forums. L'interface de Google Groups, cependant, présente de nombreux défauts qui rendent la recherche fastidieuse, notamment parce que les résultats affichés par son moteur de recherche ne discriminent pas les groupes natifs Google des groupes historiques Usenet, en sus d'une myriade d'autres liens vers des forums externes partout sur le Web. Nous avons finalement trouvé des instructions pour réduire les choix : la commande « fr.* » permet de n'afficher que les groupes de la hiérarchie francophone (mais il faut savoir que cette hiérarchie commence par ce préfixe). L'avantage des Google Groups réside dans la possibilité d'afficher le « message original », et donc d'avoir accès à l'intégralité des données et métadonnées du message (que nous décrirons plus bas). Cependant, une série de modifications de l'interface au cours de ces dernières années, ayant abouti à une version stabilisée en juin 2013, a rendu la recherche au sein des groupes très difficile. Nous avons étudié les problématiques formelles, techniques et mémorielles soulevées par l'initiative Google Groups et avons souligné les contradictions de la mission que le géant du Web s'était assignée, ne permettant pas une lecture compréhensive des archives pour deux raisons majeures (Paloque-Berges, 2013b) :

- une édition défaillante du système d'archives, créant des situations d'illisibilité générées par une mauvaise éditorialisation sémio-technique des interfaces de rendu des résultats (confusion dans les résultats, interface manquant d'intuitivité voire d'usabilité, nombreux bugs, impossibilité de communiquer avec les équipes de maintenance...) ;
- des archives incomplètes, en raison de la difficulté initiale de rassembler un corpus gigantesque ; ce à quoi s'ajoute un service spécifique de demande de retrait des messages passés par leurs auteurs originaux au nom d'un droit à l'oubli, entretenant une ambiguïté quant au motif de l'archivage, entre devoir de mémoire, initiative patrimoniale et documents historiques, Google ayant communiqué sur les trois plans.

Une fois l'outil maîtrisé malgré les obstacles, nous avons pu naviguer dans les groupes concernant notre corpus. La hiérarchie fr.* a été créée sur Usenet en 1993 grâce aux efforts de Christophe Wolfhugel[24]. Usenet, service de communication distribué et géré par ses utilisateurs, a l'avantage de documenter ses processus d'évolution et de gouvernance au sein même des communications échangées. C'est ainsi que nous avons pu accéder à une liste des groupes fr.* originaux, impossible à trouver via le moteur de recherche de Google Groups, dans un message de 2012 retraçant l'historique de la hiérarchie[25]. A partir de là, nous avons choisi les groupes qui nous ont semblé les plus pertinents pour notre étude.

En définitive, le rappel de ce parcours tortueux pour accéder aux archives en ligne nous ayant permis de constituer notre corpus, s'il peut sembler anecdotique, révèle en fait par l'exemple la situation complexe dans laquelle se trouvent les archives de listes et groupes de discussion sur le Web. Immense chantier où prolifèrent les sources, il hérite de la succession d'états transitoires et informels de l'archivage numérique natif, dû au mode même du développement de la documentation sur Internet,

23 Archivée par ailleurs sur Internet Archive [http://archive.org/details/utzoo-wiseman-usenet-archive].

24 Voir les messages appelant la communauté francophone à voter pour permettre la création d'une hiérarchie dédiée, entre le 12/12/1992 et le 07/01/1993 [https://groups.google.com/forum/#!searchin/fnet.general/une$20hierarchie$20de$20news$20pour$20la$20communaute$20francophone$20$3F].

25 Message intitulé « [DOC] Presentation de la hierarchie FR » envoyé le 02/09/2012 sur le groupe fr.usenet.divers [https://groups.google.com/forum/?fromgroups=#!search/fr.reseaux.telecoms/fr.usenet.divers/YtXV9amK4k8/OcRgCnhPAGgJ].

pour laquelle les efforts de formalisation et de standardisation fournis par les archivistes professionnels est à la fois tardif et constamment dépassé par l'introduction de nouvelles techniques d'édition numérique. Les modes (plutôt que les méthodes) d'archivage Web de documents numériques natifs encapsulent des documents dans de nouvelles interfaces, encapsulation qui peut se révéler fort problématique si elle ne prend pas en compte un nombre très important de paramètres. Les systèmes d'archives en ligne, munis de moteurs de recherche plus ou moins performants, sont des machines à remonter le temps qui modifient le rapport à la mémoire et à l'histoire : ils sont ainsi une nouvelle traversée d'un espace informationnel passé et présent.

Lors d'une communication intitulée « Problèmes de fouille dans l'archéologie du web social : anciens corpus d'Internet archivés sur le Web » (Paloque-Berges, 2013f) où nous avons présenté notre travail de post-doctorat en cours, nous avons discuté de ces « effets de simulation » créés par les dispositifs documentaires d'archivage de documents numériques natifs, qui génèrent une virtualisation au second degré de documents nés immatériels, et en avons conclu qu'ils constituaient l'un des problèmes cruciaux pour l'analyse de matériaux numériques. Le Web, et plus généralement Internet, a ainsi pu nourrir le fantasme documentaire de beaucoup, mais ne présente en fait qu'une illusion d'archive loin des modèles formels nécessaire au travail de l'historien et des chercheurs en SHS. C'est une des raisons majeures pour lesquelles nous avons dû effectuer au cours de notre travail un travail de récupération des archives et de formalisation documentaire afin de mieux comprendre comment les exploiter à la fois dans un cadre d'analyse scientifique et de préservation patrimoniale (partie 3).

2.1.4. Paroles, bruit et infobésité : le superflu des sources de la communication électronique

L'une des difficultés majeures dans l'analyse des communications électroniques est leur statut discursif informel et proliférant, leurs contenus (extraits des corps de message) appartenant aux registres de l'opinion et de l'anecdotique et faisant courir le risque à l'historien de se perdre dans les détails d'une parole sans fin. Nous montrons comment ces contenus de communication doivent être problématisés à l'articulation du discours individuel, de la construction collective et de la réalité historique.

Nous avons commencé notre recherche en considérant que les archives de communications électroniques seraient un terrain intéressant dans l'optique d'une micro-histoire des réseaux informatiques. Nous avons donc défendu un intéressement scientifique à la parole électronique médiée par les réseaux (sous la forme de courriels collectifs). Cependant, dans une économie d'abondance et de prolifération telle que celle des échanges en réseau, comment s'y retrouver ? Est-ce que tout se vaut ? Lors de la présentation de nos travaux au cours du séminaire des jeunes chercheurs du LabEx HASTEC (12 avril 2013), il a été évoqué la très grande difficulté des archivistes aujourd'hui face aux correspondances courriels. Philippe Hoffman nous a d'ailleurs fait part de sa pratique professionnelle de tri : si la documentation traditionnelle est classée, rangée et conservée, tout ce qui relève de la communication électronique tend à être mis au rebut, considérée comme un ensemble d'informations où il est trop difficile de démêler l'intéressant du « bavardage » et autres « bruissements » du langage. La discussion s'est orientée sur la question des à-côtés des textes et documents traditionnellement pris en charge par l'historien. N'est-ce pas vain de vouloir tout embrasser dans l'analyse, fantasme d'une reconstruction exhaustive du réel par l'historien ? Il faut tout d'abord essayer de mieux qualifier le discours produit en contexte électronique pour avancer sur cette question (nous revenons plus bas de manière plus spécifique sur les méthodes d'analyse des langages du courriel).

Ces archives proposent une version documentaire d'interactions relativement informelles qui pourraient être assimilées à des conversations (médiées par les réseaux). Mais derrière cette informalité de surface, c'est de l'organisation des relations sociales dans une communauté d'écriture qu'il s'agit

(Labbe et Marccocia, 2007). Plus encore, précise le sociologue Philippe Hert cette communauté se reconnaît au travers des effets de réel dérivés de l'informalité de la communication (Hert, 1999, 1996) : l'oralité, mimant le réel (l'immédiateté et la transparence des échanges) permet au groupe d'interlocuteurs de se reconnaître en tant que une communauté en la justifiant comme lieu (électronique) de débat ouvert et sans entrave. Hert propose l'idée que celui qui lit et analyse ces échanges devrait au contraire s'attacher à la dimension écrite de ces échanges pour comprendre « *la création d'effets de sens qui émanent de la matérialité du texte. Cette matérialité renvoie à la capacité structurante du texte, sa stabilité, qui obligent le lecteur à entrer dans un travail* » (Hert, 1999: 103).

Replacée dans une problématique impliquant des questions mémorielles et historiques, cette « parole » organisée dans une communication en réseau est une inscription d'une part (les traces de la communication en réseau révélée dans l'analyse des couches infrastructurelles et formelles des techniques de transmission, comme décrit plus haut), et un document d'autre part, si les communications électroniques ont fait l'objet d'archives (mêmes non formalisées). L'effet d'informalité de la parole en réseau est à la base du paradoxe qui structure les réseaux électroniques déjà évoqué : la spontanéité des échanges étant fixée dans les documents numériques (Salaün, *art. cit.*). Devant cette généralisation de ce qui a été appelé la « documentarisation » (Pédauque, 2006), comment échappe-t-on à la prolifération des discours et des données ? Doit-on y échapper ?

Nous avons étudié la réception de publics d'Internautes vétérans face à la mise en archive par Google de leurs communications passées sur Usenet en 2001 (Paloque-Berges, 2013b). Le dispositif de ces archives, rapidement décrit plus haut, implique :

- un ensemble d'archives des interactions électroniques (gardant leur structuration originale selon leur inclusion dans des groupes thématiques, et présentés sous une forme antéchronologique) ;
- une interface qui recrée un encadrement documentaire (une redocumentarisation) dans le contexte des technologies Web pour afficher des groupes de discussions Usenet ;
- un moteur de recherche ;
- une chronologie superposant des événements historiques de la « petite histoire » de Usenet et d'Internet (premier appel à projet pour la collaboration open source autour du système Linux, par exemple), à la « grande histoire » mondiale (première annonce sur Usenet de l'explosion de Tchernobyl, par exemple).

Le public qui découvre ce dispositif d'archives sur les Google Groups a montré une tendance à s'y intéresser pour des raisons mémorielles, un des premiers réflexes étant la « recherche-ego », à savoir chercher ses contributions personnelles passées à la communication collective sur les réseaux. Mais cette recherche de son soi passé de réseau est insérée dans les autres dimensions des archives Usenet : la mémoire collective (les référents techniques, sociaux et culturels créant des effets de communauté), et l'événement historique (ou plus exactement, la médiation des informations à propos d'événements susceptibles de marquer l'histoire), dimensions mises en perspective dans chaque recherche-ego, pour mieux se situer à leur articulation. Le « j'y étais » du témoin se reformule ici dans un « on en a parlé » (« nous qui communiquions déjà sur les réseaux numériques à l'époque »).

Cette situation montre bien comment l'utilisateur des réseaux du début des années 2000 a commencé à accepter l'idée qu'il était soumis, en tant qu'identité numérique en tout cas, à la documentarisation généralisée. Pour reprendre une heureuse expression d'Olivier Ertzscheid, l' « *homme est un document comme les autres* » (Ertzscheid, 2009). Certains célèbrent cela, et travaillent à retrouver des éléments de la mémoire collective qui pourrait devenir le terreau d'une socio-histoire des usages des réseaux numériques. D'autres y résistent, invoquant un droit à l'oubli, avatar d'une *damnatio memoriae*, et réclament des dispositifs d'effacement de leurs contributions passées (que Google implémentera sous la forme d'un service de suppression de l'affichage des messages d'individus

sur demande de ceux-ci). Est alors en jeu non seulement un rapport à l'identité qui structure la mémoire historique entre mêmeité ou ipséité, selon la dualité proposée par Ricoeur, mais aussi un rapport d'usage aux dispositifs qui encadrent et donnent une forme aux données du passé. Qui donne cette forme (qui documentarise) ? Le fait que cette initiative d'archivage soit portée par Google présente un cas intéressant de documentarisation par une entreprise qui se définit précisément par une mise en document numérique du monde.

Référence qui nous a été indiquée pendant le séminaire des jeunes chercheurs d'Hastec, l'ouvrage de Patrick Geary, *Mémoire et oubli à la fin du premier millénaire* (Geary, 1996) montre que cela n'est pas une problématique nouvelle : il étudie comment des opérations d'archivage (entre sélection et mise à l'écart) par les moines à l'ère médiévale sont aussi une forme de manipulation de la mémoire visant à rabattre des enjeux politiques présents sur l'écriture du passé (et vice versa). Le regain d'intérêt des quinze dernières années pour l'objet archive (Chabin, 2000), propulsé par les capacités vastes de stockage et de traitement des données numériques, doit être également pensé en termes politiques. Cela permettrait en effet de mieux comprendre des initiatives d'archivage numériques comme celle du réseau social Twitter par la Library of Congress américaine depuis 2009, qu'Olivier Ertzscheid qualifie de « *patrimoine superflu* » (Ertzscheid, in Barats, 2013 : 69-72[26]).

Si notre approche des archives est avant tout méthodologique et pratique, elle ne sert cependant pas à nourrir un fantasme d'un « tout archivage » dans la ligne droite de l'utopie d'Internet célébrant l'archive universelle, et reste consciente de l'illusion d'archive que représentent les réseaux numériques. La critique de l'archive universelle n'est pas notre objet d'étude, mais il semble que se pencher sur des cas précis d'archives numériques permette de mieux comprendre les problématiques de pouvoir qui sous-tendent la conservation de la parole électronique, ceci en prenant l'archive comme à la fois objet permettant l'analyse et objet à analyser.

2.2. Faire parler les documents numériques natifs et leurs archives

Nous revenons ici sur les grandes orientations méthodologiques en SHS repensées pour l'analyse et l'interprétation des données contenus dans les documents numériques. La recherche en SHS se positionne depuis quelques années sur une tendance à réévaluer les apports de la méthode quantitative à l'aune de la prolifération des données numériques, et nous montrons quelques uns de ses apports. Nous convoquons, dans le cadre de notre étude, une approche qualitative que nous considérons complémentaire de l'approche quantitative. Cependant, les données numériques ne disent pas tout.

Pour pallier ce problème, nous avons choisi de compléter notre investigation par une série d'entretiens avec les acteurs de cette histoire que nous cherchons à reconstituer. Travailler sur l'histoire récente des technologies de réseau présente un atout en ceci qu'elle n'est vieille que d'une quarantaine d'années, la majorité de ses acteurs étant encore vivants. Nous verrons pour chaque limite les possibilités offertes par le recours aux témoignages d'acteurs, ainsi que les nouvelles limitations que ceux-ci peuvent à leur tour poser.

2.2.1. De nouvelles relations entre méthodes qualitatives et quantitatives

L'analyse de grands ensemble de données favorisée par les technologies numériques relève d'une facilitation du traitement de ces données par les instruments logiciels de calcul. Le terme « Big Data » traduit ce nouvel intérêt pour la prise en charge quantitative des données numériques natives. Si l'histoire sociale avait déjà adopté les méthodes quantitatives de la sociologie, les historiens dans leur

26 cf. aussi la contribution, dans le même ouvrage, de Ertzscheid, Gallezot et Simonnot, « A la recherche de la « mémoire » du Web : sédiments, traces et temporalités des documents en ligne », pp.53-68.

ensemble sont aujourd'hui concernés par les processus de collection, de gestion et d'accès aux sources de grande ampleur générées par les usagers des outils numériques. Plus généralement, ils doivent s'interroger au cours d'une historiographie réflexive sur la tendance actuelle à ce que Frédéric Clavert appelle une « *mise en données du monde* »[27].

A. Approches quantitatives : l'analyse de réseau

L'analyse de réseau en particulier, héritée des méthodes statistiques et mathématiques de la sociométrie et utilisée prioritairement en sociologie ou en économie, trouve sur les réseaux socionumériques des sources abondantes pour mettre au jour les positionnements d'acteurs (usagers du numérique) dans un champ informationnel ou communicationnel, à travers l'étude des relations entre individus et groupes et de leur agrégats relationnels. Tout au long du séminaire « Autorités calculées : écritures en réseau, systèmes techniques et faire autorité » (co-dirigé en 2012-1013 avec Evelyne Broudoux au laboratoire DICEN du CNAM), nous avons exploré les nouvelles épistémologies de ce qui, appliqué aux données relationnelles des réseaux d'Internet en général et du Web en particulier, est devenu la « Web science ». La « science du Web », héritant de la sociométrie, applique ses modèles aux données du Web, en particulier les liens entre unités d'information (page Web, contenus multimédia gérés par les applications, profils utilisateurs...). Elle propose une perspective informationnelle (systèmes d'information et leur traitement algorithmique des données) pour répondre à un problème de communication stratégique : qui (se) communique le mieux dans les environnements socionumériques et comment. Ce séminaire a ouvert la voie à des interrogations de chercheurs en Sciences de l'information et de la communication autour de la relation entre les techniques numériques, les environnements des réseaux informatiques et les logiques d'écritures qu'elles supportent et transforment. La mise au jour de nouvelles formes d'autorités qui s'appuient sur des calculs algorithmiques (F. Ghitalla, A. Lauf), sur la la capacité de délégation des décisions et de l'organisation de l'information au logiciel (B. Rieder), sur la transitivité relationnelle des individus dans les réseaux socio-numériques (L. Merzeau) a guidé ce séminaire. Face aux autorités traditionnelles héritées de l'imprimé, la structuration et la circulation du savoir médié par les réseaux numériques engagent une nouvelle manière de conceptualiser ce qui fait autorité.

Le courant des Humanités numériques, présent au sein d'HASTEC dans plusieurs programmes collaboratifs, s'intéresse lui aussi à ces modèles. Le fait que sa focalisation se fasse prioritairement sur des sources analogiques numérisés n'empêche pas que les nouvelles méthodes d'analyse de réseau soient applicables à des données non nativement numérique, comme en a témoigné la séance du séminaire Digital Humanities intitulée « Analyse de réseaux et Digital Classics »[28].

B. Anciens et nouveaux enjeux du qualitatif

Notre approche n'est pas quantitative, mais nous avons dû prendre la mesure de ces méthodes dans la mesure où, comme le souligne Frédéric Clavert, elles entraînent un changement du regard qualitatif. En effet, et cela a été une discussion récurrente au cours des séances du séminaire « Autorités calculées », l'approche qualitative des sources sur Internet reste cruciale en complément du calcul des quantités de données, et ce pour plusieurs raisons. Cette approche s'intéresse à l'analyse des données au plus près du terrain, des acteurs, ainsi que des relations non plus seulement au niveau des réseaux, mais aussi à d'autres types d'activités hors-réseau et des évolutions des pratiques numériques en regard avec des pratiques pré-numériques. Comme le suggèrent (Markham and Baym, 2009), le regard qualitatif a

27 Frédéric Clavert, « Mise en données du monde, mise en données de l'histoire », billet de blog publié le 12/07/2013 http://www.clavert.net/mise-en-donnees-du-monde-mise-en-donnees-de-lhistoire/

28 Séance du 24 avril 2013 du séminaire « Digital Humanities », dirigé par A. Berra, M. Dacos et P. Mounier [http://philologia.hypotheses.org/1119] et [http://www.ehess.fr/fr/enseignement/enseignements/2012/ue/324/].

plusieurs vertus dans son rapport aux données d'Internet :

- il relativise la tendance positiviste de la recherche sur le numérique à proclamer la nouveauté irrémédiable des environnements technologiques ; il inscrit la recherche dans une histoire des résultats en SHS, et permet les comparaisons, par exemple l'utilisation du concept de « communauté de discours » (*speech community*) proposé par l'ethnographie de la communication, pour l'étude des communautés électroniques de discussion ;
- il permet de cadrer la recherche, ou plus précisément de la focaliser explicitement sur ce que l'on ressaie de comprendre ;
- il déconstruit les catégories à partir desquelles on peut avoir tendance à penser un objet recherche, en particulier dans le cas des « études sur Internet » (*Internet studies*), pour lequel « Internet » est à la fois un sujet de réflexion et un objet qui échappe si on essaie de lui garder une intégrité de surface : « *d'une part, cela veut dire qu'il faut comprendre l'architecture des éléments d'Internet étudiés en les comparant avec d'autres éléments étudiés par d'autres. D'autre part, cela implique de chercher et prendre en compte les interconnections entre l'Internet et la 'vie réelle' dans laquelle son usage prend place et qui participe à sa construction* » (Markham and Baym, 2009 : 288) ;
- il engage à anticiper les contre-arguments sur la recherche en termes d'interprétation et de représentativité des résultats, en mettant en perspective les problématiques (ce qu'on cherche) avec le terrain et le corpus (où l'on cherche, dans quelles limites, selon quels choix).

Les ambitions universalistes des méthodes quantitatives appliquées aux données numériques à travers le Big Data (l'embrassement de grands ensembles de données totalisants) sont accompagnées d'un renouveau des débats épistémologiques sur la vérifiabilité et la reproductibilité des analyses scientifiques. Les grands travaux d'équipement et d'infrastructure de la recherche, très sensibles en France notamment, encouragent ainsi à partager les corpus dans ce but – notre travail pratique, décrit en 3, s'est ainsi intéressé à cette dynamique de partage des corpus. L'approche qualitative peut participer à cette dynamique de partage en fournissant des corpus décrits dans toute la finesse de leurs données. Mais si elle recèle ce potentiel de coordination de la recherche, elle doit s'affranchir de la quête de généralisation, un « *concept qui assume un monde stable et réplicable dans lequel un ensemble de significations prévaut sur les autres* »[29] que porte souvent l'approche quantitative, pour supporter plutôt l'exploration, l'articulation et la comparaison de différents domaines (Markham and Baym, 2009 : 292).

2.2.2. Autres archives, autres témoignages : de l'utilité de « documents » non numériques

Considérer les communications électroniques comme sources légitimes pour faire l'histoire des réseaux ne peut se faire sans avoir recours à d'autres sources pour vérifier, croiser, clarifier les informations recueillies. De fait, au cours de notre travail, nous avons mené une série d'allers-retours entre les archives numériques et d'autres types de sources, à savoir, des archives institutionnelles et des témoignages d'acteur qui nous ont permis de préciser la valeur de sources des communications en réseau.

Nous avons choisi quelques acteurs apparaissant de manière récurrente dans les listes et/ou groupes de discussion de nos corpus, ou alors par recommandation de collègues[30]. Notre premier interlocuteur, Stéphane Bortzmeyer[31], ingénieur informatique, administrateur de la liste DNS (faisant

29 « "*generalizability" – a concept that assumes a stable replicable world in which one set of meanings prevail* ».

30 Merci à Valérie Schafer et Hervé le Crosnier, ainsi qu'aux acteurs interrogés, pour nous avoir ouvert leurs carnets d'adresse respectifs.

31 Entretien réalisé le 12 novembre 2012.

partie de notre corpus), mais aussi participant à l'IETF (organisme s'occupant de définir collaborativement les standards de l'Internet, en discutant notamment par le biais du courrier électronique) était omniprésent à la fois sur les listes académiques et les groupes de discussion Usenet, nous a donné à voir un premier panorama des acteurs impliqués dans le développement des réseaux informatiques en France. Il a notamment confirmé le rôle prépondérant des documentalistes dont nous avions l'intuition, ce qui ouvre de très intéressantes perspectives de recherche pour le futur[32]. De manière plus générale, il nous a aussi éclairé sur les différences de culture technique entre les listes (plus académiques, plus sérieuses) et les groupes (plus ouverts, plus ludiques, « *à la cantonade* » dira Yves Devillers, un autre des acteurs interrogés). Bortzmeyer lui-même, entré au CNAM comme ingénieur réseau en 1991, nous a révélé un pan de l'histoire d'Internet très peu connu, à savoir le rôle du département Informatique de l'établissement dans l'ouverture des toutes premières connections internationales de réseaux informatiques communicants en 1983. Liaison passant par le Centre de Mathématiques d'Amsterdam (aujourd'hui nommé CWI[33]) pour récupérer le courrier électronique et les news (Usenet) mis en place par les spécialistes des systèmes et réseaux Unix, elle a précédé la première connexion de type TCP-IP (la suite protocolaire à la base de l'Internet tel que nous le connaissons aujourd'hui) entre l'INRIA et les Etats-Unis en 1988, et lui a pavé la voie. Bortzmeyer nous a mis sur la voie des acteurs ayant été impliqués au cours de la décennie suivante dans ce développement, et nous avons réussi à interroger Yves Devillers, l'un des initiateurs des réseaux Usenet / Internet en France, longtemps ingénieur réseau à l'INRIA, Annie Renard, sa collaboratrice à l'NRIA à la fin des années 1980 et au début des années 1990, spécialisée dans la gestion des noms de domaine Internet, ainsi que Laurent Bloch, ingénieur réseau au CNAM à la même période[34]. Ce choix d'acteurs et témoins a ainsi déterminé la focalisation historique de notre travail, qui a abouti à l'écriture de l'article « Between electronic frontier and electronic agora: the role of Unix computer networks in France and Europe in the promotion of Internet's technologies and values as a technical democracy » (Paloque-Berges, 2013d).

L'apport incontestable de ces témoignages a été d'une part de nous aider à mieux comprendre les différentes réalités techniques en jeu dans ce développement complexe, d'autre part de fournir une contextualisation des enjeux socio-professionnels derrière ces réalisations techniques. En particulier, nous avons pu mieux comprendre une série de pratiques d' « *utilisations non prévues* » (Bortzmeyer) des techniques d'informatique de réseau au profit des réseaux communicants dans le cadre de ces connections pionnières tout au long des années 1980. Évoquées par l'historienne des réseaux informatiques français Valérie Schafer comme des pratiques de « bricolage », ces dernières se sont avérées utiles à détailler dans le cadre d'une recherche sur les communications électroniques, puisque c'est au niveau du détournement de protocoles de transmission pour l'envoi d'information gérées par applications mail que ces utilisations non prévues trouvent leur originalité. Plus encore, ces détournements illustrent avec force les conflits qui ont à la fois ralenti l'arrivée d'Internet en France mais l'ont aussi préparé, « *au prix de négociations, hésitations, controverses techniques complexes* »

32 Nous avons envisagé un temps d'inclure la liste Biblio-fr, très populaire dans les années 1990 et citée par de nombreux acteurs, ainsi que ses pendants sur Usenet, fr.doc.biblio et fr.doc.divers ; nous avons aussi rencontré Hervé le Crosnier, son fondateur. Mais pour des raisons de recentrage thématique sur les ingénieurs spécialisés en informatique de réseau, mais aussi parce que cette liste, l'une des seules archivées par la BNF, présentait de nombreux problèmes pour son exploitation (aux niveaux technique et juridique), nous l'avons laissé de côté. Si l'influence de l'imaginaire documentaire sur les technologies de l'information numériques est indéniable et reconnu, il reste encore à écrire sur le rôle des spécialistes de la documentation dans le développement d'Internet. Cf. par exemple le rapprochement entre le projet de documentation universelle le Mundaneum avec les technologies d'Internet et du Web [http://expositions.mundaneum.org/fr/expositions/renaissance2.0-fr] et (Schafer, 2013).

33 Centrum voor Wiskunde en Informatica (Centre de Mathématiques et de Science Informatique).

34 Entretiens réalisés respectivement les 12 décembre 2012, 4 juillet et 12 juin 2013.

(Schafer, 2012 : 80-81) entre les milieux des sciences et ingénieries de l'information et les politiques. Notre étude sur le déploiement des réseaux Unixiens en France, prolégomènes à l'adoption des protocoles de l'Internet, analyse et retrace l'une des trajectoires relevant de ces problématiques où des choix techniques se discutent de manière externaliste ; dans notre cas, c'est le paysage d'une politique nationale d'équipement et de reconnaissance hésitante des réalisations techniques marginales qui se dessine dans le récit des acteurs en opposition avec la cartographie internationale conquérante de la frontière électronique. Ce récit d'acteurs, pionniers et militants en faveur de l'expansion de la frontière électronique, a dû cependant être mis à l'épreuve d'autres sources. Le recours aux archives du département informatique du CNAM nous a permis de trancher un certain nombre d'hésitations quant à l'origine précise des premières connections Internet internationales, les différents acteurs interrogés étant hésitants sur ce point.

Le recours à des archives numériques Usenet très anciennes, remontant à 1982-1983, atteste également des premières connections comme partant du CNAM, et bientôt de l'INRIA et de l'IRCAM, mais ne montrent aucunement, à la différence du récit d'acteurs (et, dans une moindre mesure, des archives du laboratoire), les tensions socio-professionnelles et politiques ayant accompagné cette histoire. Il faut attendre l'ouverture des branches francophones, ainsi que de listes académiques où chercheurs et ingénieurs en informatique discutent dès la fin des années 1980, pour voir éclore des débats entre utilisateurs attestant « en direct » de ces tensions. A titre exemple, on pourra citer l'organisme Fnet, premier fournisseur de connections à Internet auprès de la communauté des scientifiques et ingénieurs de réseau des laboratoires publics et privés dès 1983, qui catalyse un grand nombre de critiques et de conflits relevant de la question de l'administration des réseaux (les premiers débats ayant mené à des problématiques de plus grande ampleur sur la gouvernances des réseaux numériques). L'enjeu des débats autour de Fnet est lié à l'augmentation substantielle de demandes de services réseau par de nouveaux clients et donc de la multiplication des problèmes de gestion du trafic de données réseaux[35]. Les acteurs témoignant des problèmes de Fnet font état de l'émergence d'un mécontentement chez ces clients relevant d'un manque de compréhension de l'expertise d'administration des réseaux ; c'est d'ailleurs le même type de problème qu'ils auront rencontrés face aux administrations dans lesquelles les membres de Fnet travaillent, le CNAM et le CNRS en premier lieu[36]. Si ces « témoignages de l'intérieur » nous renseignent sur la difficulté de l'ingénieur réseau à faire son travail dans un contexte où il n'est pas légitime, manquant de moyens, de soutien et reconnaissance pour ses innovations (les acteurs ayant tendance à se dépeindre comme héros de l'ombre, sans emphase cependant), la confrontation avec les archives des groupes fr. Usenet à partir de 1993 nous permet d'aller un peu plus loin dans la compréhension de ces enjeux. En effet, le début des années 1990 marque la naissance d'un proto-marché de fournisseurs d'accès à Internet introduisant une concurrence dans les prix de connexion aux services de réseau, alors que les règles d'administration des réseaux ne sont pas encore fixées. D'autre part jusque là Fnet avait une sorte de monopole officieux sur l'activité de gestion des réseaux Internet en France sans que son statut soit clairement défini pendant longtemps. La transition d'un mode d'opération informel efficace seulement dans un contexte de niche (quelques laboratoires d'informatique universitaires et d'entreprise) à la nécessité de structurer l'administration et les offres d'accès semble être au cœur des problèmes rencontrés par Fnet ; en ceci,

35 Cf. un courrier de Yves Devillers intitulé « grave pb de congestion sur corton [Action URGENTE necessaire] », et envoyé sur le groupe fnet.general le 29/01/92.

36 L'aventure de la première connexion Usenet du CNAM prend fin avec la démission de Humberto Lucas, directeur du laboratoire informatique (qui pendant longtemps sera aussi le service informatique de l'établissement). Unixien et initiateur du projet, il fait face au manque de reconnaissance et de soutien des autorités pour le projet et aux nombreuses complaintes des utilisateurs du réseau à chaque problème sans qu'ils ne comprennent ce qui était en train de se faire dans le laboratoire (tel que raconté par Yves Devillers et Laurent Bloch).

les discussions sur Usenet de la communauté réunie autour de Fnet permettent de relativiser certains éléments des témoignages.

On voit avec cet exemple les limitations rencontrées face au déchiffrage des conversations de réseau, relevant de conflits d'opinion parfois peu lisibles, mais aussi face aux témoignages d'acteurs nécessairement impliqués dans la réinterprétation de leur activité passée. Leur recoupement permet de prendre de la distance, et de contextualiser de manière plus large les conflits ayant accompagné l'évolution technique, mais aussi sociale, politique, puis économique des réseaux informatiques en France.

2.3. Messages électroniques et communication en réseau : méthodes d'approche et d'analyse

Maintenant que nous avons indiqué quelques directions méthodologiques générales prenant en compte le temps et l'espace de la mémoire technique ainsi que des considérations anthropologiques de terrain, nous nous pencherons au plus près des matériaux que nous portons à l'analyse.

Bien que la dimension collective des échanges électroniques entre ingénieurs de réseau soit l'objet principal de notre recherche, on a dû consacrer beaucoup d'attention à l'objet « courrier électronique », ou courriel. En effet, il constitue l'unité documentaire fondamentale sur laquelle reposera non pas seulement l'analyse, mais le travail pratique de préparation à la préservation et à l'analyse, qui nécessite de dégager des normes pour les intégrer aux modèles de structuration du corpus.

2.3.1. Le courrier électronique comme objet de langage

La communication électronique a relativement tôt attiré l'attention des sciences du langage, d'abord en tant que nouvelle appréhension du discursif (un discours médié matériellement par les logiciels en ligne), du communicationnel (posant des problèmes d'interaction et de formes langagières), et du textuel (puisque les formes et contenus de l'échange sont inscrits dans un document numérique en contexte de réseau). Nous préciserons ici certaines des avancées permises par cette discipline dans la compréhension de ces objets d'analyse tout en nuançant leurs propositions en fonction de notre travail.

A. Les « discours sur Internet » : entre oralité et écriture, la place du logiciel

(Mourlhon-Dallies, Rakotonoelina et Reboul-Touré, 2004) définissent le « discours sur Internet » en rapport avec son support et le préfèrent aux termes « communication » et « message », qui marquent le regard disciplinaire des sciences de l'information et de la communication, ou encore « genre » et « texte », qui marquent la méthode d'analyse littéraire structuraliste. Notre regard étant interdisciplinaire et compréhensif, nous n'écartons nous-mêmes pas ces autres notions car selon le contexte de la mise en discours médiée par les réseaux, elles peuvent venir enrichir l'analyse. A titre d'exemple, il a été important pour nous d'invoquer la terminologie littéraire pour analyser la création d'un folklore de la culture d'Internet exprimée selon des formes intertextuelles tissées au sein même des groupes de discussion Usenet (Paloque-Berges, 2011a et b). Cependant, nous sommes d'accord avec les chercheurs cités à propos du mode de désignation par défaut des discours par le biais du nom générique donné au logiciel supportant matériellement le discours : listes de diffusion, groupes de discussion, ou encore forum Web, et « salon de conversation » (chatrooms) pour les échanges instantanés.

A propos, plus spécifiquement, du courriel, l'une des qualités du regard linguistique est de ne pas postuler a priori que le courriel relève d'une radicalité nouvelle, ou tout au moins d'une dérivation des genres de l'oralité dont il serait une forme conversationnelle appartenant à la famille des conversations en face à face et des correspondances traditionnelles et d’entreprise, mais dont l'interaction se situe dans la médiation électronique de manière informelle (Anis, 1999). Si l'on peut

parler de conversation de manière générale, cela devient difficile d'utiliser ces termes pour une qualification générique. En effet, la généralisation de l'adoption du courriel et l'évidente évolution des conventions de l'échange que son usage traduit « *symbolise*[nt] *peut être la manière dont les relations sociales sont redéfinies, particulièrement dans les organisations* » et peut avoir plutôt tendance à marquer l'appartenance à une « *communauté d'écriture* » que de parole (Labbe et Marcoccia, 2007).

En effet, il est important de prendre en compte le format spécifique, contrainte par le dispositif technique, dans lequel l'échange courriel (interpersonnel ou collectif) prend place. Dans le cas qui nous intéresse, la discussion électronique asynchrone (liste, groupes, forums) ne possède pas d'unité thématique, temporelle et spatiale pouvant définir les limites d'une conversation ; les discussions sont multifocalisées, le temps hétérogène, et le lieu ne fait sens que dans la métaphore des interfaces informatiques (*ibid.*). Nous nuancerons d'ailleurs ce dernier point : si la multitude des couches logicielles et matérielles de l'infrastructure du courriel fragmente effectivement l'espace de la communication électronique, leurs points de connexion l'ancrent dans un réseau tangible en ceci que l'utilisateur peut modifier ses paramètres, et ainsi modifier sa structure, créant un véritable lieu d'action communicationnelle. C'est peut-être la limite (dans notre cas) des propositions des chercheurs cités que de ne mentionner l'importance du format technique que comme pétition de principe pour l'analyse, mais inclus seulement indirectement et en réduisant à des figures immatérielles de la communication les contraintes des formats techniques, comme montré dans la définition du genre de la « discussion asynchrone par écrit en groupe restreint », qui regroupe nos objets d'analyse : « *un faisceau de critères, à la fois des marques formelles et du dispositif énonciatif, comme par exemple les adresses et les signatures, les marques de l'échange, les jeux typographiques et la ponctuation expressive, les quasi-didascalies, le polyadressage, etc.* » (*ibid.*).

B. Une continuité avec l'échange épistolaire ?

Le courriel est également différencié de l'autre modèle auquel on tend à l'associer, la correspondance dans l'échange épistolaire, de part sa brièveté, l'absence de marques formelles d'ouverture et de clôture, mais aussi en termes de « *mise en scène des composantes de l'échange : le destinataire et le cadre spatio-temporel* » (*ibid.*), qui sont fournis dans le courriel par les métadonnées des en-têtes et non plus laissé aux effets de réel habituellement présents dans une lettre. Sur ces derniers points, nous ne sommes pas d'accord avec Labbe et Marcoccia : d'une part sa brièveté et sa dimension informelle sont très relatives, dépendantes du contexte d'énonciation (personnel ou professionnel, avec des proches ou des inconnus) ; d'autre part la génération automatisée d'informations techniques (adresse, sujet, date...) n'empêche pas leur mise en scène dans le corps des messages, mais en plus génèrent leurs propres effets de réel, comme nous le verrons plus bas. Nous avons donc décidé de continuer à considérer l'échange par courriel à l'aune des échanges épistolaires imprimés, et nous nous inspirons, tout en l'adaptant au courriel, du cadre d'analyse proposé par (Siess, 2007), qui présente cinq constituants :

- **Le cadre participatif** inclut les formes d'adressage et d'identification des interlocuteurs (émetteur et destinataire) ainsi que la situation d'interlocution – que l'on retrouve dans le corps du message (ouverture et clôture), mais aussi dans ses péritextes (métadonnées d'en-têtes avec les coordonnées administratives et organisationnelles, ou encore données de signature).
- **Le cadre normatif** implique les prescriptions normatives qui déterminent l'étiquette de l'échange mais aussi son format – que l'on retrouve dans le corps du message, mais aussi dans les métadonnées de format.
- **Le rapport de places** implique des rapports de pouvoir induits par le statut des interlocuteurs – que l'on peut retrouver dans l'adresse courriel (nom ou surnom, indication de l'organisation ou du service hébergeur de la messagerie).

- **Les buts de l'échange** sont plus ou moins explicités à un degré local (les raisons contextuelles de l'échange) ou plus global (le contexte élargi au dispositif de la communication, par exemple la prise en compte d'une communication collective qui a une histoire).
- **L'image** déployée par les interlocuteurs sous la forme d'un ethos communiquant, à savoir la présentation de soi, la compréhension de l'autre en perspective avec des normes de la communication[37].

L'adaptation de ce cadre d'analyse aux courriels requiert ainsi la prise en compte de couches informationnelles auxquelles on fait peu souvent attention, et qui relève des métadonnées qui accompagnent et déterminent l'échange sur le plan logiciel. Ces métadonnées sont difficiles à appréhender, parce qu'elles sont pour une part générées automatiquement dans l'usage de la messagerie (et restent dans les archives de courriel sous la forme de traces) et pour une part choisies par l'utilisateur de la messagerie sous la forme du paramétrage du logiciel (même si ce choix n'est pas toujours intentionnel, selon le degré de maîtrise – et de curiosité – de l'intéressé, qui peut en rester aux paramètres par défaut).

C. L'architexte : le logiciel sous le texte

Les concepts produits par la sémiotique textuelle des médias informatiques s'inspirent de la sémiologie de la signification en la recadrant dans les dispositifs techniques de la communication. Le dispositif des médias informatisés est lui-même un texte, fait d'une multitude de couches de scripts appartenant à différents niveaux de codes, en relation d'inter-dépendance, qui supporte et conditionne l'apparition du texte électronique à la surface de l'écran. L'école de sémiotique française en Sciences de l'information et de la communication a ainsi proposé le concept d' « architexte » , concept pivot dans l'analyse des usages du « lire, écrire, récrire » sur les médias informatisés, supposant l'existence d'un logiciel situé en amont de l'écriture visible (les langage naturel et graphique apparaissant sur l'interface) automatisant les conditions de production, de manipulation, d'appropriation, et enfin de circulation des textes (Souchier, Jeanneret, et Le Marec, 2003). L'architexte est un artefact et un processus de la médiation information. Concept analytique, il tente de traduire des opérations formelles et fonctionnelles participant à l'écriture sur les interfaces informatiques :

- informer les contenus : les rendre lisibles et accessibles aux usagers les moins experts des médias informatisés grâce à des paramètres de format et d'édition ;
- gérer les flux d'information : filtrage, triage et délégation à des programmes informatiques constituent ainsi un des développements nécessaires des systèmes de traitement de l'information en réseau, en parallèle des infrastructures de protocoles qui acheminent l'information à travers la connexion des réseaux ;
- introduire des modalités de régulation et d'authentification face à une information supposément « en libre circulation ».

Ces notions permettent de penser ensemble et de manière articulée les formats d'inscriptions traditionnellement séparés dans les domaines du technique et du machinique d'un côté, et ceux du textuel et du discursif de l'autre. Nous consacrerons les prochains développements à ces couches techniques infra-linguistiques, qui n'en constituent pas moins une étude des langages – des langages techniques et formels des codes et données informatiques.

37 L'étude décrite en 2.2.2. (C. Un exemple de questionnement qualitatif) s'attache en particulier à cette dimension éthique de la communication dans des dispositifs d'échanges électroniques.

2.3.2. Les langages infrastructurels de la communication par courriel : protocoles et format

Lors de la phase de découverte d'un échantillon de listes de discussion par les étudiants en charge de travailler sur la formalisation du corpus pour leur traitement analytique (projet décrit plus bas), nous avons fait ressortir un grand nombre d'interrogations liées aux formats des messages. En effet, nous avons pu constater, en marge des régularités de structure entre des courriels pris à des points différents de l'échantillon[38], des différences à la fois structurelles et organisationnelles. Ces considérations formelles, obligatoires pour préparer le terrain à la formalisation par un traitement logiciel, sont également très intéressantes dans la mesure où elles ont permis de faire la lumière sur des éléments importants de l'histoire technique de la messagerie électronique, en particulier sur la question de ses normes et standards. Le constat qui se dégage de cette étude formelle est qu'il n'y a pas de norme générale régissant l'écriture des courriels, mais l'on trouve des standards à au moins deux niveaux (pour ce qui nous intéresse dans le cadre de cette étude) qui se superposent dans l'objet mail :

- le protocole de communication, standard de transmission et d'échange de courriel de serveur à serveur, sur le modèle du réseau informatique distribué ;
- le format, qui régit le rapport du document de type courriel à ses codes de caractères (la traduction des bits numériques en texte) et à la structuration documentaire générale du courriel.

Pour l'échantillon en question, la majorité des messages est transmise en Simple Mail Transfer Protocol (STMP), qui est le protocole historique pour les courriels électroniques, et ce depuis le tout début des années 1970. Il a défini les spécificités formelles et techniques des mails selon des spécifications qui ont été adoptées par tous les protocoles de réseaux informatiques de données : l'en-tête des courriels (« from : », « to : », etc...), les possibilités d'envoi groupé (très important dans le cas des envois collectifs), la gestion des dates et heures ou encore le format des adresses des utilisateurs. Il est encore aujourd'hui l'un des principaux protocoles de communication mail.

En ce qui concerne le format, c'est le MIME qui s'impose, un standard mis au point par les laboratoires Bell (AT&T) en 1991 à un moment important de l'histoire d'Internet, ce moment d'accélération des applications et usages grand public du début des années 1990 que nous avons déjà décrits, qui est marqué également par un effort fait pour adapter les applications de lecture et d'écriture de textes électroniques au plurilinguisme. MIME, ainsi, étend les limitations formelles du nombre de caractères d'un message, permet de transférer des pièces jointes, et accepte des encodages de caractères prenant en charge, par exemple pour le français, les signes diacritiques et les accents. MIME, outre cette possibilité d'extension technique et linguistique, est multi-plateformes, c'est-à-dire qu'il est accepté par des systèmes d'exploitation différents (Mac, Windows, Linux), au contraire d'autres qui sont exclusifs (par exemple, BinHex, sur systèmes Macintosh).

La messagerie électronique, au début des années 1970, est conçue pour transférer du texte en « ASCII US simple », fondé sur l'alphabet anglais ; utilisé quasi exclusivement pendant les deux décennies qui suivent, il est devenu une limite face au développement des communications électroniques en termes d'objets transférés et de choix linguistiques. L'ASCII (American Standard Code for Information Interchange) est l'un des premiers codes d'affichage de texte sur les terminaux informatiques. Depuis son invention chez IBM au début des années 1960 (il est accepté par la communauté comme standard en 1961), il est conçu pour faire en sorte que l'on puisse « *faire se parler les machines* »[39], c'est-à-dire que l'on puisse rendre lisible du texte (en langage naturel ou formel, de

38 La liste en question est celle dédiée aux discussions sur les Domain Name Systems (DNS) ; l'extrait qui nous a intéressé dans le cadre du travail historique était délimité par son année de naissance (1993) jusqu'au milieu des années 1990 ; mais c'est une liste toujours active : les étudiants ont donc pu étudier ses évolutions formelles sur vingt ans.

39 « *ASCII was a communication standard as well as a character display standard because before that, no two brand of*

programmation) entre plusieurs types de machines différents aux niveaux matériel comme logiciel. Nous avons par ailleurs étudié le rôle de l'ASCII dans la structuration des communications en réseau, aussi bien sur le plan technique que de l'imaginaire documentaire d'Internet, dans un article écrit pendant notre post-doctorat intitulé « Ce que les "modes d'emploi" disaient d'Internet avant le Web », qui sera publié fin 2013 dans la revue *Documentaliste (Sciences de l'information)* (Paloque-Berges, 2013a). Associé au « plain text » (ou « texte brut »), c'est un format textuel minimal et léger que l'on retrouve dans la grande majorité des fichiers échangés sur Internet jusqu'au début des années 1990 et qui correspond bien aux contraintes techniques, matérielles et logicielles de la communication en réseau, en particulier la faiblesse des bandes passantes et les limitations en termes de compatibilité inter-systèmes. Le principe du *plain text* est fonctionnel et techno-centré : le texte doit pouvoir être lu par n'importe quel éditeur de texte, gardant une indépendance par rapport aux différents standards d'encodage et de formatage qu'ont les divers logiciels éditeurs de texte ; et surtout il doit perdre le moins d'information possible s'il est traduit dans un encodage différent. Il est donc facilement lisible et transmissible sur le réseau. Selon l'article *Wikipedia*, l'usage du *plain text* « *permet aux fichiers de bien mieux survivre "dans la nature"* [in the wild]*, notamment en les immunisant contre les incompatibilités des différentes architectures informatiques* »[40]. Ainsi, le format *plain text* est une sorte de norme par défaut qui appareille l'information en ligne et crée pour l'utilisateur un environnement favorable à la consultation des documents.

Cette étude formelle, concernant la partie « listes de discussion » de notre corpus, a ainsi été nécessaire non seulement pour préparer l'exploitation pratique et analytique des documents numériques, mais aussi pour comprendre l'évolution historique technique des courriels. Concernant la deuxième partie de notre corpus, les « news » des groupes de discussion Usenet, nous avons fait un travail similaire sur les protocoles de communication qui leur sont relatifs, en particulier l'UUCP, un protocole créé en 1976 aux laboratoires Bell, un des haut lieux de la communication informatique, pour des fonctions sommaires de transfert de fichier entre ordinateurs. Il sera détourné rapidement pour échanger des messages électroniques entre utilisateurs, et sera le support du développement d'un réseau social de communication en réseau crucial dans l'histoire d'Internet (Usenet). Si nous n'avons pas eu l'opportunité de nous pencher plus spécifiquement sur le format qui lui est associé (l'UUencode), dans la mesure où cette partie du corpus n'a pas servi aux tests de structuration, nous avons cependant vu le rôle très important du protocole de communication UUCP dans l'histoire du développement des réseaux informatiques en France, dans un autre article rédigé à l'occasion du post-doctorat[41].

Ceci ne constitue qu'une toute petite partie de ce qui relève des « Couches et protocoles » de la communication en réseau. Nous engageons le lecteur voulant en savoir plus à se munir de l'annuaire des réseaux publié par John Quarterman en 1990, et en particulier du chapitre « Layers and Protocols » (Quarterman, 1990 : 45-101), qui fournit un état des lieux complets de l'état des standards de communication informatique de l'époque.

2.3.3. Le potentiel de différents niveaux de la communication informatique pour l'analyse

Nous pouvons à présent regarder de plus près la complexité de l'objet courriel pour mieux en

computers could talk to each other », in « ASCII World – History » [http://www.ascii-world.com/history].

40 « *The use of plain-text rather than bit-streams to express markup, enables files to survive much better "in the wild", in part by making them largely immune to computer architecture incompatibilities*. » Article « Plain text » http://en.wikipedia.org/wiki/Plain_text.

41 « Between electronic frontier and electronic agora: the role of Unix computer networks in France and Europe in the promotion of Internet's technologies and values as a technical democracy », écrit pour une communication présentée à la conférence "Democracy and Technology. Europe in Tension from the 19th to the 21th Century" (2013, Paris Sorbonne).

déduire la richesse d'analyse qu'il engage pour un historien ou un anthropologue des techniques, mais aussi pour l'étude documentaire en sciences de l'information et de la communication et pour le travail pratique de l'archiviste à l'ère numérique. On pourra comparer cette observation à une forme de paléographie des objets numériques[42], dans la mesure où l'on tente de mieux comprendre les systèmes graphiques des écritures numériques, leurs codes, et leurs systèmes de production afin de mieux les catégoriser et les interpréter. En ceci, c'est une attention accrue à la matérialité des écritures de la communication numérique que nous défendons ici, en éclairant les processus automatiques de structuration des documents numériques dont l'archiviste doit être conscient.

Nous nous inspirons de la dualité documentaire dégagée par Marie-Anne Chabin dans son étude des archives numériques. Le « *document-trace* », est « *la conjonction entre une activité humaine et une technique d'écriture, elle-même subdivisée entre un code, ou plusieurs codes, pour transcrire ce qu'on veut exprimer, un ou plusieurs outils et une ou plusieurs matières pour matérialiser l'écrit.* […] *Le document est en quelque sorte une « sécrétion humaine » formatée par le niveau de développement technique ou technologique de chaque époque* ». Le « *document-source* », lui, donne à voir une « *information qui a été portée hier sur un support et qui est disponible aujourd'hui comme source d'enseignement pour tout utilisateur autorisé* ». La connaissance qu'il offre est « *tirée de l'information (et) a-temporelle dans le sens où elle appartient toujours au temps de celui qui la manipule* » (Chabin, 2004).

Prenons deux exemples de notre corpus. Les noms sont anonymisés de la façon suivante […], à l'exception de celui de Christophe Wolfhugel, dont la participation au développement des réseaux Internet et Usenet en France est historiquement reconnue et publique. Parce que le repérage de son nom dans les deux exemples est important pour notre propos, mais aussi parce que nous avons jugé que ses propos n'était ici pas sensibles, nous avons choisi de ne pas l'anonymiser.

1/ Le premier message est diffusé dans les premières semaines de la création de la liste de discussion « DNS » en 1993.

```
From wolf@grasp.insa-lyon.fr  Thu Jun 24 08:34:57 1993
Return-Path: <wolf@grasp.insa-lyon.fr>
Received: from grasp1.univ-lyon1.fr by mailimailo.univ-rennes1.fr (5.65c8/150391); Thu,
24 Jun 1993 08:44:02 +0200
Received: from localhost (wolf@localhost) by grasp1.univ-lyon1.fr (8.1B/8.1) id
HAA15086; Thu, 24 Jun 1993 07:34:58 +0200
From: Christophe Wolfhugel <Christophe.Wolfhugel@grasp.insa-lyon.fr>
Message-Id: <199306240534.HAA15086@grasp1.univ-lyon1.fr>
Subject: Re: sendmail IDA
To: cru-dns-mail@univ-rennes1.fr
Date: Thu, 24 Jun 1993 07:34:57 +0100 (MEST)
In-Reply-To: <199306231447.AA02385@mailimailo.univ-rennes1.fr> from "[…] - CRI Rennes 1"
at Jun 23, 93 04:47:33 pm
X-Mailer: ELM [version 2.4 PL20]
Mime-Version: 1.0
Content-Type: text/plain; charset=US-ASCII
Content-Transfer-Encoding: 7bit
Content-Length: 2373
X-Charset: LATIN1
X-Char-Esc: 29

[…] - CRI Rennes 1 said:
|Pour .uucp .span .dnet etc c'est moins clair, il existe
|multes solutions qui marchent mais pas d'accord formel. On parlera de ces
|cas de figure le 29.

Certains jours je me demande si pour .uucp, un bounce systematique
a la source n'est pas preferable. Vu l'etat de nombreuses cartes
```

42 Il semblerait que la paléographie traditionnelle commence d'ailleurs à s'intéresser aux écritures et symboles numériques, comme en témoignent une conférence sur l'histoire de l'arobase (le signe @, démarcateur des adresses courriel) par le paléographe Marc Smith à l'Ecole des Chartes (Smith, 2013) ou, plus largement, les travaux de l'anthropologue Clarisse Herrenschmidt (Herrenschmidt, 2007).

```
UUCP et des liaisons fantaisistes qui y sont decrites ca permet
a l'emetteur d'etre informe plus rapidement sur ce qui arrivera tres
certainement...

Les autres jours (ie ou je me dis qu'au-moins un message sur deux a des
chances d'arriver a destination, et a la bonne destination)
j'utilise une heuristique qui donne des resultats acceptables
(quand on parle d'UUCP rien ne me satisfait de toutes facons, ne sont
bien lotis que les sites UUCP qui utilisent leur adresse domaine).

La supposition, discutable, est qu'il est plus fiable de joindre
n'importe quelle adresse domaine que une adresse UUCP, et je traduits
cela en des liaisons avec mon relai de cout identique.

La table des chemins ainsi generee est disponible en ftp anonyme
grasp.insa-lyon.fr:pub/uumap/work/paths et les outils permettant
sa generation sont un pre et un post-processeur a "pathalias".

grasp.insa-lyon.fr:pub/uumap/work/links.pl
grasp.insa-lyon.fr:pub/uumap/work/map.pl

Ceux qui y jetterons un oeil verront que c'est d'une simplicite.

db.pl me permet de construite la table DB (db1.5 de Berkeley)
utilisee par sendmail (attention ce n'est pas de l'IDA, une mauvaise
surprise vous attend si vous utilisez cela en l'etat avec une pathtable IDA).

Enfin, pour generer tout ca de facon automatique j'utilise unpackmaps
de Chris Lewis et un cron nocture pour faire le boulot.

Seul inconvenient: si un mail pour .uucp arrive pendant la
reconstruction de la base db, le resultat est inconnu
(comprenez je n'ai jamais pris la peine d'essayer).

Voici le cron:

(note c'est tres specifique a la config que j'avais faite
de unpackmaps avant de faire ces outils, il faudrait nettoyer ca avec
une bonne recompile d'unpackmaps... Mais ca donne une idee)


#! /bin/sh
##

cd /news/out.going

if [ -s  MAPS ] ; then
   mv MAPS MAPS.old
   sleep 5     ## IO completion
   mv MAPS.old /ftp/pub/uumap/togo
   cd /ftp/pub/uumap/work
   unpackmaps -v -M"links.pl | pathalias -i | map.pl"
   db.pl
fi
exit 0

--
Christophe Wolfhugel    |    Email: Christophe.Wolfhugel@grasp.insa-lyon.fr
```

2/ Le deuxième est diffusé dans les premières semaines de la création du groupe de discussion fr.network.divers sur Usenet en 1993 (l'anonymisation a été effectuée en partie par le service d'archive initial, ici Google Groups, sous la forme de points masquant les noms associés aux adresses).

```
Xref: gmd.de soc.culture.french:11695 soc.culture.canada:11499
Path: gmd.de!Germany.EU.net!mcsun!uknet!warwick!univ-lyon1.fr!bounce-back
From: dub...@ere.umontreal.ca ([…])
Newsgroups: soc.culture.french,soc.culture.canada,fr.network.divers,fr.announce.newusers
Subject: Liste de FTP francophones
Followup-To: soc.culture.french,soc.culture.canada,fr.network.divers
Date: 16 Apr 1993 08:37:15 +0200
Organization: INSA Informatique (Grasp), Lyon, France
Lines: 34
Sender: wo...@grasp1.univ-lyon1.fr
Approved: Christophe...@grasp.insa-lyon.fr
Message-ID: <1qlk6r$n90@grasp1.univ-lyon1.fr>
NNTP-Posting-Host: grasp1.univ-lyon1.fr

        Bonjour,
```

```
        mon idee a fait du chemin. Je vais tenter d'organiser et de mettre en forme
cette liste de services FTP francophones. Mais il me faut l'aide de vous tous.

        Il serait interessant que vous me fassiez parvenir par courrier le ou les
sites FTP que vous connaissez. MAIS ATTENTION, pour une plus grande facilite
d'organisation, j'aimerais obtenir vos informations a l'adresse suivante et au
format suivant :

        *       dub...@ere.umontreal.ca

        *       Titre du courrier : FTP

        *       Contenu de vos lettres :

        - adresse sous forme de nom et de no:

        exemple: cnam.cnam.fr                              163.173.128.6

        - ce que contient le service (progr., ressources, etc...)

        - autres renseignements juges supplementaires


        MERCI DE VOTRE AIDE.

        Je vous tiendrai regulierement au courant.

--
           [...]                  |" Les pionniers d'un monde sans guerre
         Universite de Montreal   |  sont les jeunes gens qui refusent le
   Etudiant (physique et astro)   |  service militaire"
    dub...@ere.umontreal.ca       |                        Albert EINSTEIN.
```

On voit que les deux courriels se décomposent en trois parties :

- les métadonnées ;
- le corps du message ;
- une signature.

A. Les métadonnées

Ce sont des métadonnées propres aux en-têtes des courriels, et qui relèvent d'un « protocole de présentation » standardisé spécifiquement pour la communication de messages électroniques sur Internet (Quarterman, 1990 : 75). Le premier document qui définit ce standard est le RFC 822[43] dans le cadre des protocoles de communication TCP-IP (mis à jour et complété par d'autres au fil du temps[44]).

On retrouve dans l'exemple ci-dessus les métadonnées définies par les RFC :

Des données de routage

Message-ID : l'identifiant universel et unique du message, défini par la machine d'émission.
Date : jour, mois et année, heure et précision du créneau horaire, défini par la machine d'émission.
Path : définit la source de la route qu'a pris le message à travers plusieurs serveurs pour accéder à sa destination. Dans l'exemple 2/, cette route est celle des connexions NNTP (anciennement UUCP), avec son système d'adressage particulier reconnaissable grâce aux points d'exclamation.
Return-Path : l'adresse associée à la machine finale de réception.
Received : un accusé de réception pour chaque machine sur le *Path* ; il s'agit pour l'exemple 1/ des serveurs d'origine (grasp1.univ-lyon1.fr) et du serveur de relais sur lequel est installé le logiciel de gestion de liste (mailimailo.univ-rennes1).

43 Les RFC (ou Request For Comments) sont des documents qui spécifient les standards de l'Internet ou les documentent pour information ; ils sont créés et maintenus par l'organisation indépendante Internet Engineering Task Force depuis le édbut des années 1970, dès les débuts d'Arpanet [http://www.ietf.org/rfc.html].

44 La RFC 2076, datant de 1977, fait une synthèse des spécifications pour la communication par courriel et news.

Des données relatives à l'adresse source

From : l'auteur original du message.
Sender : l'émetteur du message ; dans le cas des courriels collectifs gérés par un logiciel de liste de news, il est différent de l'auteur original dans la mesure où celui-ci l'a envoyé au logiciel qui la ensuite redistribué à l'ensemble des récepteurs de la liste ou des news. Dans ce cas, il s'agit de l'adresse de Christophe Wolfhugel, qui a créé et gère les groupes de la branche fr.* de Usenet en 1993 ; la donnée *Approved* montre que le gestionnaire de news a aussi une fonction de modérateur, pouvant filtrer des messages contraires à la charte de Usenet (comme la publicité par exemple) ; la donnée *NNTP-Posting-Host* confirme le protocole Usenet qui régit la communication des news.

Des données relatives à l'adresse cible

To : le destinataire. Dans l'exemple 1/, il s'agit de l'adresse du logiciel de liste qui redistribuera automatiquement le message aux abonnés.
Dans l'exemple 2/, le *To* est remplacé par *Newsgroups*, puisque les sources multiples sont les utilisateurs abonnés à des groupes. Le courriel est ici envoyé à fr.network.divers, où nous l'avons récupéré, mais aussi à une série d'autres groupes de Usenet. Le *Followup-to* définit les cibles secondaires auxquelles le message sera automatiquement transféré (que l'on trouve aussi sous la forme *CC* dans les courriels classiques hors des news).

Des données contextuelles

Subject : dans lequel l'auteur original définit le thème de la conversation. Dans la plupart des messageries électroniques, la réponse à un message précédent génère automatiquement un *Subject* structuré en RE : (pour réponse) + le sujet d'origine (exemple 1/).
In-Reply-To : identifie le message précédent auquel le message actuel répond.

Des données de champ d'extension (*Extension field*)

Xref : spécifique aux news, il fait en sorte qu'un utilisateur qui serait abonné à plusieurs groupes de Usenet ne voit un message, envoyé dans plusieurs de ces groupes, qu'une seule fois dans sa boîte aux lettres.
Organization (ou *X-Organization*) : donne le nom de l'organisation qui héberge la machine du *Sender*.
X-Charset et *X-Char-Esc* : pour récupérer des informations relatives aux paramètres linguistiques dans le serveur cible, si elles sont disponibles.
X-Mailer : affiche le type de logiciel utilisé pour envoyer le message.

Des données de format définies par le logiciel d'émission de l'utilisateur émetteur du courriel

Content-Type : type et sous-type de données contenues dans le corps du courriel. Dans l'exemple 1/ sont définis le type de contenu (*plain text*, et non pas html comme cela va devenir le cas avec les services de messageries électroniques Web), et le sous-type définissant le jeu de caractères (ici l'ASCII)
Content-Transfer-Encoding : type d'encodage utilisé dans le corps du courriel.
Content-Length : nombre de caractères du courriel.
Lines : nombre de ligne du courriel.
Mime-Version : identifie le format courriel.

Ces métadonnées ne sont qu'un extrait de toutes celles qui informent les messages au niveau de leur présentation et transport, et abondent en variété au sein de notre corpus. En dehors des champs

classiques servant à identifier l'émetteur, le récepteur, le sujet du message et la date, dont l'importance pour l'analyse documentaire tombe sous le sens, on voit s'ouvrir un champ d'information vaste et potentiellement très utile pour la préservation documentaire d'une part, mais aussi pour d'autres types d'analyse dont nous donnerons quelques exemples.

Dans un but d'archivage pérenne et formalisé, l'identifiant universel du message est un élément primordial pour repérer le message comme objet unique, et le référer comme tel dans une base de données. Il constitue également une base pour garantir l'authenticité de la source dans un contexte juridique.

Dans le cadre d'une analyse des infrastructures techniques de la communication, les protocoles et chemins de routage permettent de comprendre la circulation des messages entre machines et entre réseaux. Cela peut enrichir une analyse sociologique de réseau, dans la mesure où certaines informations techniques permettent d'identifier les organisations (ici des universités) au sein desquelles on envoie et reçoit de l'information et servant de relais au sein d'un réseau informatique et humain. Cela peut être crucial pour croiser l'histoire des techniques avec une histoire institutionnelle, voire économique des communications électroniques. C'est la méthode que l'on a utilisé dans notre étude des réseaux unixiens (Paloque-Berges, 2013d).

Dans le cadre d'une analyse paléographique, le format des messages et le paramétrage des logiciels d'édition, de réception, d'envoi et de transfert des courriels sont utiles pour étudier la manière dont est produit et transmis un courriel, à la croisée des choix techniques de l'utilisateur (choix d'un logiciel d'édition de texte ou de messagerie électronique) et des contraintes automatiques associées à la configuration du logiciel.

B. Le corps du message

C'est dans tous les cas d'études sur la communication électronique que nous avons rencontré le seul focus de l'analyse, en termes d'analyse de contenu et de discours, dans la mesure où cela constitue l'espace d'expression le plus évident. Parmi les angles les plus pertinents proposés par la méthode épistolaire évoquée ci-dessus (Siess, 2007), les questions d'adresses, d'interlocution, de prescription, de comportement, de rapport de places, de but et d'image peuvent trouver matière à analyse dans le corps du message, et je ne m'étendrai pas dessus. Cependant, d'autres aspects documentaires et formels sont rarement pris en compte alors qu'ils participent de plein droit au conditionnement de l'interaction et à l'accompagnement de la parole interlocutoire dans l'organisation de la communication et éventuellement sa tension vers l'action.

La réponse et la citation

Les réponses intégrées dans le corps des messages présentent un intérêt pour une analyse de discours fondée sur l'étude des interactions discursives. Il signale aussi un niveau infra-discursif qui s'appuie sur des considérations techniques et relève de l'organisation du discours dans l'espace textuel et logiciel de l'écriture de courriel. En effet, l'inclusion d'une citation du message auquel l'interlocuteur souhaite répondre se fait à la fois :

- de manière automatique : le logiciel génère une commande qui marque la réponse rappelant précisément l'identité numérique de l'interlocuteur à qui on répond et étant définie par la manière dont il a choisi de s'identifier : un nom et éventuellement le nom de l'institution ou de l'organisation à laquelle il appartient ; par exemple « [nom de l'interlocuteur] - *CRI Rennes 1 said:* »
- suite à un choix du répondant, qui peut sélectionner l'extrait auquel il souhaite donner suite, mais qui peut aussi paramétrer l'intitulé du marqueur de réponse afin de le personnaliser.

Pour ce dernier cas, les marqueurs de réponse varient énormément selon que l'utilisateur laisse les

paramètres par défaut pour le format de réponse ou qu'il le personnalise. Voici quelques exemples typiques, dont le dernier représente une personnalisation humoristique rencontrée sous des formes variées ailleurs :

```
>>>>> On Sun, 17 Apr 94 14:33:53 -0100, inetf...@genos.frmug.fr.net
>>>>> (Accs Internet France) said:


[…] ecrit (writes):


D'apres  […], Jeudi 9:54


In article <1993Mar26.1...@jussieu.fr>, truo...@amertume.ufr-info-p7.ibp.fr ( […])
writes:


[…] (r...@dufy.aquarel.fr) once wrote:
```

L'inclusion des citations, parfois plusieurs et de sources différentes, dans un courriel envoyé dans une liste ou un groupe, vise le collectif par une mise en scène de la parole dans l'espace technico-discursif qui ne va pas de soi, en particulier lorsqu'elle sert une rhétorique du conflit entre participants dans une dispute en ligne. Elle peut également supporter une analyse visant à dégager des valeurs éthiques de la communication en ligne, en ce que certaines pratiques de références au discours de l'autre sont jugées inappropriées, manipulatrices, voire interdites par les chartes des listes ou groupes de discussion – par exemple citer dans une réponse collective un message qui avait été envoyé en « privé »[45].

Autres usages des caractères textuels dans les messages

Le courriel, avant les possibilités d'inclusion d'images ou autres objets médias (son, vidéo) grâce à l'HTML à partir de la moitié des années 1990, est intégralement composé de caractères textuels. Cependant, l'usage de ces caractères n'est pas limité à l'écriture en langage naturel. En effet, les symboles alphabétiques ou typographiques peuvent supporter deux principales fonctions :

- l'écriture de script en langage de programmation ;
- la structuration visuelle de l'information.

Dans l'exemple 1/, on voit en effet un extrait de code informatique en langage Perl (noté .pl). Ce script est fait pour être exécuté dans un environnement ou le langage Perl est installé, mais il peut être aussi « lu » par l'utilisateur dans un éditeur texte classique, une fonction à la base de la messagerie électronique. Ces scripts, s'ils peuvent être intéressants comme « textes informatiques » pour un historien des techniques s'intéressant aux langages de programmation, peuvent cependant poser un problème pour l'exploitation documentaire des sources. En effet, soumis à l'analyse instrumentée d'un logiciel, ils peuvent entrer en conflit avec le programme exécuté par le logiciel et produire des bugs informatiques. Nous avons rencontré ce cas avec le logiciel d'analyse de forum Calico (dont nous reparlerons) : le script Perl a été « interprété » par l'application, qui en retour a affiché des modifications imprévues sous la forme d'un surlignement systématique de toutes les données affichées après le message en question.

Ces erreurs de parcours mises à part (qui ne peuvent être traitées qu'au cas par cas conjointement avec le concepteur du logiciel d'analyse), d'autres usages des caractères symboliques doivent être pris en compte dans l'analyse du message. En effet, ils peuvent être utilisés comme support d'une structuration visuelle du message plus ou moins élaborée dans sa disposition graphique.

Dans l'exemple 2/, on constate ainsi que le symbole « étoile » et le caractère typographique

45 Paloque-Berges, 2011a et Paloque-Berges, Gueguen et Scopsi, 2013 (cf. aussi 2.2.2.C. de ce rapport).

« tiret » permettent de structurer le texte grâce à la mise en valeur sous forme de liste, afin de mieux organiser le texte dans l'espace du courriel. Cet usage classique ne présente que peu d'intérêt si ce n'est pour replacer l'étude paléographique des écritures numériques dans une histoire des formes graphiques de l'écriture (Goody, 1979 ; Herrenschmidt, 2007). Dans notre cas, cela s'inscrit en particulier dans une histoire de l'appropriation et du détournement des normes de l'écriture pour la création de formes alternatives qui remontent aux calligrammes, se développent avec les techniques d'écriture typographiques (machines à écrire) et de transmission de messages textuels (épistolaire, télégrammes, téléinformatique), et trouvent dans la textualité numérique un terrain d'expansion particulièrement riche (on pensera aux émoticônes de type « :-) » qui permettent d'introduire des nuances dans la communication écrite à distance.

Plus complexe dans sa forme graphique, l'usage des tirets et traits verticaux et horizontaux, de chevrons (« > < ») permettent de recréer des formes paratextuelles comme, outre des listes simples, des sommaires (dans le cas de longs documents envoyés par courriel, par exemple les chartes d'utilisation et autres « Foires aux Questions »), ou des tableaux. Voici un exemple extrait de la liste « DNS » :

```
From […]@univ-rennes1.fr  Tue Jun 22 07:57:24 1993
Return-Path: <[…]@univ-rennes1.fr>
Received: by mailimailo.univ-rennes1.fr (5.65c8/150391); Tue, 22 Jun 1993 07:57:24
+0200
Date: Tue, 22 Jun 1993 07:57:24 +0200
From: […]@univ-rennes1.fr ([…])
Message-Id: <199306220557.AA21663@mailimailo.univ-rennes1.fr>
To: cru@univ-rennes1.fr
Subject: Reunion du 29/30 juin.
Cc: cru-bu@univ-rennes1.fr, cru-dns-mail@univ-rennes1.fr,
        cru-news@univ-rennes1.fr


Revoici le programme du 29/30. (avec en fin le programme News plus detaille)

Cote Messagerie, Serge se charge de "fignoler" le contenu.

Maillaux, doit vous envoyer les ordres de mission, je m'etonne qu'ils ne
soient pas encore arrives.
Je vois ca avec lui.

Amicalement.
[…]

------------------------------------------------------------------------------
              Ordre du jour de la reunion du 29/30 juin.
              ------------------------------------------
              (possibilite de mener des sessions en // si besoin)

Journee 1
=========

              Operationel (info/organisation) |  Etude (ou a faire/decider)
              ------------------------------------------------------------
              Etats des lieux                 |
Messagerie    DNS     (--> Parrainnage)       |
              Sendmail(--> Parrainnage)       |
                                              |
                       <----- Services de liste de distribution --->
                           <----- Services de BAL deportees --->
                                              |
                                              |   MIME
              ------------------------------------------------------------
News          Etat des lieux du service       |
              - couverture actuelle et prevue |
                                              |
                           <---   RARE News Consortium.   --->
                      <--S'organiser pour faire "vivre" le service de News-->
                           <-- finalisation de la documentation -->
                     <-- contractualisation d'un minimum de service -->
                                              |
                                              |   News Multimedia
              ------------------------------------------------------------
```

```
Imagerie            <--- Experiences/projets/validations en cours --->
Multimedia                                      | - Diffusion de seminaires
                                                |   (projet restreint)
                                                | - reunions reparties
                                                |
                    ------------------------------------------------------
Annuaire            Etat du service (Rennes1, LyonI |
                       Toulouse, CNUSC)             |
                                                    | Strategie pour globaliser
                                                    |    le service.
                                                    | Produit(s) a utiliser.
                                                    | Regles a respecter.

                        <------------- REFLEXION sur comment progresser----->

 […]
```

Cet exemple nous intéresse pour deux raisons. D'abord, il est un moyen d'organiser les connaissances acquises (ou à acquérir) sous une forme synthétique, directement compréhensible grâce à la disposition en tableau. En ceci, il est à la fois une table d'organisation et d'instruction pour l'action, à savoir la préparation d'une réunion hors-ligne grâce à des moyens de communication en ligne qui ne se limitent pas à la simple transmission d'information. Ensuite, parce que la nature du texte numérique (sous sa forme « texte brut ») est d'être infiniment inscriptible et modifiable, il est aussi un outil d'échange récursif : les récepteurs sont encouragés à modifier des éléments et en soumettre de nouveaux (dans un autre courriel faisant partie de la discussion générale sur cette réunion), voire à commenter les éléments existants dans un espace dédié (« *REFLEXION sur comment progresser* »).

La forme la plus élaborée de réorganisation des caractères textuels dans l'espace du message est sans conteste ce qui est connu dans la culture Internet sous le nom « art ASCII ». Né avec la communication électronique, en particulier chez les cultures hackers pour signer leurs exploits sur le principe du graffiti, et apprécié pour ses possibilités mimétiques et créatives en terme d'iconographie, il a conjointement été utilisé par les entreprises de logiciels informatiques pour créer des logos dans des environnements informatiques textuels où l'image numérique était encore rare. Il continue aujourd'hui à peupler les forums Web. Voici un exemple d'un courriel envoyé sur Usenet par l'organisateur d'un concours d'informatique à l'école EPITA :

```
Path: gmd.de!xlink.net!zib-berlin.de!zrz.TU-Berlin.DE!netmbx.de!Germany.EU.net!EU.net!
julienas!boson.epita.fr!krypton!prologin
From: prol...@krypton.epita.fr (ass prologin)
Newsgroups:
fr.announce.divers,fr.comp.divers,fr.comp.sys.pc,fr.comp.windows.x11,fr.doc.divers,fr.mi
sc.divers,fr.network.divers,fr.news.divers,fr.rec.divers,fr.soc.divers,fr.test,general,s
oc.culture.french
Subject: CONCOURS EUROPEEN INFORMATIQUE
Followup-To: soc.culture.french
Date: 20 Dec 1993 20:54:34 GMT
Organization: Epita (French Computer Science school)
Lines: 49
Sender: prologin@krypton (ass prologin)
Distribution: world
Message-ID: <2f53ea$i2u@boson.epita.fr>
NNTP-Posting-Host: krypton.epita.fr

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

      *         *                                        *         *
     *           *                    |                *           *
    *             *          \         /          *      1994     *
   *   MASTER DES    *          |=====|           *     EUROPEAN    *
   * GRANDES  ECOLES *         -  |_____|  -        *     COLLEGE     *
   *   EUROPEENNES   *            \   /             *     COMPUTER    *
    *     1994      *            /   |||   \          *  CHAMPIONSHIP *
     *           *                /___\              *             *
      *         *              _____|  1  |_____        *         *
        * * *               |  3  |     |  2  |           * * *
                           _|_____|_____|_____|_

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                          Registrations & Information
```

```
internet : Prol...@epita.fr             videotext : 3615 PROLOGIN
fax      : (1)44.08.01.99                address   : 106-112 bd de l'Hopital
telephone: (1)44.08.01.74                            F75013 Paris FRANCE
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

                L E   C H A L L E N G E   P R O L O G I N


    PROLOGIN, association d'eleves-ingenieurs organise un concours
d'informatique s'adressant a l'ensemble des Grandes Ecoles et Universites
Europeennes. Celui-ci aura lieu fin Avril.

    Pour participer, chaque Ecole ou Universite doit constituer une equipe
de 5 etudiants.
Ces equipes devront posseder les connaissances les plus larges possibles
dans les modules suivants :

      * Modelisation mathematiques
      * Sciences Cognitives et Intelligence Artificielle
      * Hardware
      * Software
      * Reseaux et telecommunications
      * Systeme
      * Marketing

Pour toutes informations supplementaires, contacter, par ecrit, l'association
PROLOGIN.


~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

prol...@epita.fr
```

L'art ASCII permet notamment de faire se distinguer un message en particulier parmi la myriade que chaque utilisateur de listes ou de groupes de discussion reçoit chaque jour. Il travaille de manière explicite la question de l'identité numérique dans sa composante visuelle et communicationnelle.

C. La signature

Nous ne attarderons pas sur la signature, dans la mesure où ses propriétés formelles et communicationnelles recoupent celles que nous venons de décrire à propos du corps du message. Fichier séparé du fichier message, la signature est connue sous le nom « .sig » (l'extension du format associé au fichier). Elle se compose dans les paramètres du logiciel de messagerie, qui se charge ensuite de la joindre automatiquement au message envoyé.

L'exemple 1/ représente la version la plus minimale de la signature : nom, prénom, coordonnées électroniques, auxquelles sont souvent ajoutées des coordonnées postales et téléphoniques, le nom de l'organisation ou de l'institution à laquelle appartient l'utilisateur. Si les utilisateurs prennent souvent soin de mentionner, au sein même des signatures, que leurs opinions ne reflètent que la leur et non celle du cadre professionnel qui les emploie, selon la formule classique, faire figurer le nom de l'employeur est une pratique systématique (du moins avant la généralisation des ordinateurs et connections Internet privées) dans la mesure où c'est d'abord sur le lieu de travail que l'envoi de courriels se fait, ou en tout cas à travers un compte de messagerie comportant une indication d'employeur.

L'exemple /2 présente une autre pratique courante, celle de l'inclusion d'une citation de personnages célèbres – ce choix est cependant variable, pouvant aussi relever d'un trait d'humour, d'une plaisanterie pour initiés. L'art ASCII abonde également dans les signatures, faisant de leur composition un petit exploit : faire tenir le dessin ASCII le plus compliqué dans un espace le plus petit possible.

Nous avons par le passé beaucoup étudié ces réappropriations graphiques des caractères textuels en tant qu'ils déploient des qualités poétiques, rhétoriques, et techniques intéressantes à analyser pour mieux comprendre la manière dont les identités numériques se sont construites au travers de l'histoire d'Internet grâce à des tactiques de communication inventives. Pour plus de détails, le lecteur pourra se référer à (Paloque-Berges 2009, 2010a, 2011a, 2011b, 2013a).

En définitive, ces trois niveaux de l'écriture de la communication électronique, de la trace à la source, du contenu à la forme, doivent ainsi être pris en compte par l'analyse à la lecture humaine aussi bien qu'automatique. L'enchevêtrement des couches d'information et d'utilisation des caractères textuels rend complexe l'analyse instrumentée, dans la mesure où, sur le plan documentaire, il faut pouvoir traiter chaque couche pour ce qu'elle est, et, sur le plan formel, la structuration graphique des courriel doit être respectée dans son traitement automatisé (rendu sur la page prenant en compte les espaces et autres dispositions typographiques). Pour notre travail pratique de normalisation des documents pour l'analyse instrumentée, il a été très difficile de prendre en charge cet aspect de la richesse de l'objet courriel.

3. Un projet de mise en corpus et de partage des sources documentaires des communications

Le travail sur les nouvelles sources numériques natives implique une modification du rapport des chercheurs aux corpus, qui doivent donc s'intéresser à des aspects documentaires jusque là laissés à l'archiviste, en particulier les dispositifs d'archives dans leur genèse et leur formation (Brian, 2001). La prise en compte de ces aspects documentaires implique d'abord de mettre en forme ces documents pour les rendre interrogeables, afin de les analyser à l'aide d'un logiciel ou de les rendre accessibles à d'autres chercheurs qui voudraient à leur tour les analyser (soit pour vérifier les résultats d'analyse déjà effectuées, soir pour effectuer d'autres analyses). Au-delà, il s'agit de préparer ces sources pour qu'elles participent à une économie de la recherche de type Open data, ou, plus exactement dans notre cas, de type « open corpus ».

La logique de l'Open data (« données ouvertes ») est au cœur des réflexions actuelles sur le partage des données numériques, ici à visée de recherche scientifique[46]. Les « données ouvertes » peuvent être utilisées pour potentiellement n'importe quel type d'usage. Leurs critères fondamentaux de définition sont :

- la disponibilité et l'accès,
- la réutilisation et la redistribution,
- la participation dite « universelle » sans restriction ni discrimination d'usage, l'utilisateur pouvant réarranger le corpus et recombiner les corpus.

Ce partage des données et des corpus implique une formalisation, c'est-à-dire une redocumentarisation selon des normes et des standards qui les rendront compatibles entre différents systèmes d'interrogation et d'analyse (on parle plus précieusement d'interopérabilité).

Dans le cas de nos sources, il s'agit d'un travail où la question des normes rencontre celle de la légitimité. En effet, si ces sources ne sont pas totalement inédites (elles relèvent du genre de la correspondance, déjà pris en charge comme sources dans les SHS), elles présentent une complexité à la croisée de la mémoire et de la technique, comme nous l'avons vu dans la partie précédente. Ces discussions sont encore considérées comme des *trivia* de l'échange entre experts au sein d'une communauté, notamment en raison de leur statut informel ; et leur richesse en termes de données et métadonnées n'est pas encore pris en compte dans les analyses.

Notre projet s'inscrit dans le mouvement des Digital Humanities (Humanités numériques), qui fait la promotion de l'Open data, du partage et de l'intéropérabilité des corpus au sein du champ scientifique. Parmi les initiatives de standardisation de corpus relativement classiques, en général numérisés depuis des sources analogiques, les Humanités numériques cherchent aussi à renouveler les sources et les productions de la communications de la recherche. Par exemple, ce courant propose en ce moment même d'intégrer des formes a priori non légitimes (les billets de blog) dans la catégorie des publications scientifiques. Notre démarche est similaire : les communications électroniques des ingénieurs de réseau peuvent ainsi servir d'objet d'analyse, mais aussi de source d'information scientifique et technique. Il peut en effet être intéressant pour un chercheur d'aller voir un débat ayant eu lieu dans une liste maintenue par ses pairs depuis plusieurs années (notamment pour établir des états de l'art, mais aussi étudier les controverses dans le temps). Les listes de discussion offrent une vue dans l'évolution des travaux des chercheurs, leur manière de les communiquer et de s'organiser en collectifs.

46 L'autre volet de l'open data concernant les gouvernements pouvant ou devant mettre à disposition du public certaines informations. Cf. le Vademecum de l'Open Data publié le 17 septembre 2013 [http://www.modernisation.gouv.fr/laction-publique-se-transforme/en-ouvrant-les-donnees-publiques/lopen-data-son-vade-mecum].

3.0. Préambule : un intérêt actuel pour la gestion des archives email historiques

Nous évoquerons à titre d'exemple une initiative récente, dont nous n'avons malheureusement pris connaissance que trop tard pour analyser son intérêt pour notre recherche. À l'automne 2012, l'Internet Engineering Task Force, déjà mentionnée, publie une Request For Comment (des documents textes numériques, appelés par leur acronyme RFC) proposant à la communauté des ingénieurs d'Internet de travailler sur un système d'archivage et d'interrogation de la semi-archives des échanges ayant eu lieu sur leur liste de discussion. Comme le rappelle Gérard Le Lann, un chercheur en informatique ayant travaillé sur le projet Cyclades, modèle expérimental de réseau informatique de communication par paquets ayant inspiré Internet[47] :

> Le premier vecteur d'échanges de type numérique (qui a été déterminant pour la définition des protocoles Arpanet, puis Internet) connu dans le monde scientifique -- toutes disciplines confondues, me semble-t-il -- fut le RFC. Pour la première fois, depuis un clavier, tout scientifique concerné pouvait donner son avis sur des propositions de solutions et de normes, avec interactions itératives jusqu'à consensus.

Nous publions ici les premières ligne de l'appel à projet pour l'archivage et la recherche dans les listes de l'IETF, afin que lecteur puisse le mettre en perpective avec le travail que nous avons fourni lors de ce post-doctorat – en prenant bien en compte que notre projet a pour but de partager un corpus avec la communauté scientifique, but sensiblement différent de celui de l'IETF.

> RFC 6778: IETF Email List Archiving, Web-based Browsing and Search Tool Requirements
> Date de publication du RFC : Octobre 2012
> Auteur(s) du RFC : R. Sparks (Tekelec)
> Pour information
> Première rédaction de cet article le 31 Octobre 2012
>
> Ce RFC est le cahier des charges du futur outil d'accès aux archives des innombrables listes de diffusion de l'IETF. Le gros du travail de cette organisation repose sur ces listes de diffusion, dont la plupart sont publiques, archivées et accessibles via le Web. Cette masse d'information est un outil formidable pour comprendre les décisions de l'IETF et les choix techniques effectués. Mais son volume rend l'accès à l'information souvent difficile. L'IETF se vante de sa transparence (tout le travail est fait en public) mais avoir toutes les discussions accessibles ne suffit pas, si l'information est trop riche pour être analysée. D'où l'idée de développer une ensemble d'outils permettant d'accéder plus facilement à ce qu'on cherche.
> À l'heure actuelle, la recherche est particulièrement difficile si une discussion s'étend sur une longue période et surtout si elle s'est répartie sur plusieurs listes. Imaginons un auteur d'un RFC, un président de groupe de travail, ou un simple participant, qui veut retrouver toutes les discussions qui ont concerné un Internet-draft donné. Il va trouver des messages dans la liste du groupe de travail mais aussi dans celles d'une ou plusieurs directions thématiques, et peut-être aussi dans la liste générale de l'IETF. Certains ont chez eux des copies de toutes ces listes (copies parfois incomplètes) pour utiliser des

47 Mais n'ayant pas été adopté en France, l'Etat lui ayant préféré , en 1978 le réseau Transpac proposé par France Télécoms, réseaux qui connectera le Minitel (Schafer, 2012b)

> outils locaux. Au fait, personnellement, je me sers surtout de grepmail et je cherche encore un outil qui indexerait les données.[48]

3.1. Préparer la mise en corpus : définition, structuration, collecte, et cadre éthique

Les différentes étapes nécessaires à préparer les sources documentaires sont complexes ; nous définissons ici un vocabulaire pour que le lecteur puisse se repérer :

> **Communications électroniques** (ou médiées en réseau) : l'échange discursif et textuel médié par réseau numérique et supporté par un type de logiciel spécifique (ici les logiciels de listes ou de groupes/*news*).
>
> **Documents numériques natifs** : l'inscription de ces communications dans un type documentaire numérique, dans notre cas le courriel (avec ses données et métadonnées, c'est-à-dire son message et les instructions qui définissent ses modalités d'envoi, d'affichage, etc.).
>
> **Semi-archives** : nous proposons ce terme pour décrire les archives des documents issus des communications de listes ou de groupes telles qu'elles ont été créées à un certain moment sans suivre de normes scientifiques (pour les corpus) ou institutionnelles (pour le patrimoine) ; elles sont très hétérogènes sur le plan des intentions et des techniques déployées pour sauvegarder et mettre à disposition (quand elles le sont) ces collections de documents ; dans notre cas, il s'agit de semi-archives disponibles sur le Web où nous récupéré notre corpus.
>
> **Corpus** : indique le choix d'un certain nombre de semi-archives pour leur exploitation, analyse et partage selon les choix scientifiques de notre projet. Il peut exister des **corpus restreints**, destinés à une nouvelle délimitation des matériaux utiles dans un cas d'étude, comme par exemple dans (Paloque-Berges 2013c, 2013d).
>
> **Archives** : nous entendons ici les archives au sens le plus formalisé et le plus final, c'est-à-dire une collection documentaire standardisée par un biais institutionnel, ici selon les critères de l'intéropérabilité des corpus scientifiques.

Par ailleurs, si nous parlons de **données** pour montrer que le travail de mise en corpus prépare le travail d'analyse instrumentée (c'est-à-dire le traitement logiciel des données grâce à des instruments informatiques), nous maintenons le regard des Sciences de l'information et de la communication en nous intéressant d'abord aux données mises en formes dans des documents, c'est-à-dire des **informations**.

3.1.1. La « construction » du corpus pour son exploitation

Le mise en corpus de nos documents se situe dans la lignée des travaux en sciences du langage, « *discipline d'observation des pratiques langagières* » (avec toutes les nuances que le « langage » des objets courriels implique, comme décrit précédemment) et non pas de recueil de données par l'interaction comme en sociologie ou en anthropologie (Rigot, 2006[49]) ; elle présente deux traditions méthodologiques :

> La notion de corpus renvoie traditionnellement à deux conceptions. La première est documentaire et ne retient que des variables globales ignorant les aspects textuel et

48 Extrait de la traduction par Stéphane Bortzmeyer de la RFC 6778 publiée sur son site personnel [http://www.bortzmeyer.org/6778.html].

49 Rigot Huguette, « (En)-jeux de corpus pour la recherche en SHS. Enonces, textes et documents », in Duteil-Mougel C. et Foulquié B., 2006, pp.171-178.

> structurel. Dans ce cas, le corpus est un réservoir d'exemples langagiers ou ... une base de données textuelles. La deuxième conception, plus liée à une tradition herméneutique, prend en compte les relations intertextuelles. La définition proposée par F. Rastier fait du « corpus [...] un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés [...] de manière théorique réflexive en tenant d'une gamme d'applications » (*ibid*. : 172).

La définition de Rastier, que nous adopterons pour préparer notre corpus, insiste sur les objectifs d'usage, et donc sur le fait que le corpus est une construction en vue d'une ou plusieurs tâches, dans notre cas :

- l'analyse instrumentée du corpus par des logiciels pour compléter l'analyse qualitative par des analyses quantitatives ;
- l'intéropérabilité du corpus pour qu'il soit exploité, interrogé, analysé par d'autres chercheurs.

On voit aussi se dessiner le lien logique entre document, corpus et archives, d'une part, exploitation et usage d'autre part :

> S'il est un construit, le corpus est situé dans des pratiques qui « travaillent », documentent, catégorisent des rassemblements textuels, ainsi, quatre niveaux sont à distinguer : l'archive qui regroupe l'ensemble des documents accessibles, le corpus de référence à partir duquel des corpus d'études vont être délimités et enfin le sous corpus de travail variant selon les étapes de l'analyse (*ibid*. : 173).

Le travail de préparation documentaire des corpus implique (et nous reviendrons sur ces points de manière plus détaillée dans la suite) :

- une documentation des contextes généraux et situations particulières des objets étudiés ; ce que nous avons fait en étudiant l'histoire technique et sociale de la communication médiée en réseau sur Internet et les moments importants de transition et d'avancées en la matière sur le territoire français et au cœur des communautés de pratiques des scientifiques et ingénieurs de réseaux informatiques ;
- le respect des dispositifs juridiques et éthiques ; dans le cas d'objets langagiers produits dans le passé, au volume important et aux interlocuteurs multiples, il est virtuellement impossible de faire signer un accord pour citation ; par contre, l'anonymisation, bien que complexe pour un corpus de courriels, reste possible ;
- l'analyse des informations du recueil des données mises en forme dans des documents à leur copie sur un support fiable et durable à toutes les étapes de structuration des documents par le nettoyage, le codage et le balisage des différentes données et métadonnées, créant de nouvelles informations par le biais du processus de documentarisation ;
- la possibilité, grâce à la structuration du corpus, de sélectionner des sous-ensembles de données par des critères divers (date, auteur, thème...) ;
- la recherche d'une plateforme logicielle pour stocker, partager, interroger le corpus à partir d'une base de données ;
- le choix d'un logiciel d'analyse.

Dans ce travail, la question du balisage des corpus est la plus délicate. En effet, la manière dont les informations seront structurées dans un document interrogeable par un moteur de recherche ou un logiciel d'analyse est déterminée par le choix, le nombre et le type de repères (on parle dans les langages de structuration et d'affichage de « balises », traduit de l'anglais « *markup* »). Nous avons pris connaissance d'un travail de balisage de corpus épistolaire numérisé (et non numérique natif) pour préparer notre travail, en raison du constat de départ que présente le chercheur :

> Une grande partie des lettres de Proust, « écrites au galop », ont la même fonction que nos courriers électroniques : demandes, réponses,poursuite d'une conversation orale, etc., elles ont une fonction relationnelle privée – et ne sont pas destinées, à l'inverse des lettres de la marquise de Sévigné ou de Bayle, à être lues dans lessalons. En revanche, il est indéniable que ces objets textuels se caractérisent par une forme spécifique : en-tête et date (parfois), formule d'adresse toujours, formule d'adieu, signature, qui les classent comme « lettres » au premier regard, avant même la lecture du texte proprement dit. Inscription de la relation, et souvent des coordonnées du réel, ces indices formels, ainsi que la modalité généralement discursive, font de la lettre une parole adressée. Je propose donc de définir l'épistolaire simplement comme « classe de textes » spécifique (plutôt que d'un « genre littéraire ») (Leriche, 2006 : 264[50]).

Le travail engagé sur ce corpus épistolaire, engage, comme pour le nôtre, une double dimension :

- la création d'une base documentaire pour bien comprendre les références parfois non explicites de l'auteur au travers d'un ensemble épistolaire hétérogène (sous-genres à buts différents selon le destinataire) ;
- un balisage pour étudier le contenu ne prenant en compte que l'étude des thèmes ne peut se passer d'une étude « distributionnelle » (la relation aux correspondants) et que la structuration devra tenir compte des effets de structuration *a priori* de l'objet épistolaire (ouverture/clôture, adressage, réponses et citations, signature...).

A cela s'ajoute un parti pris méthodologique, puisque le chercheur cité suggère qu'une situation de « parole adressée » est destinée à agir sur le destinataire, et donc que l'expressivité épistolaire est subordonnée à une visée pragmatique ; il faut prendre en compte aussi les « opérations pragmatiques » du corpus (dans un logique d'actes de langage). Notre travail n'a pu aller jusque là, comme nous le verrons par la suite, mais il nous semblait important de citer cette perspective méthodologique et pratique comme un modèle possible pour le traitement des corpus épistolaires électroniques.

3.1.2. Le choix de l'XML : un langage de balisage orienté usage

Retenons pour l'instant les choix techniques qui accompagneront le balisage de nos corpus. Le langage XML (eXtended Markup Language, « langage de balisage extensible » - on peut lui ajouter des balises personnalisées) est un choix qui s'impose pour la structuration des documents du corpus, et la structuration du corpus en tant que collection d'archives lui-même. Tout d'abord, ce format est celui qui est le plus utilisé actuellement au sein des Humanités numériques, pour sa souplesse, sa simplicité, et surtout le fait qu'il ait été conçu pour structurer des documents textuels, qui restent l'un des supports majeurs d'analyse en SHS. Ensuite, c'est un format qui favorise deux exigences que nous avons assignées à notre travail. Un certain nombre de logiciels d'analyse requièrent un format d'entrée de type XML. Il s'agissait donc de nous familiariser avec ce format afin de soumettre notre corpus à l'analyse instrumentée. Puis, le XML est un langage d'affichage Web, c'est-à-dire qu'il permet de visualiser et d'effectuer des requêtes sur un document textuel à partir d'interfaces Web (connectées au réseau ou non). En ceci, le XML propose des normes favorables à une logique d'intéropérabilité, c'est-à-dire permet à d'autres services d'avoir accès aux métadonnées du corpus, et donc de les interroger depuis une plateforme extérieure et de produire des résultats. Il correspondait donc à notre but à plus long terme, à savoir rendre des archives de listes et groupes de discussion « trouvables », et donc « cherchables » sur le Web en tant que sources d'information.

On peut encoder (on dit aussi « convertir » ou « structurer ») un document XML à la main, en

50 Leriche Françoise, « Quel balisage pour les corpus épistolaires numériques ? De l'annotation traditionnelle du "document" à une analyse générique et pragmaticienne », in Duteil-Mougel C. et Foulquié B., 2006, pp.262-270

introduisant les balises nécessaires aux endroits voulus, ou bien à l'aide d'un script de programmation qui va automatiser le balisage à partir de repères trouvés dans le texte (par exemple : à chaque fois que le programme trouve la mention « From », il crée une balise « émetteur »). Pour les corpus longs, cette deuxième solution est évidemment celle qui doit être adoptée, mais elle comporte des limites (en particulier s'il n'y a pas de repères dans le texte, ou s'ils sont ambigus). Dans sa dimension formelle en tout cas, le XML s'adapte très bien à un corpus de courriels qui, comme on l'a vu en deuxième partie, présentent des métadonnées d'en-tête, repères très pratiques pour l'automatisation du balisage. En ceci, ils sont encore plus faciles à baliser que les corpus épistolaires imprimés (et numérisés), qui ne possèdent pas de métadonnées par défaut, mais qui ont quand même l'avantage de reposer sur une structure formelle récurrente dans le genre :

> Dès lors que le balisage XML autorise une identification multiple d'un « document » – les balises « auteur », « destinataire », « lieu », « date », étant considérées comme balises de « structure » –, on peut étendre la nomenclature de ces balises de structure à l'étiquetage d'autres constituants génériques comme les formule d'adresse et d'adieu, la signature, les Post-scriptum. Un autre trait formel participe de la poly-énonciation épistolaire : les citations. Notre approche se veut ici purement descriptive : repérer les éléments formels entrant dans le texte de la lettre, qu'ils soient spécifiques de l'épistolaire ou également présents dans d'autres formes de texte (Leriche, 2006 : 267).

Il permet en outre d'autres types de balisage, comme le balisage thématique qui sera facilité dans notre corpus par les « sujets » des courriels, généralement représentatifs synthétiquement du contenu de la discussion. Mais plus on affine le balisage, comme cela peut être le cas pour les thèmes et les opérations pragmatiques, moins on peut compter sur les scripts d'automatisation XML.

3.1.3. Collecte des sources et problématiques documentaires

La collecte des sources documentaires issues des communications électroniques est primordiale si l'on veut pouvoir convertir les documents ; mais cette collecte présente plusieurs problèmes techniques qu'il faut prévoir. Le fait que ces sources soient des documents numériques natifs, et non pas numérisés, ne facilite pas les choses. Ils possèdent déjà leurs propres formats, normes et structures et ne sont pas des documents « originaux » (impossible dans le cas du numérique). De plus, selon la manière dont ils ont été manipulés et formalisés au cours de leur existence (en particulier dans leur première mise en archive dans les lieux en ligne où nous les avons récupérés), de nombreuses couches de structuration ont pu s'accumuler.

Si la semi-archive est hébergée en ligne et présente un volume important, on pensera tout d'abord à utiliser un logiciel d'aspiration automatique des contenus déjà existant ou scripté pour l'occasion par un programmeur. La limitation principale que nous avons rencontrée à propos de nos gisements est que certains sites empêchent l'aspiration en bloquant la tentative de connexion du logiciel en question immédiatement (Google Groups) ou au bout d'un certain volume (par exemple 100 messages pour le site de Renater). Nous avons contourné cette limitation de manière artisanale, en récupérant à la main (copier-coller) les messages des groupes de discussion, solution très coûteuse en temps et qui a entraîné un rétrécissement de notre corpus. Nous avons cependant réussi à récupérer une liste entière par le biais de Stéphane Bortzmeyer (administrateur de la liste DNS toujours active, hébergée sur le site de Renater) qui nous a fourni un « dump d'archive », c'est-à-dire un ensemble de documents au format texte brut.

On devra également faire attention au format de sortie des documents récupérés. Dans le cas d'une aspiration ou d'un copier-coller, le format de sortie est généralement de l'HTML qui ajoute des données aux informations contenues dans les archives de message. Il faudra alors ensuite nettoyer les

données. Le cas du « dump d'archive » est le plus avantageux car il produit du document de type texte brut, ensuite plus facile à traiter, mais qui peut garder des traces d'éléments de formatage divers. La plus grande difficulté à été de gérer la multiplicité des formats d'encodage de caractères traversant le corpus. En effet, la communication électronique se faisant d'un logiciel de messagerie à un autre, chaque logiciel a son propre encodage selon ses paramètres externes (système d'exploitation de l'ordinateur, version du système, langue du système) et internes (version du logiciel de messagerie, choix de l'utilisateur entre différents types d'encodage). Il faut ajouter à cela les réencodages au cours du processus de mise en semi-archive, mais aussi les codes acceptés par les logiciels de traitement analytique auxquels seront soumis les documents[51].

3.1.4. Positionnement éthique du chercheur : quelques considérations et précautions

Cette vision techniciste des corpus doit par ailleurs s'accompagner du regard réflexif de celui qui s'intéresse à ces archives sur sa propre pratique : que ce soit du point de vue de l'ingénierie de la recherche, de l'analyse ou de l'utilisation des résultats, l'étude et le partage des archives dans une logique de corpus ouvert réclame plus que jamais une attention accrue aux positionnements éthiques de la science face informations qu'elle analyse et aux données qu'elle génère. (Latzko-Toth et Proulx, in Barats, 2013[52]) décrivent deux types d'encadrement de la recherche sur Internet :

- un cadre réglementaire : l'ensemble des contraintes formelles posées par un contexte institutionnel d'usage dans les universités anglo-saxonnes, à travers des comités de réglementation qui vérifient le respect de critères juridiques et éthiques du chercheur afin de lui donner l'autorisation de l'enquête, de l'analyse et de la publication – contraintes qui peuvent poser problème dans le cadre d'une problématisation qui évolue au fur et à mesure de la recherche (comme c'est le cas dans la théorisation ancrée) ;
- un cadre normatif : les principes consensuels internes à une discipline ou un courant de recherche, qui émettent, par le biais d'association ou sociétés savantes, des codes de bonne conduite et de bonnes pratiques (compétence professionnelle, intégrité, responsabilité scientifique et sociale, respect pour les droits, la dignité et la diversité des personne) ; le chercheur, s'il veut rendre sa recherche légitime, est invité à les suivre ; pour les recherches sur Internet, c'est la charte de l'Association of Internet Researchers (AoIR) qui fait autorité depuis une dizaine d'années.

A. Le statut d'archive comme preuve et témoignage

Les documents que nous avons étudié ne sont pas des archives au sens restreint du terme, comme nous l'avons vu, dans la mesure où elles n'ont pas a priori « *vocation à servir de preuve l'action qu'elle supporte* », une preuve recevable devant « *présenter des garanties d'authenticité et de fiabilité* » (Chabin, 2000 : 40) ce qui ne peut être le cas pour des documents numériques textuels, infiniment réinscriptibles. Même dans le cas où elles auraient été archivées par une institution garante, les discussions électroniques n'offrent pas, au niveau du corps des messages, de contenu factuel, mais plutôt du contenu d'opinion ; elles recèlent ainsi des informations précieuses sur les auteurs et leur

51 Les normes les plus couramment rencontrées dans notre corpus sont :
- Norme iso 88-59-1 ou Latin-1 ou Europe occidentale, dont le français.
- Norme iso 88-59-15 ou Latin-9 ou Caractères Latins, dont le français.
- Standard Windows-1252 ou CP1252 utilisé historiquement par défaut par Windows en dans les principales langues d'Europe de l'ouest, dont le français et l'anglais.
- Standard MacRoman, format pour les ordinateurs Apple (système Macintosh).
- Norme iso/CEI 10646 ou UTF-8, jeu de caractère universel.

52 Latzko-Toth Guillaume et Proulx Serge, « Enjeux éthiques de la recherche sur le Web », in Barats, 2013 : 32-52

environnement socio-culturel, mais seulement à titre de « *valeur secondaire de l'archive* » (*ibid*.). Cependant, au vu des métadonnées techniques et documentaires qu'elles présentent, on peut se demander si ces dernières ne peuvent être considérées comme des informations factuelles. Dans tous les cas, on peut affirmer que nos documents participent à un élargissement de l'usage du sens des archives, dans la mesure où une communauté, ici elle des scientifiques et les ingénieurs de l'informatique de réseau, peuvent leur accorder une valeur de témoignage pour l'histoire d'Internet, ou encore celle des chercheurs en SHS qui peuvent accorder une valeur documentaire et analytique à leur mise en corpus.

Nous avons, dans cette perpective, initié une collaboration avec l'équipe du dépôt légal du Web de la BNF qui s'occupe de l'archivage du Web français à but de constituer des collections patrimoniales à l'usage et des chercheurs et du public général (bien que consultables selon en les murs). Nous leur avons soumis une partie de notre corpus Usenet semi-archivé sur le Web par Google sur lequel elle a pu tester des techniques de récupération automatisée des messages. Si ce travail n'est pas achevé, il est important de le mentionner afin de montrer comment certaines institutions peuvent se porter tiers de confiance pour la sauvegarde formalisée de contenus Web.

B. Positionnements du chercheur sur le terrain Internet

Comme pour les autres terrains d'analyse en SHS, le chercheur qui prend Internet comme objet d'étude entretient un rapport à ses données de terrain qui est aussi un rapport aux personnes qui ont produit les informations desquelles les données sont tirées. Observateur, il est aussi utilisateur d'un média d'information et de communication, et en ceci son positionnement tend à rencontrer des zones grises quand il recueille ses matériaux : comment est-il perçu par la communauté qu'il étudie ? La question de la confiance lui est posée alors qu'il est introduit ou s'introduit sur son terrain.

Ce positionnement réflexif doit être mesuré à l'aune des propriétés des médias numériques en réseau en général, du Web en particulier. (Latzko-Toth et Proulx, in Barats, 2013) décrivent ainsi les propriétés à même d'affecter les problèmes et les méthodes de la recherche :

- recherchabilité : les informations du Web sont généralement indexées, et donc trouvables par les moteurs de recherche, ce qui détermine l'accès à des données qui autrement ne seraient pas ou peu visibles ;
- ubiquité : liée à la propriété précédente, elle engage la visibilité des informations, qui ne sont pas bornées à un espace particulier ;
- persistance : l'accumulation des informations est accompagnée par leur rémanence (donc au-delà même de leur suppression) sous la forme de traces, notamment dans les dispositifs de semi-archivage ;
- mutabilité : malgré leur persistance, les informations sont instables et labiles, elles peuvent disparaître totalement ou en partie (parfois une référence existe encore mais sans le contenu) ;
- invérifiabilité : l'identification des acteurs et témoins interrogés peut être complexe, en particulier dans le cas de l'utilisation de pseudonymes .

C. Citer les contenus des discours sur Internet

La citation des résultats en général, mais aussi des énoncés produits par les acteurs plus particulièrement, doit envisager les lois relatives au droit à la vie privée et au respect de la propriété intellectuelle. La tendance générale est d'accorder une attention particulière à la manière dont les acteurs perçoivent leur propre production de contenu et de traces sur les réseaux, et ne pas se reposer sur « *les statuts pseudo-objectifs de leurs écrits* »[53]. Deux conceptions se dégagent dans le rapport du

53 Voir aussi la synthèse proposée par Marie-Anne Paveau (parmi d'autres références précieuses sur l'éthique de la

chercheur à la citation.

Tout d'abord, une **critique de la publicité généralisée des contenus et traces sur Internet** : le fait qu'ils soient accessibles ne veut pas dire qu'ils sont publics par défaut. L'exemple le plus évident est celui des sites protégés par mot de passe ou les groupes que l'on rejoint par adhésion confirmée par l'administrateur ; dans ce cas, il semble important de demander une autorisation aux membres. Mais cela s'applique aussi à des espaces non verrouillés, dans la mesure où certains utilisateurs des médias en réseau considèrent qu'ils ont une conversation privée. Le fait que les informations soient instables (un contenu publié un jour peut être « dépublié » le lendemain), qu'elles gardent une forme de rémanence sous la forme de traces, ou encore qu'elles puissent réapparaître bien après leur disparition dans des initiatives de semi-archivage, participe de cette critique.

Nous avons rencontré ce dernier cas pour la partie de notre corpus issue de Usenet. En effet, la règle tacite générale sur Usenet (issue de contraintes techniques davantage que de la volonté de conserver le caractère privé des communications) a longtemps été que les serveurs par lesquels transitent les messages ne gardent pas de stock plus d'une certaine durée (une semaine ou un mois). Par ailleurs, la communication de groupes via Usenet ne s'est jamais affichée sur des interfaces communes comme celles du Web, mais seulement dans les logiciels de messagerie des utilisateurs, ce qui a pu contribuer à leur conférer un caractère relativement privé (bien que le sens du collectif et de la conversation publique soit très forte sur Usenet, mais borné à l'échelle des groupes). Cependant, des initiatives individuelles (par exemple les chercheurs de l'Université de Toronto déjà évoqués) ou commerciales, plus tard dans les années 1990 (c'est le cas de Deja News, dont le fond d'archives Usenet a été récupéré par Google pour ses semi-archives) ont rendue possible la réapparition de contenus que leurs auteurs d'origine pensaient perdus à jamais. Nous avons étudié les réactions mixtes de ces auteurs (Paloque-Berges, 2013b).

Ensuite, **l'encouragement à respecter une « intégrité contextuelle »** : à savoir non pas essayer de deviner les intentions des acteurs quant au caractère public ou privé de leur production, mais de s'attacher à caractériser la valeur de l'information produite en perspective avec sa publication possible dans le cadre d'une recherche scientifique. Il est donc préconisé, selon un principe de non-aliénation, de ne pas décontextualiser cette information de ses environnements techniques, mais aussi socio-culturels (l'esprit, les principes, les valeurs). Demander le consentement des auteurs va donc dépendre de l'évaluation du chercheur de la sensibilité des informations en ligne, de leur degré de visibilité et des normes tacites ou explicites qui encadrent leur production.

Dans le cas de Usenet, nous nous sommes par le passé imprégnés de ces éléments à l'observation des pratiques, mais aussi en consultant et analysant les nombreuses chartes (Foires aux Questions, nétiquette, guides ; cf. Paloque-Berges, 2013a). Dans le cas des listes, nous avons rencontré une résistance de certains acteurs face à la publication possible des informations ; non pas que leur contenu soit sensible, mais les listes sont traditionnellement des espaces relativement clos, qui réclament une adhésion autorisée (notamment en termes d'identité scientifique pour le cas des listes académiques).

3.2. Archiver, partager et analyser le corpus : comment structurer le corpus

Nous avons voulu ancrer notre travail dans une dimension pratique : que faire, concrètement, des documents numériques générés dans la communication par listes ou groupes de discussion ?

recherche sur les données Internet) dans le billet « Comment citer les échanges en ligne ? Courte note sur l'éthique de la recherche sur Internet », publié sur son blog *Technologies discursives* le 18 juin 2013 [http://technodiscours.hypotheses.org/590].

Comme les manipule-t-on en tant que source documentaire en prenant en compte leur matérialité numérique ? Nous partons ici du principe que l'on travaille dans une forme d'archivage au second degré : le matériau a été récupéré depuis les gisements d'archives semi-formelles du Web et il donne lieu à une nouvelle mise en archive normalisée à destination de la recherche académique. La question de la manipulation est en lien logique avec différents niveaux de traitement des sources :

- pour leur analyse par des logiciels d'analyse instrumentés, afin de permettre l'analyse croisée selon des approches qualitatives et quantitatives ;
- pour leur sauvegarde pérenne, dans un but patrimonial scientifique ;
- pour leur partage avec la communauté scientifique, en vue de leur réexploitation dans différents usages.

Nous finissons donc ce travail en décrivant les réalisations pratiques que nous avons mené à titre de tests, ainsi que les choix convoqués pour ces réalisations. Pour effectuer ce travail, nous avons mis en place plusieurs collaborations :

1/ Avec un groupe de travail de la TGI CORPUS IR (maintenant Huma-Num) dédié à la structuration de corpus linguistiques de communications numériques pour créer des archives partageables avec la communauté scientifique ; la collaboration s'est faite sur le mode d'une prise de conseils sur les standards auprès de ce groupe d'experts.

2/ Avec certains membres du laboratoire STEF (Sciences, Education, Techniques et Formation) de l'ENS Cachan, qui nous ont autorisé à utiliser leur logiciel d'analyse de corpus de type forum (communications électroniques asynchrones).

3/ Plus généralement, avec Gérald Kembellec[54], qui a accepté de nous conseiller et de nous guider dans tout ce qui relève des langages informatiques pour l'organisation des connaissances, leur traitement instrumenté et les questions d'interopérabilité des données de la recherche.

3.2.1. Un volet recherche et intéropérabilité pour la communauté SHS

Nous avons porté notre attention sur des projets en cours de grande ampleur relatifs au partage des données de la recherche scientifique au sein de la communauté académique française.

A. Isidore

La plateforme Isidore[55], sous l'égide de la très grande infrastructure de recherche (TGI) Huma-Num[56], a pour but d'assurer l'accès aux données et aux services des SHS en permettant de faire des requêtes dans des bases de données extérieures qui ont été standardisées pour accepter ces requêtes (processus que l'on appelle le « moissonnage »). Ce travail est donc essentiellement collaboratif : alors que l'équipe s'occupe de créer un moteur de recherche et une interface Web pour la communauté des SHS, la communauté, elle doit mettre à disposition ses documents en intégrant des métadonnées standardisées qui rendent possibles à ces documents d'être trouvés par le moteur d'Isidore.

Parmi les documents requêtables, l'on trouve des bases de données classiques correspondant aux publications et à des « données événementielles » de la recherche académique (appels à communication/projet/publication, annonces d'événements scientifiques...), qui relèvent depuis une dizaine d'années de la communication en ligne de la vie scientifique et à d'autres qui sont plus rares d'accès sur le Web : des données bibliographiques et des corpus de la recherche. Ce dernier type d'accès

54 ATER à l'Institut National de Techniques de Documentation du CNAM pendant notre post-doctorat, Gérald Kembellec est maintenant MCF à l'Université de Lille 3 et membre du laboratoire GERIICO.

55 [http://rechercheisidore.fr/].

56 [http://www.huma-num.fr/].

est celui que nous avons souhaité intégrer à notre projet.

B. CoMeRe

Nous nous sommes rapprochés d'un groupe de travail sur le partage des corpus en SHS au sein de HumaNum (anciennement CORPUS IR). Parmi les consortiums disciplinaires[57] qui le composent, le consortium CORPUS ECRITS inclut un groupe d'orientation linguistique dédié aux « Corpus d'écrits modernes et prise en compte de nouveaux modes de communication » (Groupe 7[58]). Nous avons rejoint ce groupe au cours de l'année 2013 et avons participé à l'assemblée générale du 1er juillet à l'Université de Grenoble, un des partenaires du projet, à l'occasion duquel le groupe a été rebaptisé CoMeRe (Communication Médiée en Réseaux).

Le but du travail de ce groupe concerne le référencement des corpus de communications langagières synchrones et asynchrones via les réseaux Internet par le biais d'une structuration XML des archives de ces communications. L'objectif général est de produire des métadonnées pour ce référencement afin de :

- renseigner les chercheurs futurs utilisateurs des corpus sur les conditions de leur constitution et faciliter l'analyse des données de ces corpus ;
- renseigner les « moissonneurs » (les services qui cherchent des informations et permettent à des utilisateurs de faire des requêtes à distance, comme avec Isidore), en déposant des fiches de métadonnées sur les serveurs Web obéissant à des protocoles de moissonnages ouverts comme l'OAI PMH (Open Archives Initiative Protocol for Metadata Harvesting) permettant de trouver les corpus et les citer ;
- suivre les recommandations de la charte Big Data (définie en 2013) : garantir la traçabilité et la qualité des données, assurer le positionnement éthique du chercheur par rapport à ses corpus et réduire les risques juridiques liés à la citation des contenus.

C. OLAC

Parmi les propositions du groupe, le recours au standard de métadonnées Open Language Archive Community (OLAC) a retenu notre attention dans la mesure où il prend en charge des métadonnées générales qui nous ont semblé importantes pour notre projet. Ces métadonnées intègrent des descripteurs permettant de spécifier le contexte des environnements technologiques de communications médiées par réseaux. La communauté OLAC met en outre à disposition des serveurs « moissonnables » permettant à des moteurs de recherche spécialisés de trouver les corpus de langue qui y sont déposés.

Fondé en 2000, Open Language Archives Community est un partenariat international d'institutions qui cherchent à créer une bibliothèque virtuelle mondiale de ressources linguistiques en développant un consensus sur la meilleure façon de manipuler les archives de ces ressources linguistiques, et en développant un réseau interopérable de stockage et de services pour faciliter la sécurisation et l'accès auxdites ressources. OLAC utilise un format XML pour l'échange de métadonnées de ressources linguistiques dans le cadre de l'Open Archives Initiative[59]. Une « archive linguistique » est définie dans ce contexte comme une collection ou un corpus de ressources linguistiques associé aux ressources descriptives (métadonnées) qui s'y rapportent. Le sens donné au mot « archive » par l'initiative OLAC (suivant les recommandations de l'initiative Archives Ouvertes / OAI, Open Archives Initiative) est un « entrepôt d'informations » et non pas l'archive préservée sur le long terme, de manière autorisée et accompagné d'une politique institutionnelle. Ce ne sont pas des

57 [http://www.huma-num.fr/service/consortium].
58 [http://corpusecrits.corpus-ir.fr/travaux-2/].
59 [http://language-archives.org/OLAC/1.1].

semi-archives pour autant : l'effort, bien qu'indépendant d'une institution, de créer des standards donne une forme d'autorité à l'initiative au sein du domaine scientifique[60].

Développé dans le cadre des réflexions menées sur la normalisation des standards, OLAC est un format très strict et très codifié qui met en jeu quinze terminologies de balise (« Term Name ») :
1. *contributor* : le chercheur ou l'équipe proposant l'archive pour le partage ;
2. *coverage* : la délimitation du corpus ;
3. *creator* : le chercheur ou l'équipe ayant créé l'archive ;
4. *date* : date de création de l'archive ;
5. *description* : la description du corpus ;
6. *format* : format spécifique aux contenus de l'archive ;
7. *identifier* : une référence unique ;
8. *language* : le langage des contenus ;
9. *publisher* : l'éditeur de l'archive, s'il existe ;
10. *relation* : ? ;
11. *rights* : informations juridiques ;
12. *source* : la/les source(s) à partir de laquelle l'archive a été constituée ;
13. *subject* : le sujet des contenus ;
14. *title* : le titre de l'archive ;
15. *type* : le type de contenus linguistiques.

OLAC permet ainsi de décrire par un jeu de métadonnées les auteurs du corpus et différentes informations contextuelles (sources originales, date de création, types de communication, formats, informations juridiques, anonymisation...) et de rendre ces informations disponibles à des serveurs de moissonnage OAI PMH (Open Archives Initiative Protocol for Metadata Harvesting), eux-mêmes requêtables par des structures comme Isidore. Associé à d'autres standards XML permettant de structurer les données en informations pour accompagner l'analyse des contenus linguistiques, il n'est pas suffisant pour formaliser à tous les niveaux les différentes couches de données des courriels décrites plus haut ; mais son approche contextuelle permet cependant de spécifier certains de leurs aspects documentaires. Le travail de formalisation des documents pour l'open data doit ainsi faire attention à fournir des descriptions normalisées, pour permettre l'interopérabilité, et doit aussi prévoir la protection de données potentiellement sensibles comme les noms et les coordonnées des personnes à l'origine des contenus des documents.

3.2.2. Un volet analyse instrumentée : formalisation des documents pour l'analyse SHS

La structuration documentaire du corpus, si nous lui avons assigné de préparer son intéropérabilité à moyen ou long terme, doit aussi se faire pour un usage plus local, qui est celui de l'analyse par le chercheur accompagnée d'instruments informatiques. Le logiciel d'analyse peut servir différentes fonctions, comme :

- aider à chercher de manière plus rapide et plus efficace des éléments dans des corpus numériques qui ont tendance à être larges ;
- mettre en lumière des récurrences lexicales, sémantiques, et plus généralement thématiques et leurs corrélations ;

60 « *Members of the archiving profession have justifiably noted the strict definition of an "archive" within their domain; with connotations of preservation of long-term value, statutory authorization and institutional policy. The OAI uses the term "archive" in a broader sense: as a repository for stored information*. » [http://www.openarchives.org/documents/FAQ.html].

- mettre en évidence des évolutions thématiques, mais aussi temporelles ;
- calculer les relations communicationnelles entre acteurs (des individus à la dynamique de groupe)...

La structuration documentaire en XML permet au logiciel de trouver les repères (« balises ») nécessaires à son traitement des données. Il faut ainsi penser cette structuration, qui est souple et peut être tout à fait adaptée au corpus d'une part, et au logiciel d'autre part, selon des critères que nous devons définir au préalable. Nous proposons ici sous une forme synthétique un certain nombre de critères que nous avons pris en compte pour structurer notre corpus et le soumettre à l'analyse. Ce travail implique de superposer aux archives une interface d'interrogation et de visualisation sur écran des corpus analysés. Nous avons porté notre choix de l'instrument d'analyse sur le logiciel Calico dont nous présenterons les qualités et les limites.

A. Rechercher dans des corpus structurés : quelques points d'entrée analytiques

Nous avons établi une problématisation générale *a priori* de l'analyse des nos corpus afin de mieux prévoir une structuration en XML. La question de départ porte sur la valeur réflexive des communications électroniques sur les outils de communication eux-mêmes : qu'est-ce qui dans les discussions permet de comprendre le rôle joué par les communautés dans le développement social et technique d'Internet en France ?

Il s'agit de repérer des thèmes de discussion récurrents ou présents à certains moments, similaires ou en variation. Pour cela, il faut pouvoir chercher des mots ou expressions clefs, analyser leur fréquence par des relevés d'occurrences, ainsi que leur distribution dans les fils de discussion et dans la participation des acteurs. Le suivi de ces thématiques doit se faire dans l'espace (des fils de conversation différents au sein d'un même groupe ou d'une même liste) et dans le temps (l'extension temporelle d'un même fil de conversation ou d'une même thématique dans des fils chronologiquement différents). A propos des acteurs, il faudra repérer la fréquence de leur participation, leur présence stratégique à certain moments et dans certaines thématiques, les interactions relationnelles entre eux, les liens qu'ils entretiennent à une organisation, ou encore la place qu'ils tiennent dans la coordination du groupe ou d'actions hors-ligne. Dans cette perspective on pourra dégager des dynamiques d'autorités au sein des relations de communications en ligne.

Au-delà de l'analyse sémantique des discours, on voit que d'autres données sont en jeu : identité des acteurs, de leur organisation, métadonnées de date et de lieu ainsi que des indications techniques qui peuvent être intéressantes selon l'objet de la recherche. Par exemple, si l'on étudie la thématique de l'accentuation dans les nouveaux formats d'encodage de textes sur support informatique (une question en effet débattue sur les groupes francophones Usenet au début des années 1990), il peut être intéressant d'analyser aussi les métadonnées de format qui accompagnent les courriels de chacun des interlocuteurs, qui détermine leur usage de signes diacritique par exemple.

Plusieurs choix s'offrent dans la manière de délimiter un corpus restreint de liste ou groupes de discussion à partir des éléments potentiels d'analyse que nous avons dégagés jusqu'à présent, que nous décrivons en soulignant les précautions à garder à l'esprit quand une telle recherche est effectuée dans le cadre de corpus de messages électroniques. Ces choix sont à déterminer avant la structuration, afin que celle-ci les reflète et les matérialise dans son système de balisage – ce qui servira ensuite à mieux chercher dans le corpus grâce à un moteur d'interrogation.

La datation.

Critère de délimitation de corpus le plus évident, il doit être accompagné dans le cas des listes et groupes de discussion d'une attention donnée au format de date et d'heure puisque la plus petite unité temporelle pour l'envoi et la réception des messages est la seconde. La multiplicité des formats

nationaux d'affichage des données temporelles peut rendre complexe la mise en archive de différentes collections, et c'est au niveau du travail d'encodage en XML qu'un format sera choisi et généralisé à toutes les archives.

L'acteur.

Avoir identifié au préalable un acteur permet de rechercher dans de multiples listes et groupes ses interventions électroniques. L'acteur peut être recherché par :

- son nom,
- son adresse électronique,
- son alias (par exemple, dans le cas de Christophe Wolfhugel déjà évoqué, le surnom « wolf », adopté par la communauté francophone de Usenet, permet de repérer certaines de ses interventions non identifiées par son nom complet).

L'organisation.

L'organisation (publique ou privée) depuis laquelle sont émis et reçus les messages électroniques peut être un angle d'entrée.

Dans le cas de corpus immédiatement identifiés comme appartenant à une communication interne, la délimitation se fait par défaut dans le nom de la liste ou sa description. C'est le cas d'un grand nombre de listes archivées sur les sites Renater et CNRS, outils de communication d'un laboratoire ou d'un service, par exemple. Il faut noter cependant que certaines des listes sont « ouvertes », c'est-à-dire qu'elles ouvrent leur inscription à des membres extérieurs soit de manière contrôlée (l'administrateur demande au membre potentiel de justifier sa demande d'inscription), soit de manière libre. Le degré d'ouverture (aux membres d'une unité, d'un établissement, d'un réseau d'établissements, à une communauté d'intérêt ou de pratiques, à tout le monde) est fonction des buts de la liste.

Si l'on souhaite prendre comme critère la participation des membres d'une organisation précise à une liste ou un groupe ne présentant pas de manière évidente de processus de fermeture, on peut s'intéresser aux adresses courriel des participants, qui peuvent être révélatrices de leur appartenance organisationnelle. La question de la représentation de l'individu par rapport à l'organisation depuis laquelle il émet des messages est à évaluer.

L'on peut aussi étudier l'implication d'une organisation en cherchant dans les métadonnées comment elle participe au routage et à la mise en forme des informations émises ou reçues (au niveau du serveur ou du logiciel de traitement des messages par exemple).

La thématique.

Entrée d'ordre sémantique, la thématique peut être relativement complexe à définir. Il peut s'agir d'un simple mot-clef ou d'une combinaison de plusieurs termes (recherchés en séquence fixe ou par occurrence de chaque terme sans contrainte de séquence). Contrairement aux autres entrées, qui suivent la structuration par défaut des métadonnées d'en-tête facilement transportables en XML de manière automatisée, il n'existe pas de moyen simple pour baliser thématiquement un corps de message. Par contre, bien que cela soit restrictif, le sujet du mail (qui est une métadonnées de courriel) peut présenter une indication thématique. Par contre, un logiciel d'analyse comportant un volet lexicométrique pourra aussi fouiller les messages et repérer des occurrences et clusters de termes qui traduiront des thématiques.

L'infrastructure technique.

Du protocole aux métadonnées de routage et de format, l'en-tête des messages contient une multitude d'informations dont l'utilité a déjà été démontrée plus haut (partie 2.).

B. Un outil d'analyse instrumentée : la suite logicielle Calico

Nous avons choisi de nous concentrer sur le test d'une suite logicielle développée par l'UMR de didactique des sciences et des techniques STEF (Sciences, Techniques, Education et Formation) de l'ENS Cachan dans le cadre de l'ERTé CALICO (Communautés d'apprentissage en ligne, instrumentation, collaboration)[61]. Le premier but de cette suite logicielle est d'analyser les processus d'apprentissage interactionnel et collaboratif dans des contextes de communication collective asynchrone (les forums, en priorité) ; mais il s'est révélé un outil intéressant pour analyser tout type de communication électronique asynchrone.

Cette suite nous a été indiquée par François Blondel qui a participé aux tests, après nous avoir entendu parler de notre projet d'analyse de listes et groupes de discussion au colloque « Pour un musée de l'informatique et de la société numérique » (CNAM, septembre 2012). Nommée Calico, elle a l'avantage de proposer une application en ligne permettant plusieurs analyses sur des corpus de type « forum » (discussions électroniques asynchrones)[62]. Nous avons pu obtenir un compte personnel pour tester tout au long de l'année nos corpus.

La suite Calico requiert que les documents à analyser soient convertis dans un format XML dédié, intitulé « XML forum », et obligeant de ranger les données dans des champs précis pour que le logiciel reconnaisse les éléments à analyser. Voici deux exemples d'encodage en XML forum d'extraits de notre corpus.

Le premier est un document complet XML (avec balises d'ouverture de clôture et spécifications du format XML utilisé). Le deuxième est un extrait d'un document XML plus long, incluant un exemple de balisage de l'interaction entre deux courriels, marqué par la balise <msgref id="identifiant du message auquel ce message est lié par une réponse Re : visible dans le <subject>"> (captures d'écran en page suivante). Le code organise les données dans des balises permettant :

- d'identifier le message par un identifiant unique relatif au document XML distinct de l'identifiant unique de message électronique, mais pouvant l'utiliser pour référence ;
- de baliser les métadonnées de l'en-tête du message (*header*) : son heure et date d'envoi, son auteur, son sujet, ainsi qu'une référence au message d'origine si le message est une réponse (<msgref id="id_..."/>) ;
- de délimiter le corps du message à proprement parler ; la structure d'XML Forum étant relativement pauvre par rapport à la richesse des métadonnées de courriel, nous avons choisi de les reporter également dans le corps du message pour qu'elles soient identifiées par le logiciel d'analyse.

61 Installée et maintenue par Emmanuel Giguet, ses différents outils sont le fruit de collaborations avec Nadine Lucas et Benjamin Huynh Kim Bang.

62 [http://woops.crashdump.net/calicorss2/index.php].

```
<?xml version='1.0' encoding='iso-8859-15'?>
<forum name='fr.network.divers-test1'
><message id='id_001'>
<header><author>dup...@corton.inria.fr (Francis Dupont)
</author><subject>probleme au CIRCE
</subject><datetime>1993-02-25 09:42:52</datetime></header><body><content type='text/html'>
<br/>Path: sparky!uunet!arakis.fdn.org!gamb.fdn.org!itesec!ensta!julienas!corton!dupont
<br/>From: dup...@corton.inria.fr (Francis Dupont)
<br/>Newsgroups: fnet.general,fr.network.divers
<br/>Subject: probleme au CIRCE
<br/>Keywords: RENATER
<br/>Message-ID: &amp;lt;2859@corton.inria.fr&amp;gt;
<br/>Date: 25 Feb 93 09:42:52 GMT
<br/>Followup-To: fnet.general
<br/>Organization: INRIA, Rocquencourt, France.
<br/>Lines: 14
<br/>
<br/>D'apres Christian MICHAU, Jeudi 9:54
<br/>
<br/>  Le CIRCE nous previent que son EA est en panne et qu'il n'est donc
<br/>plus accessible provisoirement par Renater ; il en est de meme pour
<br/>les sites en cascade derrier le circe, notamment la passerelle
<br/>de messagerie de Rennes.
<br/>  Une intervention imminente sur le site est attendue.
<br/>
<br/>PS (Francis...@inria.fr) : l'EA est le routeur RENATER du site,
<br/>le CIRCE et RENATER sont donc deconnectes.
<br/>PPS (Francis...@inria.fr) : il n'y a pas de "tickets" prevus en cas
<br/>d'incidents sur RENATER comme cela se fait dans RIPE. Les suggestions,
<br/>protestations, etc..., a ce propos doivent etre envoyees au GIP RENATER
<br/>et pas a moi ! :-)
<br/>
</content>
</body>
</message>
</forum>
```

Premier extrait du corpus converti en XML Forum

```
<message id="id_020">
<header>
<datetime>1994-04-25 13:17:43</datetime>
<author>Nicolas.Pi...@inf.enst.fr (Nicolas Pioch)</author>
<subject>Re: Execution d'un programme dans un document html</subject>
<msgref id="id_019" />
</header>
<body>
<content type="text/html">
[liste WWW-FR &lt;www...@univ-rennes1.fr&gt;]
[fr.comp.infosystemes]
 | J'utilise aujourd'hui httpd 2.18 et mosaic 2.4.
Disons CERN httpd? (beaark)
Et Mosaic for X11?
 | &lt;inc srv "|/usr/bin/date"&gt;
C'est l'ancien format des server-execs NCSA httpd jusqu'a la
version 1.1 incluse.
Seul NCSA httpd permet de faire des server-side include et
server-side exec's.
Pour info, depuis NCSA httpd 1.2, le format a change et est
maintenant respectivement
        <!--#exec cmd="/bin/date"-->
        <!--#include virtual="file.html"-->
Le nouveau format permet de faire beaucoup plus de choses, il y a
de nombreuses autres fonctions (flastmod...)
 | Quelqu'un peut-il me dire si c'est une fonctionnalite qui
 | n'existe pas sur httpd ou si j'ai un pb a rechercher
 | dans mes fichiers de config.
Pas du tout, ca existe tout a fait sur httpd.
Enfin... celui du NCSA.
-- Nicolas
</content>
</body>
</message>
```

Deuxième extrait du corpus converti en XML Forum

Calico est une suite logicielle dans la mesure où elle propose, sur une plateforme Web, plusieurs applications de traitement analytique, chacune répondant à un traitement quantitatif des données du corpus. Certaines des applications transportent les résultats des autres dans leur traitement, afin de créer des analyses croisées. La plateforme affiche tout d'abord une visualisation des messages (« ShowForum ») par unité de messages. Les outils applicatifs sont les suivants :

Colagora. Cet outil liste toutes les occurrences de termes et signes typographiques par ordre de fréquence d'apparition dans le corpus associé au nombre de message dans lequel il apparaît. Cette liste peut être réordonnée par ordre alphabétique ou par le nombre de messages les plus fréquents dans lequel un terme est cité. Associée à cette liste, une fonction permet de sélectionner des termes et de les rassembler dans des thématiques que l'on marque par une couleur choisie entre huit différentes. Cette sélection permettra ensuite de mettre les clusters thématiques de termes en valeur lors des autres analyses. Cela donne la possibilité de sélectionner plusieurs termes, pour voir comment ils sont utilisés ensemble, à quel moment du corpus ils interviennent, chez quels auteurs, utilisés comment, dans quel sens, etc. Ce repérage sémantique est utile en particulier si l'on ne sait pas encore ce que l'on cherche : on peut voir ainsi émerger des termes récurrents et les soumettre à l'analyse *a posteriori*. Cela nous a semblé être un outil intéressant pour notre entrée d'analyse sur les thématiques.

Les limites de l'outil sont la non-discrimination des termes, et donc l'inclusion d'une multitude de signes typographiques et termes de type conjonctions et articles qui introduisent une forme de bruit dans la liste. En revanche, la recherche de séquences complexes de termes est impossible, ce qui limite la sélection à des unités.

Anagora permet de visualiser le déroulement temporel de la discussion sous une forme synthétique à partir de chronogrammes. Il peut être intéressant pour analyser l'évolution quantitative de la participation à la liste ou au groupe au niveau de l'ensemble : repérer des démarrages, des fins de vie, mais aussi des creux, des absences temporaires.

Volagora offre une autre visualisation temporelle des messages, mais paramétrables selon différents critères et sous la forme de graphes. On peut ainsi afficher le nombre de messages émis à plusieurs échelles (jour, semaine, mois, semestre, année), ainsi que voir les jours de la semaine ou du mois où ont été publiés le plus de messages. Est ajoutée à cette analyse graphique de la répartition des communications sur le plan temporel une quantification des messages par longueur de message et par nombres de messages ayant reçu une réponse.

Authagora analyse la participation des auteurs en termes de quantité d'émission de messages. Pour chaque auteur, sa participation est donnée en termes de :

- contribution générale (nombre de messages émis par l'auteur) ;
- nombre de fils de discussion auquel il a participé (le fil étant défini comme une série de messages avec un sujet commun) ;
- nombre de fils qu'il a initié ;
- nombre de message auxquels il a répondu ;
- nombre de messages envoyés auquel il n'a pas été donné de réponse (messages sans fils de discussion).

Cet outil permet d'étudier les dynamiques d'interaction interprétables par la présence et les actions communicationnelles d'un auteur au sein d'un corpus.

Bobinette continue la visualisation temporelle, cette fois en permettant de lire les messages intégralement en les ouvrant individuellement dans des fenêtres superposables. L'apport de Bobinette est qu'il intègre les marqueurs colorés indiquant la présence d'un ou plusieurs termes des clusters thématiques repérés et sélectionnés avec Colagora. Il est ainsi possible de suivre l'évolution d'une thématique au cours du temps, ainsi que regarder de plus près l'utilisation des termes dans un message ou une discussion.

Concordagora fonctionne aussi à partir des marqueurs de thématiques. Pour chaque terme sélectionné et associé à une thématique, il met en comparaison de manière systématique les phrases où sont mentionnés les termes choisis, avec la possibilité de visualiser l'intégralité du message où la phrase est citée.

Calico s'est donc avéré un outil pertinent pour l'analyse de nos corpus, mais comporte un certain nombre de limites :

- la difficulté à formaliser un document XML qui sera accepté par Calico sans erreur de syntaxe ;
- l'absence d'outils d'annotation et de sauvegarde des analyses (mis à part la liste de termes organisés par fréquence d'apparition) ;
- l'absence d'attention analytique aux métadonnées techniques des messages (ce n'est pas son but initial, porté sur les contenus) ;
- une interface qui rend difficile la navigation d'un outil à un autre : les analyses étant perdues quand on passe à une autre sauf cas de transport des résultats d'une application à l'autre.

3.3. Un exercice de conversion XML pour créer un « open corpus »

Afin d'une part de nous familiariser à la structuration documentaire de type XML, mais aussi pour tester son application à nos corpus, nous avons décidé de baliser une partie de nos données à la main. Cependant, dans un but d'efficience, il nous a semblé utile de tester l'automatisation de la structuration XML sur notre corpus. Malgré un grand nombre d'outils de conversion automatique XML, il est difficile d'en utiliser un qui ne soit pas adapté, voire personnalisé pour le corpus en question, en particulier sur des corpus de courriel. Cela n'a jamais été fait auparavant à notre connaissance.

Pour cela, puisque nos compétences ne nous permettaient pas d'écrire un programme de conversion en XML structuré, nous avons engagé avec Gérald Kembellec un groupe d'étudiants sur le projet d'écriture d'un script de conversion automatique des documents du corpus en documents XML interrogeables. Ce travail a été réalisé dans le cadre du projet tuteuré final d'un groupe d'étudiants du Master professionnel DEFI (Documents Électroniques et Flux d'Informations)[63], dirigé par Camille Claverie (MCF à l'Université Paris Ouest Nanterre et chercheuse au laboratoire DICEN du CNAM).

3.3.1. Analyse des contraintes

La structuration de documents en XML implique un encodage des textes selon les balises proposées dans la grammaire du langage sous la forme d'un vocabulaire. Ce vocabulaire est aussi appelé un « espace de nommage » : une collection de noms identifiée par des références fixes (URI :

63 Master Professionnel Sciences humaines et sociales, Mention Information et Communication, Spécialité Documents Electroniques et Flux d'Informations (DEFI), Université Paris Ouest-Nanterre [http://www.u-paris10.fr/formation/master-professionnel-sciences-humaines-et-sociales-br-mention-information-et-communication-br-specialite-documents-electroniques-et-flux-d-informations-defi--414590.kjsp].

Universal Resource Identifier), ces noms se divisant en types d'éléments et noms d'attributs. Les noms, dans un document XML, permettent de faire identifier la structure logique du document par des processeurs de requête ou des moteurs de rendus dirigés par des feuilles de style.

On peut structurer le document XML « à la main », et c'est d'ailleurs pour cela qu'il a été rapidement adopté par un ensemble d'utilisateurs sans qu'il ait été besoin au préalable d'avoir de connaissances en programmation[64]. Comme évoqué précédemment, nous avons nous-même structuré une partie de notre corpus en « XML Forum », un XML adapté au traitement analytique par la suite logicielle Calico (cf. 3.2.2.B). Comme pour tout balisage de texte, l'encodage à la main devient difficile dès que le volume de données devient important. Nous avons donc décidé de trouver un moyen pour automatiser cet encodage ; comme la collecte, elle-même impossible à automatiser pour les raisons exposées ci-dessus, nous prenait un temps considérable et que nos compétences en informatiques étaient trop limitées pour ce travail d'automatisation du balisage, nous avons décidé de déléguer ce projet à un groupe d'étudiants.

Les différentes contraintes de départ étaient les suivantes :

- travailler avec un corpus de texte brut (le « dump d'archive » de la liste DNS), avec les problèmes de format que cela suppose (cf. les problèmes de superposition des encodages en 3.1.3.) ;
- créer une manière d'automatiser la conversion XML sous plusieurs formes : intégrant OLAC, intégrant XML forum, et présentant un XML personnalisé qui balise le plus finement possible le corpus ;
- faire en sorte que ce script de conversion automatique soit utilisable ensuite par n'importe qui sans besoin d'une compétence en informatique avancée, et donc à travers une interface orientée utilisateur.

L'idée générale était donc de préparer l'indexation et l'exploitation par les chercheurs de documents spécifiques, des archives textes de listes ou groupes de discussion :

- pour la recherche fine d'informations dans le corpus,
- pour l'analyse qualitative et quantitative,
- pour l'interopérabilité, dans une logique du web de données, pour la moisson des métadonnées par des moteurs.

3.3.2. Cahier des charges donné aux étudiants et réalisation

Les étudiants ont dû créer un logiciel permettant d'automatiser la conversion des documents en mode texte (un « fichier plat ») en document XML selon différents vocabulaires de balisage adaptés aux besoins de l'analyse. Le script de conversion XML va parcourir toute l'arborescence des listes de discussions contenues dans l'archive de mails, et analyser chaque mail, en déterminer sa structure syntaxique et opérer les balisages XML envisagés. Les étudiants doivent préciser l'espace de nommage (*namespace*) à partir de standards existants.

Il a été demandé que soient pris en compte dans le balisage :

- le plus possible de données des en-tête du message, dont nous avons dégagé l'importance dans l'analyse, en plus des éléments *a minima* (auteur, destinataire, date, sujet) ;
- les formes de l'indentation et retours à la ligne dans les corps des messages, afin de restituer les effets typographiques de structuration textuelle (citations / transferts et leurs chevrons, art

64 Caractéristique des langages formels de balisage (*markup languages*), comme l'avait été avant lui l'HTML (Hypertext Markup Language), premier langage de la famille des structures de balisage pour l'affichage Web, et qui a d'ailleurs assuré le succès du Web en matière de création de contenu par des utilisateurs néophytes.

ASCII / signature, tableaux...).
Le travail des étudiants a aussi dû porter sur l'affinage le plus pertinent dans la description de ces listes : plus la structure est fine, plus elle est interrogeable. Ils peuvent donc proposer d'autres éléments à baliser. Nous leur avons fourni un fichier test : l'intégralité de la liste DNS en mode texte.

En termes de modèle fonctionnel, les étudiants se devaient de produire :
Un script de **conversion automatique** qui soit facile d'utilisation par un non-informaticien :

- prêt à utiliser, avec une installation minimale,
- utilisable à travers une interface graphique orientée utilisateur.

Une **interface d'interrogation** des documents convertis en XML répondant aux mêmes critères d'usabilité que le script.
Enfin, il a été demandé aux étudiants de créer une fonction d'anonymisation automatique dans le script de conversion. Cette tâche, très difficile à réaliser sur des corpus de courriels, dans lesquels un nom peut être mentionné à n'importe quel niveau du message (des métadonnées d'en-tête et de signature au corpus du message, citations incluses), n'a pas été prise en charge par les étudiants.

A. Trois modèles XML pour la mise en œuvre du script de conversion et réalisation

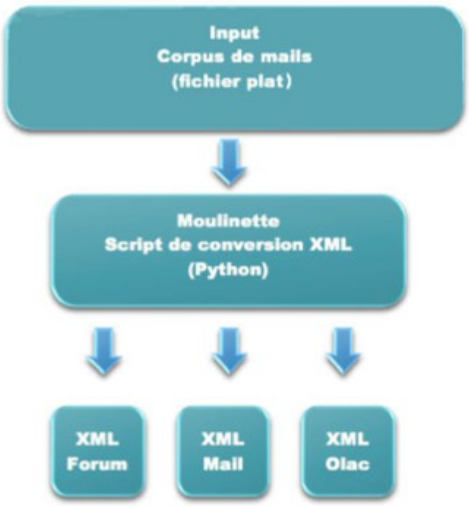

Schéma de la procédure de mise en œuvre

Trois modèles de standardisation XML ont été convoqués ou projetés :

- XML Forum : pour lecture des fichiers par la suite logicielle Calico ;
- XML OLAC : pour favoriser l'intéropérabilité ;
- XML Mail : un balisage « maison » pour tenter de produire un XML au plus près des spécificités du corpus. Il s'agit d'un balisage plus affiné, construit à partir des balises du XML Forum. Le choix des balises se justifie par la structure même du mail dévoilée, notamment par les métadonnées qu'il contient décrites plus haut.

Le programme « lit » le texte brut, le construit, en XML selon les choix de structuration des trois différents types d'XML, et produit un fichier XML final pour chaque type qui comprend l'intégralité de l'archive.

Nous ne nous attarderons pas sur la technologie utilisée pour créer le script de conversion (« moulinette »), en langage de programmation Python ; ce choix a été effectué par les étudiants en vertu des qualités du langage Python pour le traitement des chaînes de caractères, c'est-à-dire l'unité

fonctionnelle de tout texte informatisé. Nous pourrons cependant dès à présent critiquer ce choix car, si Python est multi-plateforme (compatible avec les différents systèmes d'exploitation en vigueur : Windows, Macintosh et Linux), il nécessite une pré-installation sur l'ordinateur de l'utilisateur qui requiert déjà un niveau de compétence (et de compréhension des logiques de programmation). Ce choix est le premier parmi ceux qui ont rendu le travail des étudiants très difficilement exploitable.

Ensuite, des difficultés à la conception ont eu des conséquences fâcheuses à l'utilisation et au niveau des résultats produits :

- En termes de conception matérielle du script, c'est essentiellement l'irrégularité des données qui a pu poser problème, en particulier au niveau de l'encodage des caractères et dans la confusion entre caractères appartenant au « texte » (données et métadonnées de courriel présentes à l'origine) et caractères appartenant au code (par exemple, un signe tel que le chevron « > » est très présent dans les courriel, puisqu'il est un marqueur de citation ; mais il est aussi essentiel au code puisqu'il permet l'encadrement des balises et donc leur reconnaissance par le programme). L'aspect encodage a été mal pris en main par les étudiants, produisant des documents XML remplis de « bruit » (des caractères en trop qui brouillent la lisibilité) et à nettoyer.
- En termes de conception logique, il y a eu une confusion sur l'intérêt d'OLAC. En effet, OLAC est un vocabulaire XML qui sert d'abord à donner des informations contextuelles sur des corpus (qui l'a créé, dans quel programme de recherche, sur quel type de données, etc.). Ses métadonnées sont censées être ajoutées au début de l'archive XML finale en complémentarité avec le reste des métadonnées qui elles structurent le corpus unité de message par unité ; or, elles ont été comprises au cours du projet comme un élément de balisage de chaque unité de message, ce qui n'est pas impossible mais les éléments de vocabulaire en l'état ont été mal adaptés.

B. L'interface de conversion et d'interrogation

Les étudiants ont choisi une plateforme Web open source appelée Jenkins. Le prérequis principal de cette interface était qu'elle fasse le lien entre le document à convertir, le script de conversion, et le document converti une fois passé à la « moulinette », procédure logique censée être simple à installer et utiliser pour l'utilisateur final.

Jenkins est avant tout un « outil d'intégration » (dont la technologie est Java), c'est-à-dire qu'il permet d'exécuter des commandes informatiques à travers une interface de configuration facilement manipulable. C'est une interface Web (navigateur), qui ne nécessite pas une connexion Internet car les programmes permettant de l'utiliser sont installés sur l'ordinateur. Selon les étudiants, ses qualités sont multiples :

- Jenkins est un logiciel gratuit, avec une licence libre qui autorise toute utilisation ;
- l’interface est déjà créée, avec une installation facilitée, un plus faible nombre de conditions préalables sont nécessaires ;
- les fonctionnalités supplémentaires offertes comme la gestion de l’historique, la fenêtre de sélection facilitée des documents à charger sont un plus ;
- et la mise en place de même que la modification des tâches sont extrêmement faciles à gérer grâce à son interface sobre et ses explications.

L’interface doit permettre de transformer un ou plusieurs document(s) au format texte en un fichier XML dont le(s) format(s) a/ont été définis dans la partie précédente. Le document doit par la suite pouvoir être récupéré et utilisé par l’utilisateur.

Cependant, en situation d'utilisation réelle, nous avons trouvé cette interface bien trop complexe pour un néophyte :

- depuis son installation, qui nécessite d'exécuter des commandes informatiques et de déplacer des fichiers dans l'environnement de travail pour créer les liens logiques d'exécution logicielle, ce qui ne peut se faire qu'à l'aveugle pour quelqu'un qui n'est pas familier avec la manipulation de systèmes informatiques au niveau logique ; plus encore, un certain nombre de configurations n'étaient pas adaptées à nos environnements de travail, ce qui a nécessité une plongée dans des niveaux de paramétrage bien trop complexes ;
- jusqu'à son utilisation, qui nécessite de charger des fichiers compressés et de suivre leur construction (*build*), des adaptations de commandes informatiques réclamant une compétence avancée, tout ceci à travers une console très peu intuitive en termes d'interface.

Capture d'écran d'un résultat de « construction » d'un document texte en document XML à travers l'interface Jenkins exécutant le script de conversion.

Parmi les difficultés rencontrées à l'utilisation, nous avons rencontré un problème insoluble sans une reprise approfondie du travail ; en effet, l'ensemble du système logique a été conçu pour travailler sur une machine en particulier, et avec une source en particulier (le fichier test de la liste DNS que nous leur avions fourni). Sur toute autre machine et avec un autre corpus (même de même type), le script ne fonctionnait pas, sauf si l'utilisateur décidait d'adapter le code grâce à des commandes informatiques poussées.

Enfin, en ce qui concerne l'interface d'interrogation, les étudiants ont choisi un moteur d'interrogation adapté à la recherche dans des documents XML (Solr), c'est-à-dire par balises, mais son utilisation s'est révélée de même très problématique :

- la recherche ne peut pas se faire à travers une interface utilisateur, et les résultats s'affichent dans les documents XML, ce qui n'est pas lisible ;
- chaque recherche nécessite une adaptation des paramètres de Jenkins sur le mode « compilation » (transformation d'un langage de programmation en un autre), ce qui nécessite des connaissances poussées en informatique.

C. Conclusions : des résultats négatifs pour des documents complexes

En conclusion, en raison de la difficulté du projet, du certaines confusions sur les fonctionnalités des technologies impliquées dans le projet, et du décalage entre les choix des les

étudiants et le cahier des charges, ce travail a abouti à un résultat final négatif, inexploitable en l'état, mais, nous l'espérons, riche de contre-indications pour qui voudrait retenter l'expérience.

Ces difficultés rencontrées à partir de printemps 2013 (le projet final des étudiants a été rendu mi-juin avec un rapport) nous ont également empêché de passer à la dernière étape de notre projet post-doctoral, qui devait proposer un modèle éditorial supportant la recherche et l'analyse des corpus (au préalable structurés) selon une interface graphique facile d'utilisation. Théoriquement, la conception éditoriale crée un modèle du traitement informatique des documents. Par ailleurs, nous souhaitions inclure dans ce modèle des indications fonctionnelles pour la mise en relation (par étiquetage ou autre forme de marquage) d'une base de données documentaire incluant des informations contextuelles, techniques, historiques permettant d'éclairer les discussions les plus obscures (par exemple un lexique, des articles ou liens vers des ressources Web). Cependant, les efforts consacrés à la compréhension de divers problèmes posés par la conversion XML qui dépassaient nos connaissances sommaires en informatiques n'ont pu être fournis pour cette dernière tâche.

La mise en archive de nos corpus pour le partage et l'analyse scientifiques comporte ainsi de multiples dimensions difficiles à réaliser intégralement en un an sans y dédier spécifiquement un programme d'ingénierie de recherche avec des experts en informatiques et en traitement automatisé de corpus. Objets communicationnels et documentaires complexes, comme nous l'avons vu, ils n'ont à notre connaissance jamais été documentarisés dans le cadre d'un projet d’interopérabilité de la recherche scientifique de type Open data tel que décrit précédemment, le projet CoMeRe n'en étant qu'à sa genèse et n'ayant pas expérimenté ce type d'objets. Pour le versant analyse, ils n'ont pas non plus fait l'objet d'une formalisation en vue d'une analyse instrumentée, le plus approchant que nous ayons trouvé étant le projet Calico, orienté forums Web.

Conclusion

Nous avons engagé une réflexion qui trouve sa place au cœur de celles des Humanités numériques puisqu'elle se penche sur les méthodes pour créer, préserver et exploiter les corpus numériques pour la recherche en SHS et propose une réflexion générale sur ce que les technologies numériques font aux humanités, à leurs terrains et matériaux de recherche. Cependant, cette réflexion ouvre un chantier original, encore très peu balisé, qui est celui des documents numériques natifs comme sources de la recherche en études de Sciences, Technologies et Société, en particulier en histoire et anthropologie des techniques et en sciences de l'information et de la communication. Si ces dernières, dont l'évolution suit de près celle du développement des nouvelles technologies de l'information et médias de communication, se penchent depuis une dizaine d'année sur Internet comme terrain, les premières ne s'aventurent pour l'instant que sur celui des documents analogiques numérisés (à des fins scientifiques et/ou patrimoniales), ce qui pose certaines limites en termes d'accès à des sources diversifiées pour les recherches sur les périodes les plus récentes. Cette réflexion s'est appuyée sur un des corpus les plus réflexifs qui soient en matière d'étude des sources numériques natives : des échanges courriels par listes ou groupes de discussion (nommées jusqu'au début des années 1990 des « conférences électroniques ») entre des collectifs d'acteurs ayant participé au développement d'Internet en France et s'étant vécus comme participant à une communauté de pionniers. Ces courriels, « redocumentarisés » (Pédauque, 2006) plusieurs fois, sont un exemple de documents numériques natifs intéressant car complexe à bien des titres.

Notre travail s'est attaché à mettre au jour les grandes lignes d'une archéologie du document numérique natif en étudiant les mémoires humaines et techniques qu'il a pu porter et les multiples formes par lesquelles il est passé avant d'arriver, prêt à être analysé, sur la table du chercheur. En ceci, le sens des archives numériques ne peut être restreint à celui de l'archivistique ou même de la diplomatique (même si celles-ci ont beaucoup à apprendre à l'archivage numérique, notamment en termes de domestication des technologies numériques par les standards, le droit, et autres dispositifs de normalisation et de légitimation). Au contraire, il doit être ouvert à une réflexion sur les rapports entre la formation des discours et des savoirs au cœur des contextes technologiques.

Tout d'abord, nous avons éclairé ces nouveaux sujets de l'histoire récente de la communication numérique comme des lieux de construction collective et récursive des savoirs sur les réseaux informatiques. Nous avons choisi d'envisager les systèmes de conférences électroniques comme l'un de ces « *lieux de savoir* » de l'ingénierie des réseaux informatiques sur lequel se sont appuyées la sociabilité technicienne et le savoir et savoir-faire associés. Ils sont des terrains d'étude qui intéressent des perspectives d'anthropologie des techniques au sein desquels la pensée de la technologie informatique se formule à la croisée des opérations techniques engagées dans la matérialité de l'écriture numérique et de l'imaginaire discursif des réseaux. Lieux de sociabilité professionnelle, ils intéressent également le regard socio-historique en ce qu'ils présentent l'interaction de réseaux humains et de réseaux techniques dans le déploiement d'une utopie du progrès technoscientifique à travers l'équipement en infrastructures de télécommunication de communautés d'experts. La notion ambigüe d'utilisateur des systèmes informatiques cristallise les tensions entre savoirs experts et savoirs profanes, et permet d'approcher de manière critique l'idée que les réseaux de communication numériques seraient le lieu d'une démocratie technique.

Ensuite, nous avons interrogé la matière générée par ces dispositifs de médiation des savoirs techniques en termes de lieux documentaires de mémoire. Afin de rentrer plus particulièrement dans les méthodologies d'analyse de la mémoire socio-technique portée par les documents qui ont gardé les contenus et traces de ces communications, nous avons, toujours dans la perspective d'une archéologie du numérique, convoqué le concept de documentarisation important pour comprendre le cycle de vie du document numérique natif. Nous avons montré que la mémoire documentaire des échanges de réseau n'était pas seulement une artefactualisation du passé numérique mais une manière de mieux s'organiser dans le présent. La mémoire numérique est une mémoire vivante, sans cesse réinscriptible, toujours à portée de consultation, mais aussi labile, multiforme, marquée par le risque permanent de l'obsolescence. Objets soumis au traitement qualitatif et quantitatif des informations qu'ils portent, ils engagent aussi des précautions d'usage pour le chercheur lui-même par rapport à la mémoire socio-technique qu'il fouille et excave. Ces documents portent ainsi des témoignages mémoriels aussi bien au niveau des contenus, des acteurs que des opérations techniques au niveau des logiciels impliqués dans l'échange langagier en ligne, une mémoire d'usage des technologies de la communication numérique.

Enfin, nous nous sommes attelés à la délicate tâche de transformer nos documents en corpus archivés pour être consultés et partagés et réexploités dans le monde de la recherche en SHS, dans la perspective d'une réflexion pratique et prospective. Nous avons ainsi accompagné un processus de normalisation de corpus de leur récupération dans les « semi-archives » du Web jusqu'à leur mise en archive selon les standards préconisés par les Digital Humanities pour l'intéropérabilité des données de la recherche. Nous nous sommes entourés de collaborateurs pour mener à bien ce travail d'ingénierie documentaire selon la spécificité ou l'étape du travail engagé :

- Gérald Kembellec, qui été notre premier consultant sur ce volet pratique tout au long de l'année et avec qui nous avons choisi de standardiser notre corpus grâce au langage XML, courant dans les Humanités numériques aujourd'hui ;
- le groupe 7 « Communication Médiée par Réseau » (CoMeRe) du consortium ECRITS au sein de la TGIR HumaNum, que nous avons consulté sur les standards de balisage pour le partage des corpus impliquant des contenus langagiers médiés par les réseaux numériques – ceci menant au choix du standard OLAC (Open Langage Archive Community) pour dialoguer en OAI PMH (Open Archives Initiative Protocol for Metadata Harvesting) avec les communautés scientifiques ;
- le laboratoire STEF de l'ENS Cachan qui nous permis de tester sa suite d'applications logicielles Calico, qui analyse les évolutions des interactions de participants à des discussions électroniques en termes thématique, temporel, volumétrique... à partir de communications médiées par ordinateurs de type forum, listes ou groupes de discussion ;
- le Master professionnel DEFI de l'Université Paris Ouest Nanterre, au sein duquel un groupe d'étudiants a dédié son projet tuteuré à créer un logiciel de conversion automatique de documents textes issus des semi-archives de listes de discussion en documents XML prêts pour la recherche et l'analyse. Les résultats de ce travail, négatifs, auront eu le mérite de dégager les difficultés à ce type d'initiatives : l'hétérogénéité documentaire, et notamment la superposition des couches d'encodage, qui brouille les possibilités de conversion en documents XML standards, le recours à plusieurs programmes et systèmes qui vient faire obstacle à la mise en place d'une interface orientée utilisateur non informaticien, la nécessité de faire leur place au texte encodé (lisible par la machine) et au texte affiché (lisible par l'humain), qui fait appel à une conception éditoriale indépendante des modèles fonctionnels complexes.

En définitive, ce travail sur l'analyse des communications médiées par ordinateur a permis de défricher trois nouveaux champs de recherche qui sont encore récents et qui engagent chacun des problématiques méthodologiques d'un nouvel ordre :

- la socio-histoire des réseaux informatiques en France et de ses acteurs, l'anthropologie historique de l'écriture et de la communication numériques, mais aussi, en termes d'historiographie des études en Sciences, Techniques et Société, la prise en compte de nouvelles sources issus des documents numériques natifs qui seront les archives de demain ;
- l'ingénierie des sources, corpus et archives des sciences sociales, avec les nouvelles logiques, promues par les humanités numériques, d'interopérabilité et d'ouverture des données et de leurs résultats, accompagnée par la réflexion épistémologique proposée par les Sciences de l'information, de la communication et du document ;
- et enfin, le rapport des acteurs des sciences et techniques (non plus seulement les communautés SHS) à leur propre patrimoine communicationnel, puisque l'échange électronique est devenu le premier moyen de communication des années 2000, précédant de peu la profusion de nouveaux outils des TIC dans l'accompagnement de la recherche et de la formation académique. Les leçons que nous pouvons tirer des discussions électroniques des chercheurs et ingénieurs de l'informatique de réseau avant la massification d'Internet, si elles ne sont bien sûr pas généralisables pour tout type de communication et d'organisation scientifique, peuvent cependant faire écho aux évolutions des pratiques du monde de la science depuis vingt ans.

Nous avons, pour clore ce post-doctorat, organisé une journée d'étude qui tente de nourrir ces nouveaux champs de recherche. Intitulée « Histoire et patrimoine entre archives et documents numériques: pratiques et épistémologies », elle se déroulera le 9 décembre au CNAM sous l'égide du LabEx HASTEC et en partenariat avec les laboratoires de Sciences de l'information et de la communication (DICEN) et d'Histoire des techniques (HT2S) du CNAM ainsi qu'avec le Projet Exploratoire Premier Soutien (PEPS) PATRIMONIUM financé par le CNRS et piloté par Valérie Schafer – et dont nous faisons partie depuis son lancement au printemps 2013. Le lecteur pourra retrouver le programme en annexe.

Bibliographie

Articles académiques écrits ou publiés pendant le post-doctorat (2012-2013)

« Les archives des réseaux numériques : périmètres, enjeux, défis », avec Valérie Schafer, article de synthèse soumis à la revue *Culture et Recherche* pour son dossier « Les Archives » (en attente de réponse).

« La culture Internet au risque du Web » (avec Valérie Schafer et Fanny Georges), article accepté par la revue *CIRCAV* pour son dossier « Histoire(s) de l'Internet » (accepté sur résumé, article complet en cours d'évaluation).

« Nouvelles sources numériques et logiques d'open corpus : l'intérêt d'archiver et de partager des courriers électroniques » (avec Gérald Kembellec), in *Cahiers de la Société française de Sciences de l'information et de la communication (Sfsic)*, dossier « Figures de la participation numérique: coopération, contribution, collaboration » (article accepté sur résumé, article complet en cours d'évaluation).

« Une pédagogie documentaire par le folklore : analyse des modes d'emploi de l'Internet au temps de la frontière électronique », in *Documentaliste (Sciences de l'information)* n°4/2013 (en cours de publication) (2013a).

« Un patrimoine composite : le public Internet face à l'archivage de sa matière culturelle », in I. Dragan, P. Stefanescu, N. Pelissier, J-F. Tétu et L. Idjeroui-Ravez (éd.), *Traces, mémoire et communication*, Presses de l'Université de Bucarest, 2013 (2013b).

Communications académiques pendant ou à l'issue du travail de post-doctorat

« What network computing does to communication. A retrospective analysis of early debates confronting and inventing online communication ethics » (avec Haud Gueguen et Claire Scopsi), communication acceptée à HaPoC 2013 : 2nd International Conference on the History and Philosophy of Computing 2013, 28-31 Octobre 2013, ENS ULM, Paris (2013c).

« Between electronic frontier and electronic agora: the role of Unix computer networks in France and Europe in the promotion of Internet's technologies and values as a technical democracy », communication donnée à la 6th Plenary Conference of "Tensions of Europe", ANR Resendem: "Democracy and Technology. Europe in Tension from the 19th to the 21th Century", 19-21 Septembre 2013, Sorbonne Paris. Publication acceptée pour les actes du colloque (2013d).

« Telecommunications and Borders », école d'été organisée à la Cité des Télécoms (Orange) par Pascal Griset et Léonard Laborie en septembre 2013, dans le cadre du projet ANR Resendem « Les grands réseaux techniques en démocratie : innovation, usages et groupes impliqués dans la longue durée (fin du 19e - début du 21e s.) » (2010-2014, Irice/Paris 4, CEMMC/Bordeaux 3, Triangle/Ens Lyon, Laboratoire Communication et Politique/CNRS).

« Les communications en réseau comme documents et sources pour l'histoire d'Internet. Problèmes de corpus et d'archives numériques », présentation du travail de post-doctorat à mi-parcours, séminaire des jeunes chercheurs HASTEC, EPHE, 12 avril 2013 ; présentation donnée également au séminaire annuel du laboratoire DICEN le 21 mars 2013 (2013e)

« Problèmes de fouille dans l'archéologie du web social : anciens corpus d'Internet archivés sur le Web », communication donnée au séminaire GERIICO-COPI (organisé par Lucile Desmoulins et Elodie Sevin) pour la séance « L'enquête en SHS sur Internet depuis les "origines" jusqu'au web 2.0. Quels corpus et quelles méthodes ? », Université Lille 3, 28 mars 2013 (2013f).

Références bibliographiques de l'auteur avant le post-doctorat

Entre trivialité et culture : une histoire de l'Internet vernaculaire. Emergence et médiations d'un folklore de réseau, thèse de doctorat en Sciences de l'Information et de la Communication, Université Paris VIII, octobre 2011 (2011a)

« Écritures folkloriques : expérimentation de la communication en réseau », in revue *MEI*, n°33, « Communication, Médias, Littérature », coordonné par Alain Payeur, mars 2011 (2011b)

« Tissages culturels de réseau : l'art ASCII comme écrit d'écran et de code », in revue en ligne *Réel/Virtuel* n°1, « Textures du numérique », Université Paris 1, 2010 (2010a)

« Comparer ou prévoir dans les recherches sur Internet : repenser la médiation technique de réseau à l'aune du comparatisme », in Actes de colloque en ligne du 17ème Congrès de la SFSIC, axe "Mutations médiatiques", Dijon, 23-26 juin 2010 (2010b)

Poétique des codes sur le réseau informatique, Paris : Archives contemporaines, 2009

Références bibliographiques générales

Anis Jacques, *Internet communication et langue française*, Paris : Hermès Sciences, 1999

Bakthine Mikhaïl, *Esthétique et théorie du roman*, Paris : Gallimard, 1978

Barats Christine (éd.), *Manuel d'Analyse du Web*, Paris : Armand Colin, 2013

Barlow John P., « A Declaration of the Independence of Cyberspace », 1996, texte publié sur le site *Electronic Frontier Foundation* [https://projects.eff.org/~barlow/Declaration-Final.html]

Barlow John P. et Kapor Mitchell, "Across the Electronic Frontier", 1990, texte publié sur le site *Electronic Frontier Foundation* [http://w2.eff.org/Misc/Publications/John_Perry_Barlow/HTML/eff.html]

Bourdieu Pierre, *Science de la science et réflexivité*, Raisons d'agir, 2001

Brian Éric, « Archives et mémoire des sciences : enjeux historiographiques », in *Revue d'histoire moderne et contemporaine*, 2001/5 no48-4bis, p. 44-48

Bush Vannevar, « As We May Think », in *The Atlantic Monthly*, July 1945, Volume 176, No. 1, 1945, pp.101-108

Callon Michel, Lascoumes Pierre, Barthe Yannick, *Agir dans un monde incertain. Essai sur la démocratie technique*, Paris, Le Seuil, 2001

Callon, Michel (dir.), *La science et ses réseaux*, La découverte / Conseil de l'Europe/Unesco, 1989

Campbell-Kelly William et Aspray, Martin, *Computer: A History of the Information Machine*, Basic Books, 1997

Carmagnat Françoise, « Une société électronique technicienne face à l'élargissement du réseau. Les usages d'Internet dans un centre de recherche », in Revue *Réseaux*, « Les usages d'Internet », Vol. 14 n°77, 1996

Chabin, Marie-Anne, « Document trace et document source. La technologie numérique change-t-elle la notion de document ? », in *Information-Interaction-Intelligence*, Volume 4, n°1, pp.141-157, 2004

Chabin Marie-Anne, *Je pense donc j'archive: L'archive dans la société de l'information*, Paris : L'Harmattan, 2000

Dacos Marin et Mounier Pierre, « Les carnets de recherche en ligne, espace d'une conversation scientifique décentrée », in C. Jacob (dir.), *Lieux de savoir*, tome 2. « Les mains de l'intellect », Albin Michel, 2011

Ertzscheid Olivier « L'homme est un document comme les autres : du World Wide Web au World Life Web », in revue *Hermès*, 53, 2009, pp. 33-40

Flichy Patrice, *L'imaginaire Internet*, 2001

Foucault Michel, *L'archéologie du savoir*, Gallimard : Paris, 1969

Froissart Pascal. « Internet comme objet scientifique », in *Les nouveaux cahiers de l'audiovisuel*, n°5 « Internet a-t-il une mémoire ? », Paris : Institut national de l'audiovisuel, juin-juillet 2005, pp.50-51

Garçon Anne-Françoise, *L'imaginaire et la pensée technique : Une approche historique, XVIe-XXe siècles*, Classiques Garnier, 2012

Geary Patrick, *Mémoire et oubli à la fin du premier millénaire*, Paris : Aubier/Histoire, 1996

Goody Jack, *La raison graphique*, Paris : Minuit, 1979

Guichard Eric, « Géographie de l'Internet », in C. Jacob (dir.), *Lieux de savoir*, tome 1. « Espaces et Communautés », Albin Michel, 2007

Herrenschmidt Clarisse, Les trois écritures : Langue, nombre, code, Paris : Gallimard/Bibliothèque des

Sciences humaines, 2007

Hert Philippe, « Internet comme dispositif hétérotopique », in G. Jacquinot-Delaunay et L. Monnoyer (dir.) (1999), « Le dispositif, entre usage et concept », in *Hermès*, n° 25

Hert Philippe, « Les arts de lire le réseau. Un cas d'innovation technologique et ses usages au quotidien dans les sciences », in revue *Réseaux*, « Les usages d'Internet », vol. 14 n°77, 1996, pp.85-116

Huitema Christian, *Et Dieu créa Internet*, Paris : Eyrolles, 1995

King John, Grinter Rebecca E. et Pickering Jeanne, "The rise and fall of netville: The saga of a cyberspace construction boomtown in the great divide", in S. Kielser (Ed.), *Culture of the Internet*, Mahwah, NJ: Lawrence Erlbaum Associates, pp.3-34, 1997

Labbe Hélène et Marcoccia, Michel, «Communication numérique et continuité des genres : l'exemple du courrier électronique», in revue *Texto*, 2005 [http://www.revue-texto.net/index.php?id=512]

Le Crosnier, Hervé, « Les journaux scientifiques électroniques ou la communication de la science à l'heure du réseau mondial », in Actes de colloque du CEM - GRESIC. « La communication de l'IST dans l'enseignement supérieur et la recherche : l'effet Renater / Internet », ADBS Editions, 1995

Lebeau Alain, *L'engrenage de la technique*, Paris : Gallimard, 2005

Lefebvre Muriel, *Approche patrimoniale de la communication scientifique. Les écritures (infra)-ordinaires de la recherche*, Habilitation à Diriger les Recherches en Sciences de l'Information et de la Communication, Toulouse Le Mirail, 2013

Lévy Pierre, *Les technologies de l'intelligence : l'avenir de la pensée à l'ère informatique*, Paris : Seuil, 1999 (édition originale : 1990)

Markham Annette N., Baym Nancy K., *Internet Inquiry. Conversations About Method*, Sage publications, 2009

Mandressi Rafael, « Réseaux, généalogies, contrats : collectifs savants », in *Lieux de savoir*, tome 1. « Espaces et Communautés », Albin Michel, 2007

Marcoccia Michel, « L'animation d'un espace numérique de discussion : l'exemple des forums Usenet », in *Document numérique*, 2001/3, vol.5, pp.11-26, 2001

Merton Robert, "The Normative Structure of Science" in Robert K. Merton (ed.), *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago: University of Chicago Press, 1942

Monnoyer Laurence (dir.), « Le dispositif, entre usage et concept », revue *Hermès* n°25, 1999

Moor James H., « What is Computer Ethics ? », in *Metaphilosophy*, Volume 16, Issue 4, pages 266–275, October 1985

Mourlhon-Dallies Florence, Rakotonoelina Florimond et Reboul-Touré Sandrine, « Les discours de l'internet : quels enjeux pour la recherche ? », in *Cahiers du Cediscor*, « Les discours de l'Internet », 2004

Mourlhon-Dallies Florence et Colin Jean-Yves, « Les rituels énonciatifs des réseaux informatiques entre scientifiques », in *Cahiers du Cediscor* « Les discours de l'Internet », 2004

Pastinelli Madeleine, « La mémoire et l'oubli dans l'univers de l'archive totale », in *EspacesTemps.net,* 2009 [http://www.espacestemps.net/articles/la-memoire-et-lrsquooubli-dans-lrsquounivers- de-lrsquoarchive-totale/]

Pédauque Roger T. (dir.), *Le document à la lumière du numérique*, C&F Edition, 2006

Pestre Dominique, « Penser le régime des techno-sciences en société, production, appropriation, régulations des savoirs et des produits techno-scientifiques aujourd'hui », pp. 17 à 43, in Joelle Le Marec (dir.), *Les études de sciences : pour une réflexivité institutionnelle*, Archives contemporaines, 2010

Quarterman John, *The Matrix: Computer Networks and Conferencing Systems Worldwide*, Prentice Hall, 1990

Raymond Eric S., *A Brief History of Hackerdom* (version 1.24), Thyrsus Enterprises, 2000. Version en ligne [http://www.tuxedo.org/~esr/]

Duteil-Mougel Carine et Foulquié Baptiste (éd.), Actes du Colloque d'Albi Langages et Signification (CALS), « Corpus en Lettres et Sciences sociales – Des documents numériques à l'interprétation », Texto, 2006

Ruzé Emmanuel, « Traiter les archives de la toile. Histoire d'un système d'information dans une communauté Wordpress (2003-2008) », in *Entreprises et Histoires* 55, 2009, pp.74-89

Schafer Valérie, « Le Mundaneum, un patrimoine inclassable », in *Hermès*, n° 66, 2013, p. 155-160.

Schafer Valérie, « Internet et avant ? », in Letonturier, É. (dir.), *Les réseaux*, CNRS Éditions, coll. « Les Essentiels d'Hermès », 2012a, p. 73-84

Schafer, Valérie, *La France en réseaux (1960-1980*), volume 1, 2012b, Nuvis

Serres Alexandre, *Aux sources d'Internet : l'émergence d'Arpanet*, thèse de doctorat en Sciences de l'Information et de la Communication, soutenue à l'Université de Rennes 2, 2000

Siess Jurgen, « 'Les missives sont écrites pour inventer le réel' : l'épistolaire dans la perspective de l'analyse du discours », in *Filol. Lingüist. Port.*, n.9, p.369-386, 2007

Smith Marc, « La véridique histoire de l'arobase », séance du 29 janvier 2013 du cycle de conférences « Du rare à l'unique », Ecole des Chartes
http://www.enc.sorbonne.fr/actualite/vie-de-l-ecole/conference-la-veridique-histoire-de-l-arobase-par-

Souchier Emmanüel, Jeanneret Yves, et Le Marec Joëlle (dir.) (2003), *Lire, écrire, récrire. Objets, signes, pratiques des médias informatisés*, Paris : BPI Centre Pompidou

Turner Fred, *Aux sources de l'utopie numérique : De la contre culture à la cyberculture*, C&F Editions, 2012

Turner Fred, "How Digital Technology Found Utopian Ideology: Lessons From the First Hackers' Conference", in David Silver and Adrienne Massanari (eds.), *Critical Cyberculture Studies: Current Terrains, Future Directions*, New York University Press, 2006

Voirol Olivier, "La lutte pour l'interobjectivation. Remarques sur l'objet et la reconnaissance », in Estelle Ferrarese (ed.), *Qu'est-ce que lutter pour la reconnaissance?*, Lormont: Édition Le Bord de l'Eau, 166–186, 2013

Wellman Barry, « Studying Internet Studies Through the Ages », in M. Gonsalvo et C. Ess (éd.), *The Handbook of Internet Studies*, Wiley-Blackwell, 2011

Zacklad Manuel, « Réseaux et communautés d'imaginaire documédiatisées », in Skare, R., Lund, W. L., Varheim, A., *A Document (Re)turn*, Peter Lang, Frankfurt am Main , 2007, pp. 279-297

Annexe :
Programme de la journée d'étude organisée dans le cadre
du contrat-postdoctoral

Cf. pages suivantes.