RxFAQ: Improving the RxNav FAQ Page Through Analysis of Customer Service Emails

Allison Cruise, Associate Fellow 2020-2022

July 2021

Project Sponsors: Olivier Bodenreider, Jenny Rewolinski

Contents

[Abstract 3](#_Toc80087993)

[Objective 3](#_Toc80087994)

[Methods 3](#_Toc80087995)

[Results 3](#_Toc80087996)

[Conclusions 3](#_Toc80087997)

[Introduction 3](#_Toc80087998)

[Objective 3](#_Toc80087999)

[RxNav, RxNorm and Related APIs and Applications 4](#_Toc80088000)

[Procedures 4](#_Toc80088001)

[Studying RxNav and Related Information 5](#_Toc80088002)

[Collection and Organization of Customer Service Emails 5](#_Toc80088003)

[Exploration and Analysis of Emails 5](#_Toc80088004)

[Figure 1 6](#_Toc80088005)

[Figure 2 7](#_Toc80088006)

[Interviews, Meetings and Communication with Team 8](#_Toc80088007)

[Outcomes 8](#_Toc80088008)

[Analysis Document 8](#_Toc80088009)

[Figure 3 9](#_Toc80088010)

[Suggestions Document 10](#_Toc80088011)

[Discussion 10](#_Toc80088012)

[Recommendations 11](#_Toc80088013)

[Works Cited 12](#_Toc80088014)

[Appendices 12](#_Toc80088015)

[Appendix 1: File Cleaning Walkthrough Document 12](#_Toc80088016)

[Raw data 12](#_Toc80088017)

[Logical structure 15](#_Toc80088018)

[Conversations 15](#_Toc80088019)

[A repeat customer’s opinion on whether a memo will seem to us like a reply or a new conversation might not suit the purposes of the present analysis. 16](#_Toc80088020)

[Reduction of noise 16](#_Toc80088021)

[Appendix 3: Suggestions Document 16](#_Toc80088022)

# Abstract

## Objective

The goal of this project was to utilize up to ten years of archived customer email data in a data-driven analysis of service. The ultimate purpose of this analysis is the improvement of the FAQ page on the RxNav website. Through analyzing the past customer service data, a quantitative approach to determining the most frequently asked questions can be achieved.

## Methods

Ten years of customer service data was read, and seven years’ worth of data was deemed to be relevant for the project. This data was then split out into columns, and each relevant question and answer was categorized with tags, to be used for future sorting and analysis. The most frequently asked about products were also collected through this method.

## Results

A total of 524 questions, from 361 different askers, were categorized with 237 individual tags. Fourteen different products from the RxNav website, UMLS, and NLM, were identified as being asked about in the customer questions. The final analysis of the questions and answers was provided in a spreadsheet. A document with suggestions for the organization of the FAQ page was also created by the Associate Fellow responsible for this project.

## Conclusions

The analysis of this customer service data can be used to support data-driven changes to the FAQ page on the RxNav website. The tagging of the relevant questions and answers will allow for future sorting to determine most frequently asked questions. The suggestions document, with the perspective of the naïve user who conducted this project, can also provide guidance for FAQ and website updates. The spreadsheet provided for this project can easily be added to in the future for the potential development of a knowledge base.

# Introduction

## Objective

This project was conceived by members of Lister Hill Center and MEDLARS Management Section, who develop and support the RxNav application. The original project proposal was to create a knowledge base for RxNav by analyzing up to ten years of collected customer service email correspondence received and sent through the RXNAVINFO mailing list.

The realized final project reported here involved analysis of emails received and answered from 2014 to 2021. The goal of the project came to be improvement of the FAQ page available on the RxNav website. Due to the interrelated nature of the RxNav application, the content of the project was not limited to the RxNav application alone and required some knowledge of the RxNorm dataset and RxNorm API, the RxMix application, and other APIs like RxClass and the Drug Interaction API.

Through exploration and analysis of seven years of customer service email correspondence, a final document was produced which categorized every relevant email in the dataset, and a summary of recommendations for the RxNav website documentation and RxNav website FAQ page was provided.

## RxNav, RxNorm and Related APIs and Applications

The RxNav website (rxnav.nlm.nih.gov) is home to the RxNav application, a browser for accessing drug information from multiple sources. RxNav displays data from RxNorm, RxTerms, MED-RT, DailyMed and more. It is free to access and use and can be viewed with any web browser. The RxNav website provides documentation to help users navigate RxNav.

There is also a locally installable application available from the RxNav website called RxNav-in-a-box. This software is free to download and requires a UMLS (Unified Medical Language System) Metathesaurus License agreement, which can be requested through the UTS (UMLS Terminology Services) website. RxNav-in-a-box contains a downloadable version of RxNav, RxClass, RxMix, and the RESTful companion APIs (RxNorm API, Prescribable API, RxTerms and RxClass APIs, and the Drug Interaction API).

The RxNorm API, which is included in RxNav, is a web service providing access to the RxNorm dataset, “a normalized naming system for generic and branded drugs; and a tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems” (“RxNorm Overview”). Documentation for RxNorm is provided through the UMLS website, and documentation for the RxNorm API is available on the RxNav website.

Other products supported by LHC and MMS available on the RxNav website include:

* RxClass, an API and web application for navigation of drug class hierarchies
* RxMix, a browser-based interface which allows you to use functions of the RxNorm, RxTerms, RxClass and Drug Interaction API in interactive or batch mode
* RxTerms, an API for the RxNorm-derived drug interface terminology for prescription writing and medication history recording
* Prescribable RxNorm API, an API for accessing the prescribable content subset of drugs in RxNorm
* Drug Interaction API, an API which provides access to drug-drug interaction information from ONCHigh and DrugBank

All of the described applications, APIs, their documentation, and related products (like UMLS) are included in the customer service correspondence data and were considered for analysis and recommendations.

# Procedures

Once the project sponsors and Associate Fellow Allison Cruise agreed to work together on this project, an initial kick-off meeting was held where the project was discussed, team members were introduced, and first steps for the project were planned. The goal to improve the FAQ page was introduced as early as this kick-off meeting. The customer service email dataset was introduced, but not discussed in length or provided to the Associate Fellow at this time. The team agreed that the first step of the project should be becoming familiar with the applications, APIs and documentation on the RxNav website, as well as the documentation for the RxNorm dataset.

The project timeline was organized in three broad phases: Training on RxNav, RxNorm and other documentation, analysis of customer service email data, and write-up of final recommendations for FAQ content based on that analysis.

## Studying RxNav and Related Information

Allison and Technical Information Specialist In Hye Cho conducted an hour long virtual meeting during which In Hye introduced RxNav and RxNorm, and provided a demonstration of the RxNav browser and the RxMix tool. In Hye also showed Allison where to find the relevant documentation she would need to explore for the project.

After this meeting, Allison independently explored all of the documentation on the RxNav website, as well as the documentation for the RxNorm dataset. After exploring this documentation, Allison and In Hye Cho met again, joined by Jenny Rewolinski, for a more in-depth explanation of the API documentation. During this second meeting Allison was able to ask questions from her own exploration of the documentation.

Reading the documentation for RxNav and related APIs and tools provided necessary background information for understanding the customer service inquiries and team member answers. This process also allowed the Associate Fellow to see the current status of the documentation and make note of potential improvements for that documentation.

## Collection and Organization of Customer Service Emails

Phase 2 of the project was facilitated by Phill Wolf, Aparna Hemchandra and Nicholas McGraw. Customer service emails were collected in an archive dating back ten years, to the beginning of the RxNav services (“Statistics”).

The archived emails were saved in text files, which were cleaned by Phill Wolf who deleted unnecessary text, removed duplicate messages, and ensured the emails were in chronological order. The dates of the emails, subject lines, to and from fields, and the bodies of the emails were preserved for data analysis purposes. A document detailing how the emails were cleaned is available in Appendix 1.

Because of the large size of the final data file, the files were split into 9 smaller text files for easier access. Once these files were legible and available, the exploration step of phase 2 began.

## Exploration and Analysis of Emails

The first step in the exploration of the customer service email data was to read through all of the emails, from most recent to oldest. The archives included a total 5,031 messages, which included multiple questions and answers, as well as some messages not relevant to the project. While reading the emails, a ‘walkthrough’ journal was kept, where notes were taken about email content, recurring subjects, questions and concerns, and thoughts for the next, more thorough analysis of the data.

One outcome of reading all of the archived emails and taking notes was an increased understanding of which content would be relevant toward the end goal of the project (improving the FAQ page) and which content could be ignored during analysis. Another benefit of this step was the ability to make an informed decision about how much of the older content to analyze. After reading all emails, Allison advised the team that the last two, and oldest files, contained information either too outdated to merit inclusion in the FAQ page, or duplicate information which was sufficiently covered by the other seven files. Therefore, further analysis was done only going back to 2014, with the first three years of service excluded. This seven year set of emails 2,360 total messages.

By becoming familiar with the email content, a first example of potential analysis was created by analyzing the emails in the first of the seven text files, which included approximately 92 relevant questions:

### Figure 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Asker | Thread title | Questions | Answers | Notes | Categories/Tags | Products |
| 1 | Customer Name | Subject line of email | Paraphrase of customer question | Paraphrase of provided answer | Any other notes or links | Tags for the question content | APIs or tools mentioned |
| 1 |  |  |  |  |  | Tag for 1 | Second project if mentioned |
| 1 |  |  |  |  |  | Tag for 1 |  |
| 2 | Second customer | Second subject | Second question | Second answer |  | Tag for 2 | Products for 2 |

Figure 1 is an example of what the initial example analysis document contained. Each question was provided with an ID number, allowing it to be linked to multiple tags. Customer names were entered so the paraphrased content could be easily linked to the original archived text if further information was needed. The questions and answers columns contained paraphrased versions of the relevant question and the answer provided in the email exchange. The notes column was added for any additional information the analyzer might want to include. The initial categories/tags column was ill-defined, including overall tags for both question and answer content. The products column was similarly ill-defined, and contained any products mentioned in the questions or answers. Following feedback from the team, the analysis document was revised, and example analysis was conducted on the first and half of the second file, which contained approximately 114 relevant questions:

### Figure 2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Asker | Thread title | Questions | Answers | Notes | Categories/Tags (QUESTIONS) | Categories/Tags (ANSWERS) | Products |
| 1 | Customer Name | Subject line of email | Paraphrase of customer question | Paraphrase of provided answer | Any other notes or links | Tags for the question content | Tags for the answer content | The product the customer was using or asking about |
| 1 |  |  |  |  |  | Tag for question 1 | Tag for answer 1 | Second project if mentioned |
| 1 |  |  |  |  |  | Tag for question 1 | Tag for answer 1 |  |
| 2 | Second customer | Second subject | Second question | Second answer |  | Tag for question 2 | Tag for answer 2 | Product for question 2 |

As Figure 2 shows, the second analysis included tags for both the questions asked, and the answers provided. The second analysis also included the following definitions below each column header:

* ID: Assigned ID tag for each question and the corresponding categories/tags for both questions and answers.
* Asker: The name of the user asking the question as it appears in the email thread.
* Thread title: Subject line of the email thread.
* Questions: Paraphrased version of the actual question the user is asking. Individual questions are split out and tagged.
* Answers: Answer given by team member in the email thread.
* Notes: Any additional notes/field for tagger's use.
* Categories/Tags (QUESTIONS): Categories/Tags assigned for each user question.
* Categories/Tags (ANSWERS): Categories/Tags assigned for team member answers.
* Products: Products referenced/inquired about by USER in question - NOT including products referenced in answers.

The second analysis document included the addition of the more detailed definitions above, as well as a separation of the tags for question and the tags for answers. These changes were added with the intent to make the content of the analysis more meaningful and easier to understand.

The team provided feedback for this second attempt at the analysis, and following their responses, Allison moved forward to analyze all the email content, and complete the third and final phase of the project, providing the analysis document and recommendations. Details of this final analysis will be presented in the outcomes section below. The analysis document itself is available in [Appendix 2](https://www.nlm.nih.gov/about/training/associate/associate_projects/Cruise_Appendix2.xlsx).

## Interviews, Meetings and Communication with Team

Throughout the project phases, meetings were conducted, typically on weekly or biweekly schedules as needed. Due to COVID-19 these meetings were all conducted virtually using the Zoom platform.

Regular communication about the project was also done through the Microsoft Teams platform, where all the RxFAQ team members could participate, and through email. Jenny Rewolinski set up a OneNote project notebook which included background information about the projects and its phases. This notebook could be accessed by any team member with the link.

During the second phase of the project, interviews were also conducted with team members who provide customer service through the RXNAVINFO mailing list. Phill Wolf, Lee Peters, and In Hye Cho agreed to be interviewed. These interviews were casual, half hour conversations about the interviewee’s experiences with customer service, and any recommendations they had for the project. These interviews were a helpful way to understand the expectations of the team members, identify potential themes to look for during analysis, and to gain a deeper understanding of the difficulties of the customer service process and how an improved FAQ page could be useful.

# Outcomes

## Analysis Document

The final analysis document, including emails from 2014 to February 2021, included 524 relevant questions from 361 different askers. There were 165 different tags used for the questions, and 124 tags for the answers. In total, 237 individual tags were used across both the question and answer categories, meaning 51 tags were used in both the question and answer categorizations. A total of 14 product categorizations were used in the Products column and are listed in Figure 3.

### Figure 3

* RxNorm Dataset
* RxNorm API
* RxNav
* RxMix
* RxClass
* RxTerms
* RxNav-in-a-Box
* UMLS Account Issues
* Interactions API
* RxNav
* Prescribable API
* NLM Account Issues
* RxCUI History API

During the analysis process recurring questions were identified. A basic system was used wherein the rows next to repeated questions stated ‘see (number corresponding to previous same question).’ This was done to help identify these questions later on, and to save the time of the analyzer.

While not all questions received were about RxNav or tools and APIs on the RxNav website, related products like the RxNorm dataset and UMLS were included in analysis, since this data could be potentially appropriate for the improved FAQ page.

More in-depth analysis of the findings presented in the analysis document will be conducted by LHC and MMS team members in the future. The in-depth analysis will help identify recurring tags, which can then be used to explore recurring themes and frequent questions, which will help improve the FAQ page in a data-driven way.

While this analysis is not yet available, we can see from the basic ‘see x’ system utilized that 17 questions were repeated, which were noticed and noted by the analyzer. One of the most common questions was regarding pricing for the RxNorm API (i.e., “Is there any pricing information for the RxNorm API?”). The common nature of this question allows us to identify a need for information on pricing of the API, and potentially a need to improve, or make salient, the Terms of Service documentation.

The next most common question identified by the ‘see x’ method was regarding UMLS password resets, and UMLS account issues. These questions, which are about the UMLS license system and not RxNav or its related products, represent a potential opportunity to point users to the UMLS website to get help, which could allow team members to spend more time with RxNav-related questions.

More suggestions for ways to improve the FAQ page were provided in a brief document which will be explained below. The original suggestions document is also available in Appendix 3.

## Suggestions Document

During the project, the UMLS FAQ page was referred to as an example of a potential way to organize the RxNav FAQ page (“UMLS”). This FAQ page’s organization helped to inform the organization of the suggestions.

• Building Queries – It is suggested to provide some common query workflow examples using frequently asked about pathways, to highlight most used API functions, and to explain common misunderstandings identified in analysis (for example, the results file document from RxMix needs to be unzipped with third party software).

• Licensing/Restrictions– As noted above, questions and licensing and usage restrictions are common.

• Contact/’Getting Help’ – Providing contact information could lessen the number of questions received which would be more appropriately asked elsewhere.

• Common Data Elements/TTYs – Under this header were listed some of the term types and content types which were frequently asked about in the customer service emails.

• Data Freshness – This header would contain information about how often information is updated and what data can and cannot be downloaded. These are common questions.

• Other Common Topics: Other frequently mentioned topics, like how the approximate match function operates, were listed here.

• Other Suggestions: Adding a banner encouraging users to share their use case (to help make answers more relevant) and making the link to the FAQ page more prominent on the RxNav website were suggested.

# Discussion

This project was an opportunity to explore RxNav and its related products and APIs, as well as to take an in-depth look at the customer service provided for those products and make suggestions which could improve or ease the service process.

The project took a quantitative, data-driven approach to improvement of the FAQ page. Initial feedback from the final analysis document suggested that the team members were not surprised by the common tags and themes identified. As the team members actively answer customer service questions nearly every day, they are familiar with what is asked. However, this data will not only help to justify updates to the FAQ page, but will help reveal less obvious past questions which may be worthy of inclusion or that could spark improvement to the documentation in other ways. The analysis data could also be added onto, and potentially developed into a knowledge base or wiki which will help future customer service providers when answering customer service inquiries.

An unexpected outcome of this project was the Associate Fellow’s improved understanding of how APIs function. Reading through the API documentation and the questions and answers for the APIs provided new knowledge of how and why APIs are used, frequent problems, and how to answer questions about APIs.

Reading the questions and responses also provided insight into how to write detailed, clear questions, and thorough answers. This experience eventually allowed the reader to interpret questions with limited, or even incorrect information, based on previous similar questions which had been parsed and answered. This ability to interpret questions will be useful when providing service in the future.

As mentioned, the thorough answers provided by various team members over the years, and follow-up questions and answers in the email threads, were insightful as well. Reading and analyzing effective answers will likely improve Allison’s customer service skills, especially when handling technical questions.

Finally, there are the primary goals of this project, analysis and categorization of customer service data in a quantitative method with the intention to improve an FAQ page. This experience has already proven useful for similar projects the Associate Fellow has conducted involving categorization of written comments, which involved creating and utilizing a standard set of categories in a similar method to what was used for the RxFAQ project. This quantitative process could be applied in different environments, with different content, such as items checked out from a library, comments submitted to a feedback form, or analysis of article titles and abstracts provided in literature searches.

# Recommendations

The next step for this project will be to use the final analysis document to draw further conclusions about what information should be included in the improved FAQ page. This may involve using pivot tables to discover the most common tags and explore the content attached to them.

The supplied suggestions document, provided from the Associate Fellow’s naïve perspective, can also help start conversations about the final form of the new FAQ page and potential changes to the RxNav website.

The customer service process will come to include the Microsoft Dynamics platform in the future. This platform allows for the creation of knowledge bases. The information provided by this project could be used as a starting point for a customer service knowledge base and could inform organization of such a resource.

Potential improvement of this project would include a more standard approach to the tags, with a corresponding definition document to help future taggers make decisions about how to categorize questions.

# Works Cited

“RxNorm Overview.” *U.S. National Library of Medicine,* National Institutes of Health, [www.nlm.nih.gov/research/umls/rxnorm/overview.html](http://www.nlm.nih.gov/research/umls/rxnorm/overview.html).

“Statistics.” *U.S. National Library of Medicine*, National Institutes of Health, [rxnav.nlm.nih.gov/Statistics.html](file:///C:\Users\cruiseab\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\DT11UF9K\rxnav.nlm.nih.gov\Statistics.html).

“UMLS - Frequently Asked Questions.” *U.S. National Library of Medicine*, National Institutes of Health, [www.nlm.nih.gov/research/umls/faq\_main.html](http://www.nlm.nih.gov/research/umls/faq_main.html).

# Appendices

## Appendix 1: File Cleaning Walkthrough Document

### Raw data

#### File per month

The listserv generated 133 text files, LOG1001.txt through LOG2102.txt, each corresponding to one month of RXNAVINFO archives.

Most files were retrieved via email and saved by Outlook; two were extracted by CIT. The files retrieved via email and saved by Outlook have a five-line header:

From: NIH LISTSERV (Commands Only)

Sent: Monday, March 1, 2021 9:40 AM

To: Wolf, Phill (NIH/NLM/LHC) [C]

Subject: File: "RXNAVINFO LOG2102"

(blank line)

Following the header, the files saved by Outlook and the files provided by CIT follow the same pattern:

Each file contains a series of messages. These are apparently in chronological order, from earliest to latest.

#### Message

Each message is preceded by a row of “=” signs (73 of them). Then come email headers and a message body. These appear to be in a format described by IETF RFC 2822, “Internet Message Format”.[[1]](#footnote-2)

The email headers, perhaps numbering in the hundreds, include some notable entries:

Mime-Version: 1.0

Content-Type: multipart/alternative;

boundary=45e3b4ea73ba442e05a7be42f0786058e8fd22012c8d25f5d2b7e2f37dad

Message-ID: <rFOvJWEHSNWwDsKI1Uq6qA@geopod-ismtpd-1-5>

Date: Mon, 1 Feb 2021 22:34:30 +0000

Reply-To: RxNorm Navigator Information <RXNAVINFO@LIST.NIH.GOV>,

Andreia Cruz <andreia.cruz@EHEALTH-CONF.ORG>

Sender: RxNorm Navigator Information <RXNAVINFO@LIST.NIH.GOV>

From: Andreia Cruz <andreia.cruz@EHEALTH-CONF.ORG>

Subject: 13th International Conference e-Health,

July 2021: submissions until 1 March 2021

Comments: To: Lachlan <\*\*\*>,

NLM RxNorm Information <rxnorminfo@nlm.nih.gov>

Long headers may have been “folded” onto the next line. Unfortunately, the Listserv archives also contain numerous cases of improper “folding” in which only the header label appears on the first line, and the second line is not indented:

Message-ID:

<DM8PR09MB714493E1C6115BFE6957548F829D9@DM8PR09MB7144.namprd09.p...

The Listserv archives also contained about a dozen corrupt headers that prevented parsing the message. Those were corrected by hand.

After a blank line, the message body follows.

#### Message body

The message body is in one format or another, in keeping with the “Content-Type” message header. Nearly 100 distinct Content-Types are found among the RXNAVINFO messages! These include attachments (such as image/png). The most common, probably significant content, types are variants on “multipart”, “text/plain”, and “text/html”.[[2]](#footnote-3)

3595 CONTENT-TYPE: MULTIPART/ALTERNATIVE

1955 CONTENT-TYPE: TEXT/PLAIN;CHARSET=US-ASCII

1442 CONTENT-TYPE: TEXT/HTML;CHARSET=UTF-8

1367 CONTENT-TYPE: TEXT/PLAIN;CHARSET=UTF-8

1290 CONTENT-TYPE: TEXT/HTML;CHARSET=US-ASCII

1012 CONTENT-TYPE: MULTIPART/RELATED

378 CONTENT-TYPE: TEXT/HTML;CHARSET=WINDOWS-1252

366 CONTENT-TYPE: TEXT/HTML;CHARSET=ISO-8859-1

353 CONTENT-TYPE: TEXT/PLAIN;CHARSET=ISO-8859-1

305 CONTENT-TYPE: MULTIPART/MIXED

267 CONTENT-TYPE: TEXT/PLAIN;CHARSET=WINDOWS-1252

258 CONTENT-TYPE: TEXT/PLAIN;CHARSET=ISO-8859-1;FORMAT=FLOWED

Each content type follows distinct rules set forth in a separate IETF RFC, such as RFC 2387 “The MIME Multipart/Related Content-Type”. Furthermore, MIME multipart message bodies recursively include parts with different mime types.

#### Transfer encoding

Moreover, while “Content-Type” indicates how to interpret the bytes of a message body, “Content-Transfer-Encoding” indicates how to discover what those bytes are. Here, for example, is the body of a message sent by NLM staff in response to a question. At first glance, it is gibberish. Headers indicate Content-Type “text/plain” and Content-Transfer-Encoding “base64”. The text does not look very plain until after you un-base64-encode it.

Content-Type: text/plain; charset="utf-8"

Content-Transfer-Encoding: base64

SGVsbG8sDQoNClRoYW5rIHlvdSBmb3IgeW91ciBpbnRlcmVzdCBpbiBSeE5vcm0uIFRoZSBSeE5h

diB0ZWFtIGlzIGNj4oCZZCBvbiB0aGlzIGVtYWlsIGFzIHRoZXkgY2FuIHByb3ZpZGUgYmV0dGVy

IGluc2lnaHQgYWJvdXQgQVBJIGVycm9yIG1lc3NhZ2VzLg0KDQoNCkJlc3QgcmVnYXJkcywNCklu

IEh5ZSBDaG8NCk5hdGlvbmFsIExpYnJhcnkgb2YgTWVkaWNpbmUNCk5hdGlvbmFsIEluc3RpdHV0

ZXMgb2YgSGVhbHRoDQppbi1oeWUuY2hvQG5paC5nb3YNCg0KRnJvbTogTGFjaGxhbiA8bGZyYW5j

a3hAZ21haWwuY29tPg0KU2VudDogTW9uZGF5LCBGZWJydWFyeSAyMiwgMjAyMSA1OjM4IEFNDQpU  
(...)

We can decode the base64-encoded body with the “base64” Linux command:

$ echo 'SGVsbG8sDQoNClRoYW5rIHlvdSBmb3IgeW91ciBpbnRlcmVzdCBpb...' | base64 -d

Hello,

Thank you for your interest in RxNorm. The RxNa

### Logical structure

Considering the variety of content-types and content-transfer-encodings, it is not practical to interpret the files as a text stream, e.g., with awk.

We used the Sun/Oracle implementation of J2EE javax.mail. We separated the Listserv files at the “===” boundaries, corrected the illegal header folding, and submitted each Listserv message block to the javax.mail library for interpretation. We checked the logs for errors, hand-edited some corrupt Listserv archives, and re-ran the process.

The javax.mail library translates a Listserv text block to a “MimeMessage” with headers and content; the content may be text, some kind of non-text blob, or “MimeMultipart”; the latter, in turn, contains parts annotated with a MIME type. The content of each part may be text, non-text blob, or “MimeMultipart” again, and so on and so forth, infinitely!

There are, in turn, two flavors of MimeMultipart in the Listserv archives. Some MimeMultipart instances give a choice of parts (“multipart/alternative”) in which one of the parts might be plain text and another might be HTML text; or all parts might be graphics of various dimensions. Other MimeMultipart instances contain a sequence of parts (“multipart/mixed”) in which one part is usually the gist of the message (perhaps as a nested “multipart/alternative”) and other parts are attachments or signature graphics.

### Conversations

We might like to group messages by “conversation” and remove duplicate occurrences of text from each conversation.

In the realm of e-mail, messages may be related into conversations (or “threads”) by Message-ID and In-Reply-To headers, or by Subject. (Some messages have a “Thread-Topic” and “Thread-Index”, but these are useful only within an institution that uses Microsoft mailers.)

* A message may have a “Message-ID” header. It may also have a “In-Reply-To” header that refers to another message’s Message-ID. Presumably an email tool, such as Outlook, attaches such a header when the user chooses to “Reply”.
* Messages may also be related into conversations by Subject, allowing for some variety of “Re:” prefixes. We disregard the Subject because it does not make strong distinctions. For example, “Re: RxNav” (in some combination of upper and lowercase) occurs 663 times in the non-encoded portions of the RXNAVINFO archives.

Besides citing another message with In-Reply-To, replies may restate, or “quote”, some or all of the text of the original message. No standard governs quotation. The original may be above or below the reply. The original may be inset by “>” marks or not. Furthermore, the writer of the reply may have edited the quoted text, e.g., to keep only relevant parts or to interleave answers amidst the questions!

A conversation may be a linear chain of replies or it may be a haze of multiple replies to the same message, or replies that respond to multiple messages in one message.

#### Problems with Conversations

A logical reply might be written without pressing the “Reply” button.

A new conversation might be started by choosing to “Reply” to an old message.

### A repeat customer’s opinion on whether a memo will seem to us like a reply or a new conversation might not suit the purposes of the present analysis.

### Reduction of noise

#### One unified history

We process messages in reverse-chronological order, without any regard for the artificial boundary between Listserv months.

#### Non-authoritative messages

We take an answer-centric approach. We use only messages “From:” an NIH address. (It is not possible to identify NLM from an email address as we may use “@nih.gov” addresses.) We assume that such “answer” messages are meaningful on their own, either by using complete sentences or by including (quoting) another message. Also, using only messages from NLM eliminates spam.

We eliminate messages from certain NLM addresses that do not answer RxNav questions: “MOR Service Account” and “NLM UMLS Custserv”.

We would like to eliminate messages that were not “To:” a non-NLM, non-listserv address. These messages are internal chatter among RXNAVINFO attendants. Unfortunately, many Listserv archive emails do not have any “To:” header.

#### Attachments

We discard any MIME part that has a Content-Disposition other than “inline”.

#### Duplicate renditions

Some messages include multiple renditions of the same content. We take “text/plain” in preference over “text/html”.

Many messages include signature (or other) graphics. We omit all MIME parts that are not text.

#### Redundant original text

Original messages, “embedded” in replies, are a mirage formed by repeating original lines of text with a “>” prefix. There is nothing reliable about these quoted messages.

We make a crude effort to keep authoritative responses and remove originals that are duplicated by authoritative responses. Specifically, we remove any message-body line whose text (excluding leading “>” marks) occurs with a prefix of at least one “>” in any chronologically later message. This has the salubrious side effect of weeding out “signature” lines and other often-repeated boilerplate.

## [Appendix 2: Final Analysis Spreadsheet](https://www.nlm.nih.gov/about/training/associate/associate_projects/Cruise_Appendix2.xlsx)

## Appendix 3: Suggestions Document

• Building Queries – common workflow examples, like how to find ATC from NDCs, highlighted. Most used API functions (and most recommended [in the answers tags] API functions) highlighted. Common misunderstandings explained (for example, result file needs to be ‘unzipped’)

• Licensing/Restrictions– using data in commercial applications, citing, proprietary sources, API pricing and API keys, API call limit

• Contact/’Getting Help’ – listserv, FDA, DailyMed, Drugbank, UMLS etc.

• Common Data Elements/TTYs – severity, brand names, US market only, disease info, allergies, start/end date meaning, ‘status’ of NDCs and RxCUIs.

• Data Freshness - Update schedule, automating updates, and ‘downloading’ data ‘tables,’ (for example, how do I download the ‘whole rxclass database’?) source code, schemas, uptime/downtime

• Other Common Topics: Approximate match, VA classes map to products, not ingredients! Exceeding limit on the API calls? First check the data for duplicates! Why is this product not present? (out of scope, error in NDC input, retired etc.) How do I get data added to the API?

• Other Suggestions: Potentially add a note/banner encouraging users to share their use case, especially when asking for help. Make the link to the FAQ really prominent – I would recommend the FAQ being accessible in the footer, or having its own button in the header, and being clickable from any of the pages in https://rxnav.nlm.nih.gov/index.html.

1. https://tools.ietf.org/html/rfc2822 [↑](#footnote-ref-2)
2. $ grep -hE '^Content-Type:' LOG\* | tr 'a-z' 'A-Z' | perl -ne 'chomp; s/\s\*;\s\*/;/g;print qq($\_\n)' | perl -ne 'chomp; s/\s\*=\s\*/=/g;print qq($\_\n)' | perl -ne 's/"//g;print' | grep -v 'NAME=' | perl -ne 'chomp;s/BOUNDARY=[^;]+//g;print qq($\_\n)'| perl -ne 's/;;/;/g;print' | perl -ne 's/;$//g;print' | sort | uniq -c | sort -nr | head -n 12 [↑](#footnote-ref-3)