Multi-layered archives of electronic conversations: how the past of the Internet becomes critically present again on Usenet Web archives

Camille Paloque-Bergès

Multi-layered archives of electronic conversations: how the past of the Internet becomes critically present again on Usenet Web archives

Camille Paloque-Berges (HT2S, Cnam). Short paper for the RESAW 2015 Conference, Aarhus, June 2015.

Humanities and social sciences have been interested in Computer-Mediated Communication through the Internet since the 1990's (Wellman, 2001). Linguists were among the first to create corpora of born-digital language interactions, and the French Research Infrastructure "Corpus" dedicated a workgroup dedicated to "Network Mediated Communication" standardized and interoperable corpora[1] (Chanier et al. 2014). While we focus on CMC archives as sources within our research[2], we also argue that these collections require the attention of heritage institutions outside of the academic field. CMC material, embedded in network environments as traces of past conversations, participate in the formation of an archive in the sense of the foucaldian archive: a layered, discontinuous chain of discourses giving access to the episteme (the conditions of knowledge) of digital networks (Foucault, 1982). An interdiscourse that appears superficially as an aggregation of opinions, it actually reveals how human language interact with computer technologies in many technical and social mediations.

Among CMC genres, mailing-lists and newsgroups are not dependent on the Web, which explains why they have been ignored in the "appraisal process" (Niu, 2012) underlying institutional born-digital archiving, mostly focused on Web media. We think that they are, in fact, weaved into the Web in a way that reveals its multi-layered nature – at theoretical and practical levels. Their historical and heritage value, as well as their technical constraints when collected for building corpora, are arguably interdependent from Web archive sites – a first stage of "vernacular" (non institutional) archiving process. One specific example concentrates general issues raised by Web-archived CMC: the case of the Usenet archive, a text-based collection of group conversations gathered from the most prominent ancestor of digitally-based social networking. This paper is based on two studies we did on Usenet archives as an object of analytical and methodological research:

1/ a genealogy of its successive archival processes, with the archived collections ending up in the Google Groups web service, put in perspective with an analysis of the reception of this first heritage initiative by Usenet users themselves (Paloque-Berges, 2013);

2/ our own collection and analysis of a corpus of newsgroups' messages in order to research a chapter of the French history of computer communication networks, with a socio-technical approach inspired by the sociology of innovation refined with the perspectives of infrastructure and code studies; followed by an attempt at engineering the corpus for structuration, standardization and interoperability purposes (to be published)[3].

1. Accounting for the historical and heritage value of CMC material: from Internet history to Internet heritage

Mailing-lists, a collective form of emailing processed through automated software (since the mid 1970's) and newsgroups, a specific form of collective emailing native to Usenet (since 1979), were defined in their heydays as "electronic conferences" testifying of the first social uses of computer networks (Quarterman, 1990). Both stakeholders, protagonists and witnesses (Rheingold, 1993; Hauben and Hauben, 1997; Hafner and Lyons, 1998; Huitema, 1996; Mounier, 2002) and historians or sociologists of the field (Abbate, 2001; Flichy, 2001) have underlined their role in structuring and deploying the modes of cooperation in the history of computer networks has been[4]. We work at defining source value of the born-digital documents that were produced through these interactions, for the technical, social and cultural history of computer networks, but also for Internet studies more broadly (for instance: Paloque-Berges,

[1] We joined this workgroup in 2012, for our post-doctoral research purposes (see note 3).

[2] Comprised of communication history of computer networks, methodologies for studying born-digital archives.

[3] This research was funded by a post-doctoral research contract (2012-2013) at the Laboratoire d'Excellence HASTEC (History and Anthropology of Knowledge, Technologies and Beliefs) and operated within the French community of universities Hautes Etudes Sorbonne Arts et métiers (Hésam) based in Paris.

[4]

2012, 2013).

Pioneer initiatives in archiving this material for heritage purposes are to be found within Internet-centric communities, especially when the Web started to offer a favorable environment for gathering and accessing mailing-lists and newsgroups archives[5]. Prominent organizations in the technological development of the Internet (IETF, RIPE or W3C) have operated such web-based archiving, facilitated by automated archival processes by mailing-lists management systems like Listserv or Sympa, easily integrated within websites. IETF took one step further in 2012 by issuing a Request for Comment (RFC n°6778) for "Email List Archiving, Web-based Browsing and Search Tool Requirement", a project still under construction.

The case of newsgroups is slightly different from mailing-lists'. Because of its general access purpose, Usenet's newsgroups accommodate a global conversation expanding well beyond the themes of Internet historical cooperation (mostly gathered in the comp.* – for computers – and sc.* – for science – groups). Its massive and international scope might explain why no institution has tried to archive it – although the Library of Congress' decision to archive Twitter content since 2009 testifies that such an interest could technically occur. However, heritagization processes have started outside of the professional and institutional archivist field. In the first decade of Usenet, an implicit rule was to keep newsgroups archives on administrators' servers no longer than two weeks, for resource management limitations. Isolated initiatives, such as the collection gathered at the Zoology department at University of Toronto with material going back to 1981, offered the ground for the first archival file curated and provided by the Internet archive foundation[6] in 2011. Before that, Déjà News Research Service (a Usenet newsgroup subscription-based provider for professionals) led a first Usenet massive archiving initiative in 1995, in order to experiment and offer its users a tool for searching in past conversation through an innovative Web interface.

The archive was bought by Google in 2001 and integrated in their new online collective discussion services, Google Groups. This initiative raised the interest on the part of Internet active users, especially early adopters and more particularly old-timers who had been using Usenet since the 1980s. Popular technology-oriented online media, like Wired or Slashdot, covered it, with enthusiastic focus on the Usenet timeline provided by Google as an entry point to the archive[7]. Two referential levels structure this timeline: how global history events appeared and were discussed on Usenet (for instance, the 1984 Chernobyl catastrophe); how technical and cultural computer history landmarks emerged and were shaped within Usenet communities (for instance, the first call for participation in the Linux project in 1991). This convergence shows how Google was set to legitimize its position as a major player to be in Internet history by calling on early adopters' memory of Internet use – Usenet itself being a major meeting point for Internet players until the rise of the Web. This strategy, however, was scrutinized and criticized by users themselves, at the occasion of numerous debates in the Slashdot forums. On the one hand, it was praised; indeed, the archive was used for what we call "ego-searches" (finding one's contributions in the past conversations) and "alter-searches" (finding common references of events known to Usenet's community). These reflexive practices helped build the collective cultural memory of Usenet (Paloque-Berges, 2012, 2013). One the other hand, it was criticized for unearthing and bringing out of the shadows information relative to individuals, a concern towards the risks associated with displaying personal data and opinions. Actually, Google soon provided a feature for the removal of messages, as an early example of experimental device to solve "right to be forgotten" issues. The ambivalent status of Usenet conversations, public in that newsgroups were open access, but private in the sense that the vast majority of Usenet was composed of small, confidential groups, was thus revealed. It is one of the first and most fundamental impediments to an archival in compliance with the rights of individuals, as it is in terms of Internet research ethics, that calls for a "contextual integrity" when facing the problem of publicity or privacy of network mediated discourse (Latzko-Toth et Proulx, 2013).

2. Testing the potential and limitations of web-based newsgroup archives for building corpora

[5] Before that, they were stored on personal computers and exchanged through rudimentary technologies like File Transfer Protocol (FTP).

[6] Internet Archive [http://archive.org/details/utzoo-wiseman-usenet-archive].

[7] « Get started with Usenet on Google Groups - Groups Help » [https://support.google.com/groups/answer/6003482].

We experimented the practice of handling and exploiting Usenet archives for historical and heritage purposes, with the constraint of standardizing a corpus so it can be operated for uses beyond our own. Our research project tackled the subject of French computer networks history studied through the networks-based exchanges of their participants (researchers and engineers in the field of network computing). Our corpus was comprised of several academic mailing-lists as well as newsgroups from the fr.comp.*[8] Usenet hierarchy. Our methodology for analysis was based in a socio-technical perspective borrowed from the sociology of innovation (Paravel and Rosental, 2003), which we adapted and refined according to the specificity of our material. We thus operated a multi-layered analysis for studying these born-digital and Web-based archives, embracing content, software infrastructures and encoding standards – following advice from Infrastructure studies and Critical Code studies. As such, our focus was the different layers of source material that can be found in electronic conversations, from content to header metadata. We argue they have value both in terms of discursive content analysis and in software and infrastructure analysis.

The Usenet archive is embedded within the Google Groups systems, with a forum-type threaded structure and search engine. These interface and infrastructure require critical analysis, as ambivalence lies at the core of the system. The Usenet archive is considered dynamic: it is incremented as more Usenet activity is produced, similarly to the Google Groups themselves. The embedded structure produces a mutualization between Google Groups and Usenet in terms of interface display and search functions, with no visual separation between the two services and their associated archives. For instance, searching for one keyword outputs results without discrimination in regards to their respective belonging. Thus, a prior knowledge of Usenet's complex hierarchy of thematic groups is crucial to finding relevant material. The same applies to search functions: the user should know of trick search queries such as "fr.*" in order to find all the groups falling under the francophone branch category. We found no detail manual of use, and no way to communicate with the administration through faulty contact forms and dead ends links. There is a general sense of under-maintenance (except from a few changes to the system through the years, resulting in disallowing the most important features[9]). Beyond these impediments, Google Groups do not allow the automated crawling, indexing or retrieving of the archive, blocking any attempts to do so, for obvious reasons regarding the commercial and security issues of Google information systems.

We retrieved by hand all the messages in a 2-year period of a particular group of our corpus, in order to: first, operate a qualitative analysis on all aspects pertaining to messages, both at the level of infrastructure and content; second, test quantitative analysis software[10]; third, experiment how to standardize the corpus for interoperable search and research within the academic community. One particular feature of the archive was relevant; that is, accessing "original messages", making the whole of the message visible (metadata and body content). Our socio-technical methodology required the study of the traces left by infrastructure, and metadata does offer a great view of these traces: along classical header data (sender, receiver, subject, date and time), information on format, protocols and routing are crucial to analyze how communication is materialized. Metadata also proved crucial for the two other tasks. Indeed, as we chose to use XML for structuring our test corpus, metadata elements fit right into the mark up structure, facilitating the automated process to do so. We thus experimented two ad hoc XML structuring: first, with the help of a given template in order to run the quantitative analysis software ("XML Forum"); second, through the frame of the OLAC[11] standardized XML recommendations, as they aim at providing accessible online corpora fit for interoperability.

Eventually, we encountered a few difficulties preventing us to further the fulfillment of our experiment into effective, open, accessible and searchable corpora: anonymization (the names and personal information not being restricted to the metadata categories); noise in the content resulting from

[8] This notation shows that in the francophone hierarchy (fr.*), there is a number of subcategories, such as fr.comp*, which hosts groups dedicated to computer-related themes in the francophone branch (for instance fr.comp.infosystemes).

[9] Such as an overview of quantity of messages per month and year, with links to the actual content categorized chronologically. The difficulty to navigate diachronically in the archive has increased over the years.

[10] The online software suite Calico, developed a team of researchers in Education sciences in order to analyze web forum interactions. It requires the encoding of corpus into « XML-Forum », an XML structure native to Calico (Erté Calico, at UMR STEF [http://woops.crashdump.net/calicorss2/index.php].

[11] The Open Language Archive Community (OLAC) is an international collaborative network working at standardizing natural language-based corpora for research purposes.

the original uses and circulation of messages through different mail readers with their own format parameters. Such a project, far from being an easy task even though we are handling plain text files, requires the expertise of professional digital archivists and programmers in order to be executed properly, even more if the goal is to put back the corpora online for sharing with the community.

Ultimately, Usenet Web archives sites (and web archived CMC in general) are themselves a layered archive of the digital past and present Internet, as we showed through the history of its archival on the Web from the mid-1990s. An object of digital cultural and technical memory, their values lie in interactions inscribed in discourse, but also the technical framing at the interface and infrastructure levels. They are weaved into the history of Web, if not entirely constrained by it. One could argue that a systematic archiving of all digital-born networked conversation is an impossible, if not trivial, task. However, we find critical value in studying the archival processes, whether vernacular or institutional, beyond our own interest in writing the history of communication computer networks. Indeed, the archival of CMC as a new form of recording the everyday online conversations inevitably questions the fear of general network surveillance and the binding of the informal word to digital inscription with resilient memory. Effectively, our relation to digitally archived network communication is marked by an unbalance between the difficulty to grasp the conditions of recording user documents and discourse online for heritage and historical purposes and an acute conscience of the economical and political stakes attached to the exploiting of digital data.

Bibliography

Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. "The CoMeRe Corpus for French: Structuring and Annotating Heterogeneous CMC Genres." *JLCL - Journal for Language Technology and Computational Linguistics* 29, no. 2 (2014): 1–30.

Foucault, Michel. *The Archaeology of Knowledge*. Vintage, 1982.

Hauben, Michael, and Ronda Hauben. *Netizens: On the History and Impact of Usenet and the Internet*. Wiley-Blackwell, 1997.

Huitema, Christian. *Et Dieu créa l'Internet*. Paris : Eyrolles, 1996.

Latzko-Toth, Guillaume, and Serge Proulx. "Enjeux éthiques de la recherche sur le Web." In *Manuel d'Analyse Du Web*, C. Barats (eds.). Paris: Armand Colin, 2013: 32–52.

Hafner, Katie and Lyon, Matthew. *Where Wizards Stay Up Late: The Origins Of The Internet*. Touchstone / S & S International, 1998.

Mounier, Pierre. *Les maîtres du réseau*. Paris : La Découverte, 2002.

Niu, Jinfang, « An Overview of Web Archiving », *D-Lib Magazine*, Vol. 18, N° 3/4, March/April 2012

Paloque-Berges, Camille. "La mémoire culturelle d'Internet : le folklore de Usenet." *Le Temps des médias* n° 18, no. 1 (June 8, 2012): 111–23.

Paloque-Berges, Camille. "Un patrimoine composite : le public Internet face à l'archivage de sa matière culturelle." In *Traces, mémoire et communication*, I. Dragan, P. Stefanescu, N. Pelissier, J. F. Tétu, and L. Idjeroui-Ravez (eds.). Presses de l'Université de Bucarest, 2013: 279–86.

Paravel, Véréna, and Claude Rosental. "Les reseaux, des objets relationnels non identifies ?" *Réseaux* n° 118, no. 2 (April 1, 2003): 237–70.

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

Rheingold, Howard. *Virtual Community - Homesteading on the Electronic Frontier*. MIT Press, 1993.

Wellman, Barry. "Computer Networks As Social Networks." *Science* 293, no. 5537 (September 14, 2001): 2031–34.