Mailing list archives as useful primary sources for historians

Alexandre Hocquet, Frédéric Wieber

Mailing list archives as useful primary sources for historians : Looking for flame wars

Alexandre Hocquet & Frédéric Wieber
AHP-PReST, UMR 7117, Université de Lorraine, Université de Strasbourg, CNRS

Dedicated to Jan Labanowski

Abstract

This paper aims to show the potential of mailing lists archives as primary sources for studying recent History of science. In order to focus on the debates regarding software within the computational chemistry community in the nineties, the corpus we chose consists in a scholarly mailing list. It is a typical corpus from its time, conceived, constructed and maintained by a community. The threaded conversations of the list also constitute a unique rhetorical form in its organization which is technically bound to the Internet-based media of that time.

We first of all present the issues at stake within our research topic and show how relevant is such a corpus to address them. We then discuss the "ethnographic" characteristics and the structure of the corpus. Why has this mailing list been created by computational chemists? Why this list has been successful in the community? What and how scientists have been writing on the list? The structure of the corpus is interesting regarding duration: it is constituted of "flame wars", that is outbursts of heated, short and dense debates, in an ocean of evenly distributed polite messages.

We then show how the relevant flame wars are located and extracted when producing a graphical representation of the number of messages per day over time. Once flame wars are isolated, the arguments exchanged by practitioners are studied precisely in order to show the argumentative structure of the debate and the different positions of different actors.

Keywords

scientific software, computational chemistry, mailing list, flame war, threaded conversation, computer-mediated communication, scientific communication

1. Introduction

Computer-mediated communication, especially since the dawning of the Internet, has proved a valuable source of information and analysis for the humanities and social sciences. This kind of digital corpora gives access to day-to-day exchanges among actors (Markham, 2004), and furthermore, they can also be viewed as structured text and metadata (from software, protocols and formats) that allow the analysis of computer-mediated communication practices (Hansen et al. 2010). Among them, mailing lists have already been the subject of various studies in sociology and communication sciences (Marshall, 2007), (Bury, 2005). Today, in these fields, the interest of studying an electronic mailing list lies in its quasi-invisibility (Beaulieu & Høybye, 2011). Unlike studies of the uses of an "innovative" communication technology, mailing lists allow to focus on the actors everyday exchanges in their uses of a mundane tool.

From the Historian's point of view, mailing list (or usenet) archives can constitute, for recent periods of time, interesting sources that are complementary to more traditional ones (Paloque-Bergès, 2018). An example of treatment of such corpora can be to analyze innovation a posteriori (Paloque-Bergès, 2017) (Leslie, 2016). This particular type of Internet archive is also part of “obscure sources and less visible networks, stoking new life into vernacular terms such as the Net” (Driscoll and Paloque-Bergès, 2017) that allows to document communication practices on the Internet in past (scientific) communities (Paloque-Bergès, 2018). In this respect, the history of mailing lists is another kind of history of the Internet, and mailing lists archives are other kinds of corpora than web archives (Hale et al., 2016) (Nanni, 2017). The difference is that mailing lists are a also a sort of computer-mediated communication, and thus the materialization of an online community and social relationships (Baym, 1998).

The other use of mailing lists we retain in our work is to take advantage of said banality and variety of "anonymous" actors to analyze the debates of that time, similar to a microstoria approach, or, from a History of Computing point of view, a History of people (Ensmenger, 2004). In this paper, we try to show that this type of corpus, constituted of a sort of middle layer of discourse between orality and formality, can give valuable historical information about definite communities, and can thus be of interest for the Historian in general.

As we will argue in this paper, in the specific context of History of Science, the informality of this kind of natively digital corpora allows the unveiling of tensions between the actors, unlike the corpora of published scientific papers. Mailing lists disputes are “the kind of digital material that [people] do not want to archive” (Brunton, 2017). Yet, archives of mailing lists allow to follow precisely the arguments exchanged among scientists regarding debated issues in their field. It thus gives access to something that would be difficult to reconstruct with more traditional material. Some Historians have attempted to depict the issues at stake in academic disciplines through the lens of the analysis of scholarly mailing lists (Paloque-Bergès, 2018). In a study of the Humanist

mailing list, Nyhan (2016) addresses the issue of academic disciplinary identity in the Computing Humanities (and later, Digital Humanities) field. She extracts terms and phrases occurrences along the list time span to assess how academics perceive their field. It is very similar to our work regarding the object studied, the community and its materialization in an online forum, and the search for what is at stake in the community. We want to emphasize however that we depict a situation of tensions. Beyond the whole archive of the entire messages, we focus specifically on threads as a structural unit. We argue that this unit is adequate to analyze debates, and we insist that the most heated debates are the more interesting to reveal the tensions: we are looking for flame wars.

As historians of science, we are interested in the computational chemistry community. "Computational chemists", gathered around the uses of computers in chemistry, belong to a scientific field which started to grow in the eighties, whose aim was to develop "computational tools and techniques [which] offer a new method of attack in the continuing effort [in the chemical community] to obtain chemical information" (Counts, 1987). Thus, the computer is a pivotal element of this scientific community. This is the reason why the actors, that use the computer on a daily basis, have access to electronic mailing lists media even if this communication technology is relatively new in the early nineties. Paradoxically, the epoch we study is relatively new in terms of development of electronic mailing lists, yet the daily use of computer by our actors give them enough ease with this medium relatively early (Pisanty & Labanowski, 1996).

The activity of developing computational tools in this particular scientific community is a source of multiple tensions among the scientists involved as well in the development, the distribution and the maintenance as in the use of software. Developing, using and distributing software leads these actors to reflect on many things among which what scientific activity should or may be, what kind of relationship there is between scientific methods and the software implementing these methods, how their coding work is rewarded, which form of intellectual property to resort to, and whether to commercialize their research products. They also wonder about their ideal concept of the openness of science (Wieber & Hocquet, submitted).

These tensions have to be understood in the broader context of relationships between computational chemistry and the industry in times of mobilization of American universities to produce innovation (Berman, 2012). In the 1980s and 1990s, when the Personal Computer and the workstation democratized computation in the laboratory, a new era of "desktop modeling" (Johnson & Lenhard, 2011) coincided with a growing demand for modeling software, especially from the pharmaceutical industry (Richon, 2008). Molecular modeling software packages became a huge potential market for hardware manufacturers to sell graphics terminals and computing power to the Big Pharma (Hocquet & Wieber, 2017).

To make explicit the tensions among computational chemists raised by software, we rely on the archive of a particular mailing list, the so-called "Computational Chemistry List" (CCL). In the context of this article, we discuss our use as a corpus of the archive of this list. We will first describe the structural characteristics of mailing lists and threaded conversations, and how they represent an interesting genre of corpus. We will then introduce the CCL and the reasons why this specific mailing list is adequate to study the issues at stakes within our study. Finally, we will report a semi-quantitative analysis of the archive of the list that supposedly allows to extract the most interesting threads, and we then provide a qualitative analysis of one example of an interesting "flame war" thread.

2. Corpus charcateristics and structure

2.1. Mailing lists and threaded conversations

A mailing list is an asynchronous Internet forum based on the concept of threaded conversation, using the email address as the identification of the person. It is technically using a « mail exploder » program which redirects the emails that someone is sending to all subscribed people. It is thus a forum where people have to subscribe to read/listen, but not necessarily to write/speak up (although it is in general mandatory). This is actually the case for the CCL where the rules, defined by the moderator, allow anyone to contribute.

The structure of the conversations in mailing lists is hierarchized in a unique way. People send emails to a whole group of subscribed people, and successive replies to an original email define a "thread", i.e. a set of successive emails sharing a common subject header. This thread defines the topic of a so-called "threaded conversation". It must not be confused with an email epistolary correspondence, a genre that shares its structure with classic epistolary correspondence, but lacks the "public sphere" approach of debating in public. The form of communication of threaded conversations is “many to many". Even in the pre-web Internet, most virtual communities have relied on asynchronous threaded conversation platforms as a main channel of communication: Usenet newsgroups, email lists, web boards, and discussion forums, all contain collections of messages in reply to one another (Paloque-Bergès, 2018).

The informal conversation style supported by the basic post-and-reply threaded message structure has proven enormously versatile, serving communities ranging widely in focus and goals. Modern incarnations of threaded conversation are embedded in Facebook wall posts, blog comments, YouTube comments or Wikipedia talk pages (Hansen et al., 2010).

Mailing lists as threaded conversations present some typical characteristics. First, mailing list is a push technology, where the conversations arrive directly in your email client. Second, the identity of the poster is declared through the email address (admittedly, to the extent of what an email address means in terms of identity). Third, contributions to a mailing list are not editable. Once a message is sent, everybody will read it. Fourth, mailing list is an older technology than web forums, one that prevailed the world wide web and that actually does not need it to function. Mailing lists are typical of a pre-web Internet. Using the email in the early nineties is nothing innovative but is also not a common mode of communication between scientists. The CCL grew because the computational chemists were using computers on a daily basis. As the CCL founder and maintainer wrote it in 2001: "Can you imagine that the CCL is more than 10 years old? We were here before the Web and hype..." (CCL message, 2001.05.17-008).

Through the Listserv program and the Internet, the CCL mailing list was thus shaped by the computational chemists, and, in a process of typical “mutual shaping of users and technology”, where social practices and technological innovations influence each other (Boczkowski, 1999), the computational chemistry community organized itself around the CCL.

2.2. The Computational Chemistry List (CCL)

The CCL was created in January 1991, by Jan Labanowski, a computational chemist, by then an employee of the Ohio Supercomputing Center (OSC). The purpose of the list was to gather a fledgling community of researchers, and has been immediately successful. Computational Chemistry was a field in its infancy, the new "users" of computational tools were lacking education in the field (Counts, 1989), and the "developers" of those tools found a unique way to disseminate the products of their research. The primary goal of the CCL was to "educate and get educated" (Labanowski, 2007).

After a few years of having occupied disk space and bandwidth within the OSC, the CCL stopped being supported and maintained by this institution, and it then became a hobby project of Labanowski, first through grants from the NSF, then with the contributions of (some of) its members through donations and sponsoring.

As shown by a survey in 1994 (CCL message, 1994.05.16-009), about a half of the subscribers was residing in the United States, more than a half was working in academic institutions, a quarter of them being graduate students, and a tenth holding senior positions. This represented a wide spectrum of members from academia, government agencies and corporate structures (from the pharmaceutical industry, but also from software or hardware manufacturers), and from different levels of the academic hierarchy (Pisanty & Labanowski, 1996).

This survey has been undertaken in the first years of the CCL and only gives a snapshot of the list in 1994, a time when the CCL was growing in terms of number of messages and number of subscribers. The survey is not independent as it has been performed by the moderator and colleagues, in a time when the growing CCL needed to justify its existence in order to apply to grants and become financially sustainable. Yet, this enthusiastic account is valuable to grasp the characteristics of the CCL at that time. Even though no other quantitative survey of the list exists, it is straightforward to roughly assess its vitality along the years by observing the evolution of the number of daily messages. After a few years of steady growth, it reached a plateau in the middle of the 1990s until the middle of the 2010s.

The CCL, as an English speaking list created in a US supercomputing center was very American, but was also of global scope, as the CCL could allow for European, Japanese and soon people from all around the world to join in. The CCL as a mailing list was also a new and unique mode of communication between very different kind of scholars, not only geographically but also culturally and hierarchically. The CCL was the arena where all the people linked to molecular modeling software one way or another could debate. Its informal mode of communication and the ease to join the CCL community were key to the success and growth of the CCL as a medium, but also as an arena where issues at stake in the community could be discussed freely. Though it is hard to assess the representativity of the population of the CCL subscribers in terms of social or professional or geographical profiles within the computational chemistry community, it seems to be safe to say that grossly each social professional or geographical profile has a loquacious enough character among the CCL subscribers to speak up.

Typical roles in online forums can be observed in the CCL community. In the middle of a vast majority of "lurkers" (i.e. subscribers that read the messages but very rarely intervene), contributors tend to specialize. Some ask many questions, thus opening many threads. Some never open a thread but happily answer to people asking for help, other tend to jump into a conversation once a thread is beginning to grow (Pisanty and Labanowski, 1996).

The participants' roles on the list are even more diverse. For example, among developers of a program owned by a corporate company, some are researchers at academia whereas others are employees of that company. Among users, academic chemists coexist with corporate researchers pertaining to a R&D team in a pharmaceutical or other company. There are also the people in charge of selling software, the marketing force, the people in charge of the maintenance. The molecular modeling software lies at the middle of all these people. It is a boundary object (Star & Griesemer, 1989), with a true identity, but plastic enough to represent something different to every different kind of persons involved, and tailored through arrangements between the different actors.

Scientific software as a boundary object binds the actors together and is the raison d'être of the list, and above all, is the subject of most conversations.

Beyond the issue of the representativity of the computational chemistry community within the CCL, another question arises: the translation of the scientific community practices, discourses and actions into an arena that mirrors it. The CCL, though informal, is an arena where relationships of power or authority exist within the list participants. A few anthropological studies of mailing lists have shown how issues of gender (Bury, 2005) or more generally of differences of status (Marshall, 2007) can interfere in a Internet based conversation. Yet, these relationships are not exactly the same on the CCL as in the mundane scientific life within a laboratory, or in a conference, or in the process of publishing a paper... They lead to new forms of relationships of power in the debates. For example, the technical coding expertise, especially for new languages or operating systems, gives authority on the CCL to the young geek provided he has sufficient wit and boldness to engage into a debate with more seasoned fellows (CCL message, 2008.02.02-004).
In a paper about the birth of Bitnet and Listserv (the technical ancestors to mailing lists software), Grier and Campbell use the idea of “presentation of self” of Goffman to describe the informality of the new medium (Grier & Campbell, 2000). First, from the linguistic point of view, the email as a form of speech has been described as quasi-orality, lacking the formality of written language, especially in its infancy, and even though it is technically a sociotechnical device using (computer-mediated) writing. But quasi-orality is not only informality. It is also an attempt to recreate the sense of a community, a place where several people can exchange, something more difficult in written language (Hert, 1999). The mailing list is also, according to Goffman dramaturgical approach, a “backstage”, a place where the participants of the list interact within the community without acting in front of an audience (the public of a conference, the jury of a thesis defense, the readership of a manuscript...). The backstage does not equate with a “private space”, though. It is, as a matter of fact, quite public. More important is the fact that the backstage is the place where the representation on the frontstage is prepared, to keep with the theatrical metaphor.

It is also to be noted that another, deeper, layer of “backstage” exists when private communications between participants interfere within the thread. Private emails, or even phone calls between participants involved in a threaded conversation may affect the thread “frontstage”. As pieces of information or arguments are exchanged in such private communications, unbeknown of the other participants (and unbeknown of the reader of the mailing list archive), the evolution of the thread can be affected.

2.3. The influence of the moderator: topics and topicality

In a paper revisiting Goffman interactionist framework in the era of social media, Hogan views the facebook posts (or social media in general) as an exhibition (in the museum) more than as a performance (onstage) (Hogan, 2010). The mailing list as a medium shares with the Goffmanian stage metaphor the idea of interaction (via messages) instead of an exhibition, yet the Hogan museum metaphor is relevant in that messages are submitted to a community, and are not addressed to anyone in particular, and communication is asynchronous. Moreover, this exhibition relies on a "curator". In our case, the administrator of the list filters and organizes content, through technical decisions, terms of service and moderation.

The personality and position of the person in charge of the moderation of the list is pivotal for the success of the list. Many scholarly mailing lists were created during the nineties in a wide scope of scientific areas (Hyman, 2003) but very few survived as a useful and lively place of debate. On the one hand, many died of attrition, due to lack of interest or lack of debate. On the other hand, many died violently, because of too passionate and controversial topics leading to mail explosions followed by unsubscriptions (Labanowski, 2007). The right balance that permits a list to survive and to flourish is an essential characteristic of the way the list is moderated. It is also pivotal for defining what is the exact use of the list as an arena. Through the careful definition of the terms of service, through the moderation of spirited debates, through the constant recall of what is recommendable and what is not on the list, the moderator influences what is said and how it is said. Labanowski acts as a censor to enforce the set of rules of the CCL, as he reminds regularly (CCL message, 1995.01.20-017).

The topics encountered in the CCL could be grossly ordered in three categories (Pisanty & Labanowski, 1996). Apart from academic events announcements, the main kind of topic is asking for help for software use. Sometimes welcomed, when the archived question actually helps the community for a publicly available software, sometimes frowned upon when the software is a commercial one, the CCL is the quickest way to get help when you are working in front of your networked computer (CCL message, 1997.05.22-011). Debates about trending scientific topics are also not rare, a kind of subject that can mix pure abstract theories or models with technical considerations like hardware specifications or programming languages (CCL message, 1995.05.15-013).

Commercial announcements, or debates about them, belong also to the CCL core, and the fact that those are clearly and precisely authorized with strict rules reveals that it was not obvious to the whole of the community if and how they should be allowed. The specific terms of service regarding commercial announcements, as designed by Labanowski, is very revealing of the peculiarity of the CCL as a scientific mailing list. Commercial talk (like announcements of new software release) is not banned at all (as it is in most academic forums): software releases are acknowledged as an

important part of computational chemistry everyday life. But it is neither completely without rules: vaporware announcements (i.e. announcements about software that is not yet officially released) or "meet us at the booth" announcements (promotional events) are banned, to ensure the CCL is not becoming an advertising channel for corporate software (Labanowski, 1999).

This is a valuable example of how a moderator/administrator shapes the debates of the list by defining the rules, policies and etiquette: as a matter of fact, very few academic lists allow commercial talk or advertising in their policies, and the fact that the CCL does is a unique instance of allowing software and commercialization as a debated issue.

As mentioned earlier, keeping a strict definition of topicality is a key to the success of the list. Yet, from a historian perspective, topics that evolve into passionate debates, or even disputes, if proved unhealthy for the list itself, are the most valuable pieces of information. Given that discussion is possible and even encouraged if not considered off-topic, then the most controversial tensions within the community generate interesting threaded conversations where a variety of actors within the community can interact.

In a similar manner that the “scientific controversy” is a useful tool for the STS scholar to learn about the scientific, political and social matters at stake, the “flame wars” (the threaded conversations in which the topic is controversial enough, or one of the posters is provocative enough to degenerate into a self-sustaining avalanche of posts (Turnage, 2007)) are a valuable source of information. If some of the mailing lists members appear to be “trolls”, i.e. provocative posters wanting to disrupt the harmony of the community by posting on controversial topics, then Gabriella Coleman argues that the troll forces the community to react, and their teasing is thus something that allows the community to discuss and debate about sensitive matters, forcing the members out of a polite stance and thus forcing them to reveal otherwise concealed opinions (Coleman, 2012). Flame wars reveal tensions within the community, and these tensions are unveiled by the very actors and their debates.

3. Localization and analysis of flame wars

3.1. Data and corpus

The CCL, as part of its program to reach self-funding, has created a website opened to sponsoring. One feature of this website is an open archive of the messages of the list. This archive is open to all and is continuously updated. Each web page representing a posted message lists a subject header (as chosen by the original poster), a timestamp, an email address and the body of the message including the poster signature that further identifies the poster.

The corpus is thus constituted by messages, and these posts are not only pieces of text but they are also metadata: author (as an email address), date (as a timestamp), title (as a message subject

header). There are also further metadata hidden in the archive webpage html code, like sequencing of messages. The signature at the end of the message has a hybrid status between text (some may be quite verbose with the author fave quotes, or an ascii work of art) and metadata, as they represent a further trace to identify the poster (email addresses in the nineties may well be not very explicit). A signature helps to learn more about the poster's institution and social status.

As a complement to a human reading and exploration of the mailing list, mining the threaded conversations can prove a valuable tool to unveil the most fruitful conversations that engage in the most interesting debates from the historian's point of view. As compared to modern "Big data" digital corpora, the thousands messages per years of a mailing list only represent hundreds of megabytes, which is considered small data, by nowadays standards, and can be handled by home computers. But such a corpus shares the same issues as in "Big data" corpora: the retrieval of the corpus material may cause problems. For example, format of time or text encoding may change over time, metadata (especially subject headers defining threads, and thus topics) may be incomplete or inconsistent, volatile or multiple email addresses can be used by a same actor.

The thread of a conversation can be defined in two ways: the sequencing of messages as defined by the metadata or the whole set of the messages sharing a common subject header. Both ways can be problematic. First, the archiving is not consistent over the years in terms of metadata. This is due to the evolution of mail exploder software over the years, but also to the apparition of spam from the mid-nineties on (Brunton, 2013). Given the fact that the CCL is open to any poster (and not restrained to subscribers), Labanowski has been in constant struggle over the years to deal with a flooding of unwanted irrelevant messages submitted to the CCL, and has been forced to imagine tricks to keep them at bay. From a few ads a week in 1995 to thousands of daily spam messages a few years later, the consequence has been to tinker several times the posting protocol, and subsequently the very structure of messages once archived on the website.

Second, unlike in web forums, the definition of a thread (and a topic) in a mailing list is not straightforward and unique: some posters, while answering to a previous message regarding a certain topic, may alter or modify or simply change completely the original subject header. While analyzing, we face the issue of having to manually define if a post belongs to a thread or not, based on considerations about topicality.

Exhaustiveness can also be an issue. The mailing list archive is not absolutely complete. The incessant fight of the moderator against spam is a detriment to the reliability of message posting. To reconstitute the integrity of the list archive, the fact that messages are organized in threads decrease the probability that an important message may be missed. If the missing message is important for the thread and the debate, then it is very probable that it is quoted by another message. Of course, an isolated and mundane missing message that implied no answer will stay unnoticed, but we argue

that we precisely focus on messages that provoke debate and yield arguments in the heated threads. Such emphasis reduces the risk of unnoticed messages. In mailing lists where archival is very organized (like the popular operating system Debian users mailing list), the messages archives can be sorted chronologically or by thread according to the explorer of the archive. To study flame wars in a mailing list in lieu of the whole mailing list archive thus allows to address the issue of incompleteness differently. Some authors chose to study several mailing lists on the same topic (Beaulieu & Høybye, 2011) in order to compare discourses, procedures, but also ways of archiving.

In order to mine the list archive, it has to be transformed into an operable corpus. This is done by scraping the web list archive, then transforming it into a database structured around all the text messages and their related metadata. Once achieved, the corpus can be used in different ways. In all these possible treatments, the idea is to use the thread as the basic unit that structures the way the corpus is processed, because the thread appears to be the most interesting basic structure of the corpus.

For example, in this manner, the computer processing can help to pragmatically locate interesting threads, based on the parameter of the number of messages per thread, which defines the length of each thread. The idea, here, is to consider that the longer the thread, the most probable it consists in a debate that shows tensions within the community. The valuable threads can then be detected by plotting the number of messages as a function of time: the more a thread is engaging, the most likely it deteriorates into a flame war, the more messages it aggregates, the more valuable the thread can be.

We can define an average thread length value which corresponds to a density of messages per day. We verified that a high density of messages is correlated to the presence of long threads. We represent the density of messages vs. time in Figure 1.

[insert Figure 1 - **see Figure 1 in Appendix 1**] [Figure 1 caption: Figure 1 plots averaged thread length on the CCL vs. time. Each peak corresponds to a sudden outburst of a heated thread. The bottom part encompasses a timespan from the creation of the CCL to 2002. The upper part corresponds to a zoom from 1999 to 2001. The last peak indicates the flame war we analyze qualitatively.]

In this graph, each peak represents an average of thread length for each message in a given period of time. A peak in average thread length can thus be seen as a way to detect (potentially interesting) flame wars, and save the fastidious reading of the whole archive.

Most threads consist in two messages: a question and the answer, or a question and the summary of all answers received via private emails. Actually, it was considered bad etiquette in the early nineties to engage in a debate, due to precious bandwidth consumption. This is reflected by the fact

that the first actual flame war, as reported by the first peak on the graph, is occurring no sooner than June 1993, two years and a half after the starting of the list. By the late nineties, though, inflated threads burst periodically every few months.

If we focus on 2001, the timeline unveils four emerging threads. A closer look at April and July allows to unveil two other threads that, unlike the other ones, extend over a larger time span. The localization of these six long threads can be viewed as an operational way to search what topics were debated in 2001.

3.2. Qualitative analysis: a typical flame war

The computer-based identification of long threads is then complemented by a qualitative analysis of their argumentative structure, and of the debated ideas within them, thus revealing the tensions within the community. This is one example of this type of analysis that we propose now by discussing a specific heated conversation thread about a popular software package policy. This thread starts on December the fifth, 2001 and forty-five posts from thirty-three subscribers are sent during ten days. More precisely, after a few posts are exchanged on Wednesday the 5th and the following weekend, this flame war really begins, as a cascade of messages, on Monday the 10th and terminates on Friday the 14th.

The first message, on Wednesday the 5th, is an announcement which seems to be innocuous. A Ph.D. student in Taiwan indicates that the results his laboratory has obtained concerning a benchmark, for PC computers, of a "popular electronic structure program" (thereafter referred to as Foobar[1]) are made publicly available on a webpage (CCL message, 2001.12.05-008). This opening post by a Taiwanese Ph.D. student in a global, yet mostly American, mailing list where senior researchers gather (some of them even iconic) is an example of the CCL as an arena where dialogue exists across hierarchies.

The benchmark, and notably one specific technical question, is of interest to several people: they ask about the makefile to compile the Foobar code with a specific compiler. The "backstage" nature of the list is here manifest. The actors' request involves a very technical and pragmatic question associated with the possibility to run the program in-house on specific PC's architectures. To make sure he is not infringing any rule or etiquette, the initial poster writes: "We'll post the detail [of the makefile] on our website after we make sure that it won't violate the license agreement of [Foobar]" (CCL message, 2001.12.07-008).

But after a few hours, in his last message, he posts: "We have got the information from [Foobar, Inc.] that distributing the modified version of makefile or the instructions is violation to the license agreement" (CCL message, 2001.12.08-004). This is this piece of information that will set the

1 Foobar is the name we use to anonymize this “popular electronic structure program”. We will name Foobar the software package and Foobar, Inc. the corporation which commercializes it.

thread on fire. We can see here how a private communication, in this case between the poster's laboratory and Foobar, Inc., affects the evolution of the thread. After four days of peaceful posts on the list, the flame war begins when this private information is made public. As shown by Figure 2, which portrays the dynamical organization of the thread, an avalanche of messages is then posted from Monday the 10th until Friday the 14th.

[insert Figure 2 - **see Figure 2 in Appendix 2**] [Figure 2 caption: Figure 2 portrays the flame war thread structure. Each node represents a anonymized post (indicating CCL archive reference and date). Each edge represents the citation of a previous post. Grayed out posts are the ones cited in the text]

This flame war is a clear instantiation of the central character of molecular modeling software in this community. More precisely, the discussion is here launched on what the license of scientific software can or must allow. Different criticisms put forward that Foobar licensing policy is too restrictive. For example, a Swiss academic metaphorically translates Foobar policy into the automotive business: "[...] the company's policy, translated to the automobile business, appears to be: "OK, we'll sell you the car (program), but you have to produce the proper key (makefile) yourself.. if you copy the key from someone, we'll sue you... maybe we can give you the key for the trunk". Rather strange way of doing business" (CCL message, 2001.12.10-005). This deliberately provocative post, or "trolling" in today's vocabulary, acts as a "fruitful troll" by forcing the community to react, and even the CEO of Foobar, Inc. can't help to respond.

He replies to the criticisms by pointing out that, unlike many other software vendors, Foobar, Inc. provides the source code of Foobar as well as [...] "makefiles for supported platforms and compilers" (CCL message, 2001.12.10-007). He then adds that making the program run on other platforms, with other compilers and makefiles which have not been tested by the company will lead, if made public, to unreliable versions of the program being used and then to problems for the technical support of Foobar.

The intervention of the CEO illustrates how diverse are the actors who participate to the list. Foobar is at that time a very popular and influential software, absolutely central in the field (Foobar topics are discussed on an almost daily basis on the list). The CCL gathers users, developers and vendors, from academia and the industry, of this popular software. The CEO's defense is articulated around the readability of the source code, which is seen as being fundamental for scientific software because it allows epistemic transparency, and around the question of the support of Foobar.

The thread then splits into two topics: 1/ the issue of the technical support and user-friendliness of Foobar; 2/ the articulation between the availability of the source code, the possibility (or not) to implement the software on different platforms, and the stability and robustness of the software

associated with its protection by Foobar, Inc. The structure of the thread, as shown by Figure 2, is then reticulated into two sub-threads corresponding to these two topics. Even though a CCL subscriber receives the emails chronologically, a mixture of changing subject headers (from "Foobar benchmark" to "Foobar flamewar" via "Foobar makefile policy"...) and indented quoting into the message body allows the subscriber (and the historian) to navigate into the subthreads. Not only are they different in topics, but the posters too are different. It thus allows to evaluate who (among the CCL posters) is interested in which topics.

Regarding technical support and user-friendliness, the discussion starts with a post from a Mexican academic. He emphasizes that most of the questions he has asked to Foobar technical support have finally been answered when asked on the CCL list (CCL message, 2001.12.11-015). If a community of users constitutes a more effective support than the official support, why, he wonders, "[Foobar] doesn't have a more open policy to allow end users to communicate improvements [...]"? The restrictive policy of Foobar is viewed by many frustrated users as a hindrance to end-users improvements in a situation of lack of technical support. The CCL is by the way full of posts of users asking for Foobar support to the community (for example, CCL message, 1997.05.22-007) and of acrimonious replies that the community should not take the burden of support of the salesproduct of a private corporation (CCL message, 1997.05.22-011).

This discussion then leads to the issue of the user-friendliness of Foobar. Another academic, this time an end-user experimentalist instead of a theoretical chemist, considers that an effort has to be made in order to provide a user-friendly interface (CCL message, 2001.12.11-022). His post shows that computational tools and software are being democratized at that time in Chemistry, in particular because they can be implemented on computers which are cheaper and then more accessible. Several messages then discuss this democratization phenomenon. The suggestion of the experimentalist chemist is for example criticized because it will lead to so-called "black-box" software. Some computational chemists reject such black-boxes because of their epistemic opacity.

This sub-topic involves a tension between end-users and lead-users (and developers), the former feeling left out by the lack of user-friendliness and the latter fearing a blackbox syndrome. The CCL is thus a unique arena where these antagonists viewpoints can confront. The diversity of the list not only shows in terms of geography, hierarchy or professional status. It is also present in terms of relationship to software.

The second ramification of the thread shows that "open", in the sense of providing epistemic transparency by making the source code readable, is not always satisfying. The debated question is here to know who can contribute to Foobar, and how it can be used. The issue is made salient with the clash between the availability of the source code, the possibility (or not) to implement the software on different platforms (CCL message, 2001.12.13-006), and the stability and robustness of

the software associated with its protection against modification by Foobar, Inc. (CCL message, 2001.12.13-018). The discussion ends with posts asking for the availability of a comprehensive test suite in order to verify the compilation, as a desired sound scientific practice. It is also interesting that, although the first young poster of the thread was silenced by the start of the flame war, the posters asking for different scientific practices to achieve robustness are junior scientists that find in the CCL a unique arena to express their strong views, and they do not hesitate to confront senior scientists or even the CEO of the corporation himself, something that would be more difficult in other academic spaces where hierarchies are more salient. As a matter of fact, the technical ability (to program in newer languages, or to build a makefile for benchmarks...) gives the younger scientists sufficient authority on the CCL to express themselves, bypassing usual hierarchies, creating de facto hierarchies on the list based on technical expertise rather than scientific status.

The flame war is finally cooling when the further ramifications into sub-sub-threads also makes it harder to follow and the final blow is given on the last day before the weekend when attention level drops.

4. Conclusion

Our aim in this paper has been to highlight the interest, for Historians, in working on mailing lists as specific Internet archive. The instance of flame war we have discussed is, in this context, more interesting as an example of a debate within a list than for the specific arguments exchanged within it (these arguments are discussed in Hocquet & Wieber (2017)). It clearly shows the diversity of the actors debating on the list: computational chemists from different countries, junior as well as senior scientists, end-users, lead-users and software developers and vendors. Each different profile has the possibility to participate in an open debate, which is launched by what we have named a "fruitful trolling". The threaded structure of the debate and the relative informality of the communication medium are prone to reveal the major tensions within the community as regards the central question of software development, licensing, use, support and maintenance. The archive of the list constitutes in this sense a capital historical material.

Yet, in order to make sense and utilize all the possibilities of this archive, others treatments could be achieved. When several flame wars have been located, a diachronic analysis of the content of these debates could be realized in order to show how the tensions raised by software can evolve with the scientific, economic or technical context. Furthermore, along such qualitative analyses of flame wars, a distant, quantitative analysis of the threads based on text mining could also be conducted in order to show what topics appear and disappear on the list over time, and how they relate one to another according to occurrences of words (Rockwell & Sinclair, 2016). The threads can also be regarded as networks. In a work about "very large scale conversations", Sack (2000) offers the idea of a semantic network, that is, a network that lexically maps the semantic proximity of words or

phrases used in the corpus. Within the frame of a study, like ours, that focuses on threads as structural units, a social network analysis could help to depict which threads/topics are engaged by whom and which threads/topics are interconnected by communities of contributors, and this could be achieved by using subject header metadata in order to define bipartite (or bimodal) networks involving threads/topics on one hand and email addresses/actors on the other (Hansen et al., 2010). This work is currently in progress.

Finally, because the CCL archive is constituted of short and dense debates in an ocean of evenly distributed polite messages, the study of threads should be complemented with a way of figuring out how to make sense of this routine ocean of messages to extract information from this "background noise". The flame war we have analyzed is a short-lived debate: it lasted ten days from the first to the last post. Our analysis of this one-off event can also be set in a wider context, by using other archival material. We have put forward that the CCL is an arena that mirrors the community of computational chemists, yet it does not constitute an exact reflection of it. If its informality allows a clear expression of tensions within the community, and its archive constitutes a recording of the debates, the CCL is only one particular backstage, which participate in the construction of the frontstage. Therefore, the pieces of information we harvest in our analysis have to be articulated with other sources, from other backstages (such as laboratory archives) as well as from the frontstage (such as scientific papers or op-eds), to better understand the wider context of the tensions we unveiled. This articulation should permit to understand these tensions both epistemologically and socially.

The difficulty of articulating all these dimensions also lies in the different temporalities of the technical, political and economic contexts. The short timing of a flame war is set in a wider context changing over longer periods of time, that do have an influence over the changes in scientific practices, which possess their own different temporalities. The issue is how to articulate these multiple temporalities within a unique narration, with its own linear temporality, and an element of answer can be found in the different ways to analyze our corpus that would allow to gain information on these different changes.

5. References

Baym, N. K. (1998). The Emergence of On-Line Community. In S. Jones (Ed.), *Cybersociety 2.0: Revisiting Computer-Mediated Communication and Community* (pp. 35–68). Thousand Oaks: SAGE Publications, Inc.

Beaulieu, A., & Høybye, M. T. (2011). Studying Mailing Lists: Text, Temporality, Interaction and Materiality at the Intersection of Email and the Web. In S. N. Hesse-Biber (Ed.), *The*

Handbook of Emergent Technologies in Social Research (pp. 255–274). New York: Oxford University Press

Berman, E. P. (2012). *Creating the Market University: How Academic Science Became an Economic Engine*. Princeton University Press.

Boczkowski, P. J. (1999). Mutual shaping of users and technologies in a national virtual community. *Journal of Communication*, *49*(2), 86–108.

Brunton, F. (2013). *Spam: a shadow history of the Internet. Cambridge*, MIT Press.

Brunton, F. (2017). Notes from/dev/null. *Internet Histories*, *1*(1–2), 138–145.

Bury, R. (2005). *Cyberspaces Of Their Own: Female Fandoms Online*. Peter Lang.

CCL Archive. Message: 1994.05.16-009.

CCL Archive. Message, 1995.01.20-017.

CCL Archive. Message: 1995.05.15-013.

CCL Archive. Message: 1997.05.22-007.

CCL Archive. Message: 1997.05.22-011.

CCL Archive. Message: 2001.05.17-008.

CCL Archive. Message, 2001.12.05-008.

CCL Archive. Message, 2001.12.07-008.

CCL Archive. Message: 2001.12.08-004.

CCL Archive. Message: 2001.12.10-005.

CCL Archive. Message: 2001.12.10-007.

CCL Archive. Message: 2001.12.11-015.

CCL Archive. Message: 2001.12.11-022.

CCL Archive. Message: 2001.12.13-006.

CCL Archive. Message: 2001.12.13-018.

CCL Archive. Message: 2008.02.02-004.

Coleman, G. (2012). Phreakers, Hackers, and Trolls: The Politics of Transgression and Spectacle. In M. Mandiberg (Ed.), *The social media reader* (pp. 99–119). New York: New York University Press.

Counts, R. W. (1987). What is computational chemistry? *Journal of Computer-Aided Molecular Design*, *1*(1), 95–96.

Counts, R. W. (1989). The educational foundation of computational chemistry. *Journal of Computer-Aided Molecular Design*, *3*(1), 95–96.

Driscoll, K., & Paloque-Berges, C. (2017). Searching for missing "net histories." *Internet Histories*, *1*(1–2), 47–59.

Ensmenger, N. (2004). Power to the People: Toward a Social History of Computing. *IEEE Annals of the History of Computing*, *26*(1), 96–95.

Grier, D. A., & Campbell, M. (2000). A social history of Bitnet and Listserv, 1985-1991. *IEEE Annals of the History of Computing*, *22*, 32–41.

Hansen, D. L., Shneiderman, B., & Smith, M. (2010). Visualizing threaded conversation networks: mining message boards and email lists for actionable insights. In *Proceedings of the 6th international conference on Active media technology* (pp. 47–62). Berlin, Heidelberg: Springer-Verlag.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace: fifteen years of british universities on the web (pp. 62–70). ACM Press.

Hert, P. (1999). Quasi-oralité de l'écriture électronique et sentiment de communauté dans les débats scientifiques en ligne. *Réseaux*, *17*(97), 211–259.

Hocquet, A., & Wieber, F. (2017). "Only the Initiates Will Have the Secrets Revealed": Computational Chemists and the Openness of Scientific Software. *IEEE Annals of the History of Computing*, *39*(4), 40–58.

Hogan, B. (2010). The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society*, *30*(6), 377–386.

Hyman, A. (2003). Twenty years of ListServ as an academic tool. *The Internet and Higher Education*, *1*(6), 17–24.

Johnson, A., & Lenhard, J. (2011). Toward a new culture of prediction: Computational modeling in the era of desktop computing. In A. Nordmann, H. Radder, & G. Schiemann (Eds.), *Science Transformed?: Debating Claims of an Epochal Break* (pp. 189–200). University of Pittsburgh Press.

Labanowski, J. K. (1999, April 17). Rules. Retrieved February 26, 2018, from http://www.ccl.net/chemistry/aboutccl/rules/index.shtml

Labanowski, J. K. (2007, January). Free Speech, Quality Control, and Flame Wars. *Academe*.

Leslie, C. (2016). Flame Wars on Worldnet: Early Constructions of the International User. In *International Communities of Invention and Innovation* (pp. 122–140). Springer, Cham.

Markham, A. (2004). Internet Communication as a Tool for Qualitative Research. In D. Silverman (Ed.), *Qualitative Research: Theory, Method and Practice* (2 edition, pp. 95–124). London ; Thousand Oaks, Calif: SAGE Publications Ltd.

Marshall, J. P. (2007). *Living on Cybermind: Categories, Communication, and Control*. Peter Lang.

Nanni, F. (2017). Reconstructing a website's lost past Methodological issues concerning the history of Unibo.it. *DHQ: Digital Humanities Quarterly*, *11*(2).

Nyhan, J. (2016). In Search of Identities in the Digital Humanities: The Early History of Humanist. In J. Malloy (Ed.), *Social Media Archeology and Poetics* (pp. 227–242). MIT Press.

Paloque-Bergès, C. (2017). Usenet as a web archive. Multi- layered archives of computer-mediated communication. In N. Brügger (Ed.), *Web 25: histories from the first 25 years of the World Wide Web* (pp. 227–250). New York: Peter Lang.

Paloque-Bergès, C. (2018). *Qu'est-ce qu'un forum internet ?: Une généalogie historique au prisme des cultures savantes numériques*. OpenEdition Press.

Pisanty, A., & Labanowski, J. K. (1996). Electronic mailing lists and chemical research: a case study. *TrAC Trends in Analytical Chemistry*, *15*(2), 53–56.

Richon, A. B. (2008). An early history of the molecular modeling industry. *Drug Discovery Today*, *13*(15–16), 659–664.

Rockwell, G., & Sinclair, S. (2016). *Hermeneutica: computer-assisted interpretation in the humanities*. Cambridge, Massachusetts ; London, England: The MIT Press.

Sack, W. (2000). Conversation Map: An Interface for Very Large-Scale Conversations. *Journal of Management Information Systems*, *17*(3), 73–92.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420.

Turnage, A. K. (2007). Email Flaming Behaviors and Organizational Conflict. *Journal of Computer-Mediated Communication*, *13*(1), 43–59.

Wieber, F., & Hocquet, A. (submitted). Computational Chemistry as "Voodoo Quantum Mechanics": models, parameterization, and software. *Foundations of Chemistry*.

APPENDIX 1

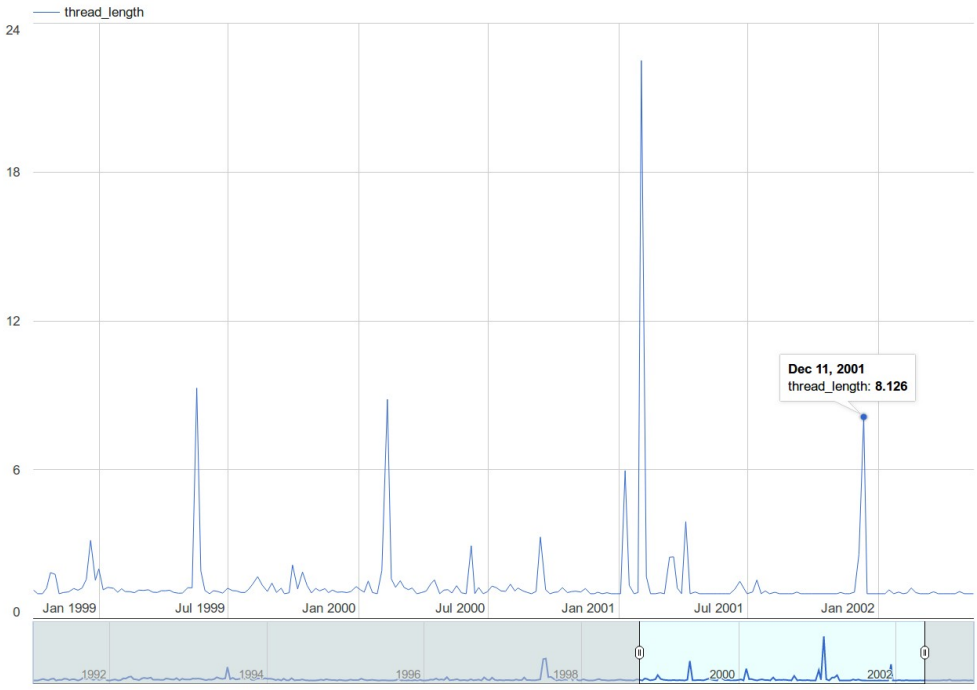

Figure 1

APPENDIX 2 : Figure 2