NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Journal Article Tag Suite Conference (JATS-Con) Proceedings 2017 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2017.

PubMed: Redesigning citation data management

Authors

Kathleen Gollner and Kathi Canese.

Affiliations

National Center for Biotechnology Information (NCBI)

Over the last couple years, we have drastically changed the systems and process used to manage PubMed citation data. It began with revising long-standing NLM policies and reducing reliance on manual citation corrections, then culminated with the release of the PubMed Data Management (PMDM) system in October 2016. With PMDM, we introduced a single system for managing citation data with a UI for editing citation data. In this brave new world, the responsibility for correcting citation data shifted from NLM Data Review to PubMed data providers. Any errors reported in PubMed citations are now forwarded to the publisher — a strategy that publishers have enthusiastically upheld. Here, we outline how the systems and process for managing PubMed citation data have changed, and detail the outcome of these changes since PMDM was launched.

Introduction

PubMed, a database maintained by the National Library of Medicine (NLM), includes more than 27 million citations for biomedical and life science research literature. Thousands of citations are added to PubMed each day. But, as PubMed continued to grow, it was increasingly evident that a new approach to managing PubMed citation data was required.

Over the last two decades, the scope and scale of PubMed has changed. First, there were MEDLINE citations — a collection rooted in NLM's first retrieval system, MEDLARS, initiated in 1964. MEDLINE citations are, largely, obtained from journals selected by the Literature Selection Technical Review Committee (LSTRC). Citations from these journals are indexed with the Medical Subject Headings (MeSH), a controlled vocabulary developed and maintained by NLM.1 When PubMed was initially released in 1996, it was designed to provide web-based access to MEDLINE.2 But, in the years since, PubMed has expanded to include additional sources of citations.

Today, PubMed contains citations from MEDLINE, OLDMEDLINE, PubMed Central (PMC), and the NCBI Bookshelf. A significant number of citations created by NLM prior to the MEDLINE collection have been added to PubMed; these are collectively referred to as OLDMEDLINE citations.3 In addition, most full-text articles in PubMed Central (PMC) have a corresponding citation in PubMed. These include citations for articles published in PMC participating journals, as well as many of the articles supported by funding agencies that require deposit in PMC for compliance with public access policies.4 There is also a PubMed citation for most books, book chapters, or documents added to the NCBI Bookshelf.5 Citations deemed out-of-scope for MEDLINE may be included in PubMed as well. These citations are likely one of two varieties: for journal issues published before the first MEDLINE-indexed issue, or for articles published in MEDLINE-indexed journals on topics outside biomedical and life science research. Lastly, PubMed includes citations for ahead-of-print articles published in either MEDLINE-indexed or PMC-participating journals.6

The policies governing what citations are included in PubMed are many and varied. But the process for managing the citation data was designed for the original citation collection, MEDLINE. While PubMed has included citations from non-MEDLINE sources for some time, the process for managing the data remained largely unchanged.

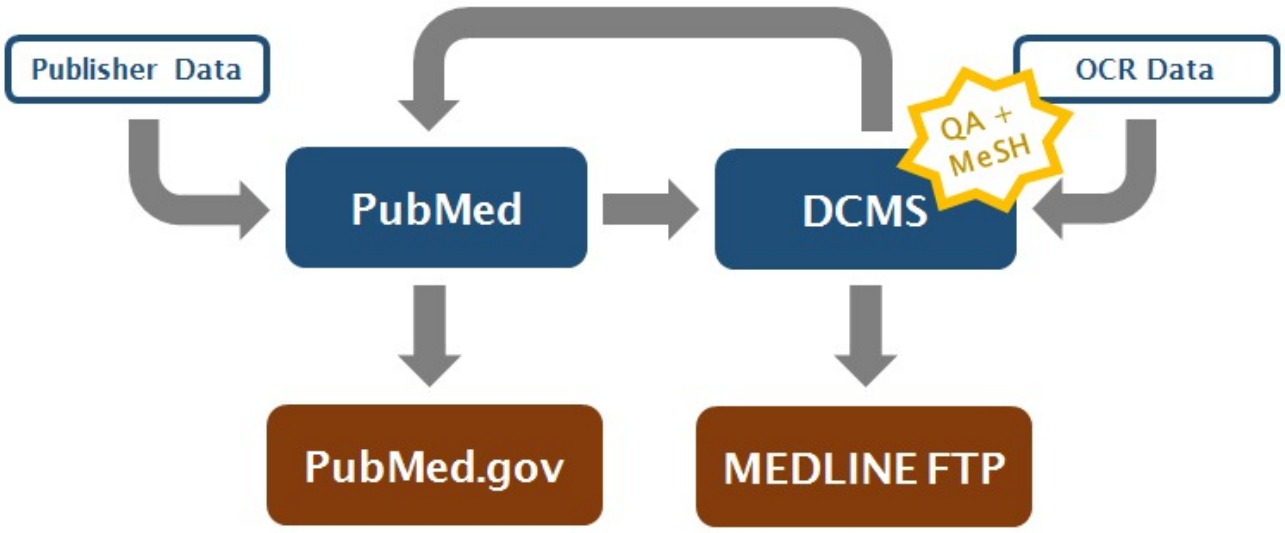
Additionally, the MEDLINE collection is growing as well.7 As the number of MEDLINE citations continued to rise, it placed increasing demands on MEDLINE processing. Each MEDLINE-destined citation was reviewed by NLM

staff to ensure the data was complete and correct. And the same NLM staff were responsible for handling any errors reported in citations, for both MEDLINE and non-MEDLINE citations. It became increasingly difficult to keep pace with citation data corrections, as a result.

Over the last two years, the NCBI PubMed team and the NLM Indexing Section team have redesigned the process of managing citation data. The cornerstone of the redesign, the PubMed Data Management (PMDM) system, was recently released, alongside a number of significant NLM policy and process changes. The story of this redesign is outlined here.

A Tale of Two Systems

Previously, NLM relied on two systems for managing PubMed citation data: the PubMed system and the Data Creation and Maintenance System (DCMS). The PubMed system, managed by the National Center for Biotechnology Information (NCBI), stored and indexed citation data for PubMed search. The DCMS, managed by the Library Operations (LO) and the Office of Computer and Communications Systems (OCCS) at NLM, facilitated data review and topical indexing of citations for MEDLINE. Here, we provide an overview of the systems and process used to manage PubMed citation data before the redesign.



Managing PubMed Citation Data, Previously

Receiving XML submissions

The process began, and still begins, with data provider submissions to PubMed. This data provider may be the publisher or a third-party organization that submits the citation data to PubMed on the publisher's behalf. Each data provider applies to submit citation data to PubMed in XML. They must demonstrate their ability to prepare complete and correct citation data in keeping with the PubMed Publisher DTD. Once approved by NCBI PubMed staff, an account is created for the data provider, and each eligible journal is assigned to their account. Going forward, the data provider can submit citation data for their designated journals; submissions for any unassigned journals will be rejected.

Citation submissions were loaded to the PubMed database twice a day, and then indexed for PubMed search each night. All new citations were typically available in PubMed search the day after submission.

The recently loaded citations were shipped to the DCMS once a day. Ahead-of-print citations were an exception, though; they were not included in the export until they were updated to reflect the final, published article. The export to the DCMS initiated the MEDLINE indexing process for citations in MEDLINE-indexed journals. But all citations,

regardless of whether they were destined for MEDLINE or not, were eventually shipped to the DCMS and, subsequently, bound to the MEDLINE indexing process.

Indexing for MEDLINE

Once the citations arrived in the DCMS, they were queued for data review by NLM Data Review staff. Most citation data fields in each citation were manually reviewed to ensure these citation data were a complete and accurate reflection of the published record. Any errors in article titles or author names, for instance, were amended. In addition, NLM staff added data — like grants, databanks, and links between associated citations, e.g. errata and retractions. NLM policy did not allow data providers to include these data in their citation submissions to PubMed; they relied on NLM Data Review staff to add these data. Finally, NLM Indexing Section staff assigned the relevant Medical Subject Headings (MeSH).

For MEDLINE-indexed journals that did not have an eligible XML data provider, an NLM contractor created the citations by OCR scanning the print issue of the journal. The citation data was loaded to the DCMS, reviewed for completeness and accuracy, and then indexed with MeSH. Only a small minority, 6%, of MEDLINE-indexed journals rely on the OCR process. 94% of MEDLINE-indexed journals have an approved XML data provider.

Exporting updated citations

Each day, all citations added or modified in the DCMS were exported to PubMed and posted to the MEDLINE FTP. The new and updated data would reflect in PubMed search the following day, after PubMed indexing completed. In addition, the new and updated data would reflect in the MEDLINE FTP, where licensees, organizations who had signed a license with NLM to use the MEDLINE data for their own products and services, could collect the latest citation data.

In summary, the process for managing PubMed citation data was divided between two separate systems. The PubMed system received initial submissions from data providers, then stored and indexed citations for PubMed search. The DCMS handled data review and topical indexing for MEDLINE. But all citations — not just MEDLINE-destined citations — were eventually shipped to the DCMS. This created significant challenges in the ongoing maintenance of PubMed citation data.

Considering the Pain Points

This tale of two systems reflects the many changes in scope and scale of citation data handled at NLM. While PubMed was initially created to provide web-based access to MEDLINE citations, its role has since evolved, expanding with a growing MEDLINE collection and extending to additional sources of research literature — PubMed Central and NIH-funded manuscripts, for instance. But the systems and process relied on for managing the data had not kept pace. This introduced a number of challenges in the day-to-day management of PubMed citation data. Three of these challenges are detailed here.

Data asynchronicity, loss, and corruption

Every day, large volumes of citation data were relayed back and forth between PubMed and the DCMS. With one system loading new data daily and the other system modifying data daily, the two systems were never truly synchronized. The daily transfers were part of a constant and arduous effort to keep the citation data in either system up to date. But with each transfer, there was the risk of losing or corrupting data. NLM staff monitored the citation data transfers for potential problems, but, nevertheless, some updates would be overwritten, omitted, or otherwise corrupted.

Onerous correction procedures

The most public of the problems: how an error in a PubMed citation was addressed. As long as a citation was ahead-of-print, it could be corrected by the data provider. But once a citation reflected the final, published article, it was

ultimately sent to the DCMS. In the DCMS, only select NLM staff could edit the citations and only in keeping with the MEDLINE workflow. If a user reported an error in a citation in the queue to be reviewed for MEDLINE, then they would be turned away by NLM Customer Service. They would be told that the citation should be complete and correct by the time it was indexed for MEDLINE. Unfortunately, due to delays in MEDLINE indexing, it could take several months or, in some cases, much longer before the citation was corrected.

In addition, not all citations are MEDLINE-destined data. If a user reported an error in a citation that was not in the queue for MEDLINE, then the correction request would be added to the list for NLM Data Review staff to address in turn. With only a small team of Data Reviewers able to correct the citations in the DCMS, it could take a fair amount of time for these errors to be amended as well.

Inconsistent public data sets

Two different citation data sets were made available to the public. When obtaining citation data from PubMed through E-Utilities, the data for any PubMed citation was available. These data were delivered in the PubMed DTD. When obtaining citation data from the MEDLINE Licensees FTP, the data for a subset of PubMed citations were available. These data were delivered in the MEDLINE DTD.

The PubMed DTD and the MEDLINE DTD were not quite the same. The MEDLINE DTD described the data available in the DCMS, while the PubMed DTD described the data available in PubMed. The data available in PubMed included both the MEDLINE data fields, as well as a few others, e.g. a list of publication dates and a list of article identifiers. Those who relied on data from both sources — organizations who licensed MEDLINE data and collected ahead-of-print citation data from E-Utilities, for instance — had to resolve these inconsistent outputs.

Planning for Redesign

In light of these challenges, the PubMed team at NCBI and the Indexing Section team at NLM Library Operations (LO) initiated an effort to drastically modify the systems and process relied on for managing PubMed citation data. The goal was to design a more efficient and effective workflow, one that would streamline the process of adding and modifying citations. More specifically, the following goals were pursued:

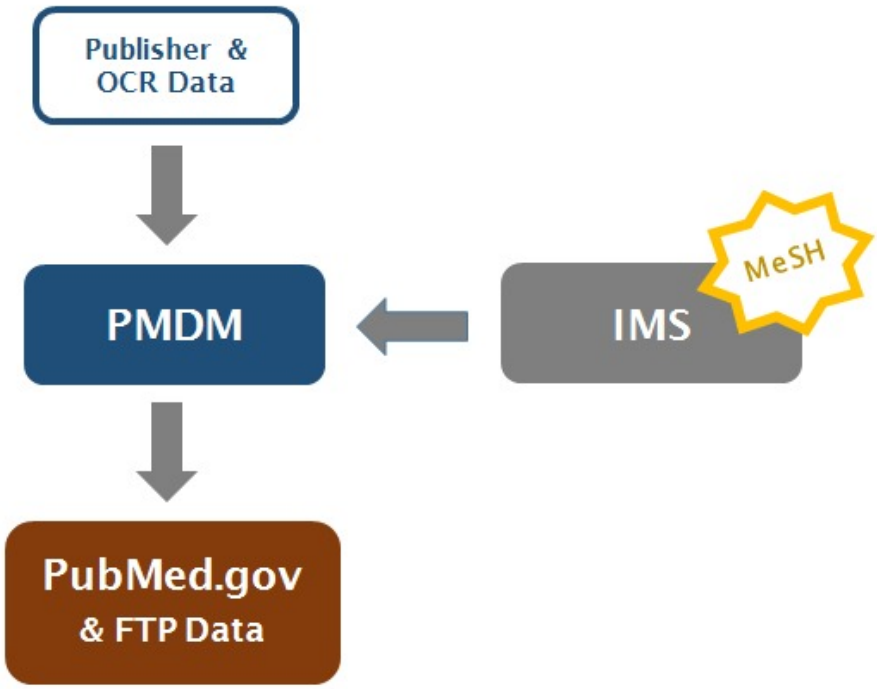
1. Developing a single system to store and manage citation data, with a separate system for managing Medical Subject Heading (MeSH) assignments.
2. Minimizing the reliance on manual data corrections by NLM Data Review staff by:
    1. Revisiting NLM data correction policies.
    2. Automating data corrections.
    3. Inviting data providers to submit additional data.
3. Allowing data providers and publishers to edit their citation data.

Essentially, the process for managing citation data would be divided from the process for indexing with MeSH. This new arrangement would allow the two processes to occur independently. In addition, this new arrangement would include new measures to facilitate more efficient corrections. One by one these goals were realized, and a brave new paradigm was established.

A Brave New World

During the last two years, the systems and process for managing PubMed citation data were drastically redesigned. A significant number of policy and procedure changes preceded the release of PMDM. This section describes the specific changes enacted during the redesign, as well as the observed impact of the changes since launch.

Managing PubMed Citation Data, Redesigned



Designing a new system

The PubMed Data Management (PMDM) system was designed to receive, manage, and export citation data. The initial process is unchanged; data providers submit citation data in XML as usual. But, now, the OCR data is submitted in this stream as well. When NLM staff OCR scan print issues of MEDLINE-indexed journals, the citation data are prepared in the PubMed Publisher DTD, and submitted to a designated FTP account. These citation data are loaded to PMDM alongside citation data from publishers, third-party providers, and PMC.

Citation data submissions are loaded to PMDM five times a day. As soon as the citation data has been loaded, the citations are available for editing using the PMDM UI. Once a day, all new or modified citations are indexed for PubMed search and exported to the FTP site.

The FTP site is no longer restricted to licensees and no longer limited to the MEDLINE data. Ahead-of-print citations, never before available through the FTP, are now offered. And all citation data is delivered in the PubMed DTD. Regardless of whether the data are obtained from the FTP or E-Utilities, the same data are available in the same DTD.

With this new arrangement, the workflow for MEDLINE indexing was revised. Data review is conducted using the PMDM UI. Indexing is completed independently of the data review process in the Indexing Management System (IMS). Concurrent with the design of PMDM, a new indexing system, the IMS, was built. The IMS provides a platform for MEDLINE indexing with the Medical Subject Headings (MeSH) vocabulary. Unlike the DCMS, the IMS does not include any functionality for maintaining other aspects of the citation data. When Medical Subject Headings (MeSH) are assigned to a citation in the IMS, the MeSH assignments are exported to PMDM and updated in PubMed search.

Minimizing manual data corrections

While the new system was in development, there were ongoing efforts to reduce reliance on manual data corrections. During the early stages of the redesign, the NCBI PubMed team developed the “DCMS Tracker”, a tool for comparing the citation exported from PubMed to the citation received back from the DCMS. The Tracker highlighted what had been changed by Data Review, allowing for reflection on how citation data were being revised.

Insights from the DCMS Tracker inspired several significant changes. First, it contributed to re-evaluations of NLM data correction policies. Some routine corrections made by NLM Data Review staff reflected long-standing protocols, like standardizing the title case for all citations. NLM Library Operations eventually decided to retain the title case provided in the data provider's submission to PubMed, as that directly reflected the publisher's data. Other data correction policies were subject to similar reviews.

Second, the DCMS Tracker helped identify corrections that could be automated. These corrections are now made when the XML citation data are initially loaded to PubMed; punctuation errors, for instance, are corrected automatically. Duplicate colons are removed. Missing periods are added. Adding these changes to the PubMed loader ensured that the corrections reflected as soon as the citation was available in PubMed, eliminating the need for a manual correction by NLM Data Review at a later time.

When the DCMS Tracker was first implemented, 87% of citations were modified by NLM Data Review during the MEDLINE indexing process. Within eight months, during which the policy changes and automated corrections were implemented, the number of citations modified was dramatically reduced. Only 18% of citations were corrected by NLM Data Review.

In an effort to further reduce manual citation data review, NLM revised their policies on the data fields data providers were eligible to submit. Previously, data providers were not permitted to submit most publication types, grant information, databank and Clinical Trial numbers, or the links between associated citations, e.g. between a retraction and the retracted article. But, one by one, these NLM policies were changed, and the PubMed Publisher DTD was modified to accommodate submissions. Now these data types can be included in citation data submissions to PubMed. If the data providers include these data in their submissions, the data will be included in the citation as soon as it is live in PubMed, saving NLM Data Review from manually reviewing and adding the data to these citations.

Allowing publishers to edit data

Finally, the most significant NLM policy change of all. With the introduction of the PMDM system, the data providers and publishers would be able to use the PMDM UI to edit their citation data. In fact, the responsibility for correcting citation data would shift from NLM Data Review staff to PubMed data providers.

To accommodate this new policy, the PMDM UI was designed to allow data providers to edit citations in their designated journals. Similar to the permissions arrangement for submitting citation data in XML, each data provider would only have access to citations in journals assigned to their account by the NCBI PubMed team. But since some data providers are third-party providers, working on behalf of client publishers, the PMDM UI allows data providers to create subaccounts for additional users with access to all journals or a specific subset of journals in the data provider's set. A third-party provider can grant access to representatives for each of their client publishers, and those publishers will only have access to the journals specified by the provider.

While publishers are expected to correct their citation data in PubMed, they are also expected to follow best practices, as established by the International Committee of Medical Journal Editors (ICMJE) and other publishing organizations[8], and to uphold current NLM policy.

In the months since PMDM was launched, both NLM Data Review and PubMed data providers and publishers have been using the PMDM UI to edit citation data — and with remarkable success.

Assessing the Initial Impact

Over the last five months, the new PMDM system has been enthusiastically adopted by both core user groups, PubMed data providers and NLM Data Review. The data providers and publishers, in particular, have been using PMDM to correct errors in citations. Here, we describe the initial impact of the new approach to managing PubMed citation data.

Data Providers and Publishers

When we announced the official release of PMDM, many PubMed data providers were quick to show their support. As told by a couple of our data providers:

> I have now managed to do many long-forgotten amendments in our transmitted sendings to you, it feels like heaven! … It was high time for us publishers to be able to manage these amendments in PubMed directly ourselves without asking anyone – júhú!

> I'm glad to hear that NCBI launched PubMed Data Management (PMDM) system, which will improve our citations' quality and smooth the revising process.

But, while this support was encouraging, we were hoping to see the enthusiasm translate to editing citations in PMDM. We did. As of mid-March 2017, over 12,100 citations have been edited by 175 external PubMed data providers, i.e. excluding internal data providers, PMC and OCR. Of the journals assigned to external data providers, 93% have a data provider who is actively using PMDM to correct citations. An "active" data provider has made at least one edit to a citation. This is a conservative measure, as it requires that the data provider has a citation that requires a correction. 79% of the data providers who are deemed inactive (7%) are responsible for only one journal. With less content published, they are also less likely to have a citation that requires correction. But, to ensure data providers use the system when required, we continue to reach out to alert them of or assist them with corrections, as needed.

And what have data providers and publishers been editing? Author lists, mostly. This is not surprising as author errors are the most frequently reported error in PubMed citations. Often, the errors are caused by common markup errors in XML submissions, like missing authors or inverted names, e.g. <LastName>Jane</LastName> and <FirstName>Smith</FirstName>. In the case of author affiliation fields, errors have been accumulating for some time. After a NLM policy change in 2013, data providers were permitted to submit affiliation data for all authors and investigators. NLM Data Review subsequently stopped reviewing the data in the affiliation fields; they relied entirely on the data provided by the data providers. 9 After the release of PMDM, several data providers have made a concerted effort to clean-up several years' worth of errors in the affiliation fields of the citations. While the author lists have been the most frequently edited, data providers and publishers have been correcting every editable citation data field — from fixing typos in article titles to adding Clinical Trial registration numbers.

In addition to correcting citations in PMDM, data providers have been allowed to include additional data fields in their submissions, specifically more publication types, grant information, databank and Clinical Trial numbers, and links between associated citations. Some data providers have begun to include this data in their submissions. Since January 2017, 35 providers supplied new publication types, 3 for grant information, 6 for databank numbers, and 23 for associated citation links. While these numbers are small now, requiring NLM Data Review to continue to add this information, we expect that more data providers will be able to supply this citation data in the future. To begin including this data in their submissions, data providers often need to make significant changes to their systems and process. This may take some time.

NLM Data Review

Meanwhile, NLM Data Review staff continue to conduct data review for MEDLINE-destined citations, but this work is now conducted using the PMDM UI and their workflow has been streamlined. They no longer review most citation data fields in every MEDLINE citation. Instead, they focus on specific data fields: publication types, grant information, databank and Clinical Trial numbers, and links between associated citations. If these data fields are not supplied by the data providers in their submissions to PubMed, then these data will added to the citations by NLM Data Review during MEDLINE processing.

With the release of PMDM, the responsibility for correcting errors reported in citations shifted from NLM Data Review to PubMed data providers. This change has helped reduce the Data Review workload. Previously, citation data errors reported to NLM Customer Service were forwarded to NLM Data Review staff — unless the citation was

in process for MEDLINE, in which case the user was told to wait for MEDLINE indexing to be completed. But since the release of PMDM, citation data errors are reported to the publisher, and the publisher is responsible for handling the correction. Between September and December 2016, NLM Data Review saw a 72% reduction in the number of NLM Customer Service requests forwarded to them.

Looking Forward

The redesigned system and process for managing PubMed citation data is barely six months old. But initial assessments of the impacts of the changes are very encouraging. Most PubMed data providers are actively using PMDM to correct errors in their citation data. NLM Data Review, relieved of a comprehensive review process, can now focus on adding valuable data to citations. Going forward, we will continue to encourage PubMed data providers to incorporate the additional data fields in their initial submissions to PubMed. We will also continue to work with data providers and publishers to ensure errors reported in citations are addressed. But, if these first few months are an indication, the outlook for this new collaboration is promising.

Significantly, the redesign has centralized all PubMed citation data. This simplifies the process required to manage the data, and more readily facilitates additional improvements. When the citation data was dispersed across multiple systems, it was difficult — at times, prohibitively difficult — to pursue certain system and process improvements. But, now, with all the citation data in one place, we can more readily pursue changes that will allow for even more efficient and effective management of PubMed citation data.

Conclusion

As the number of citations in PubMed continued to rise, new and improved approaches to managing these data were required. The original arrangement could not accommodate the growing number of MEDLINE citations and the additional sources of citations — PubMed Central and NIH-funded manuscripts, for instance. A drastic redesign was initiated in response.

Over the last couple years, we have transitioned to a new system and process for managing PubMed citation data. A single, centralized system for citation data was introduced. Revised policies and automated procedures reduced reliance on manual data corrections. Finally, the PubMed data providers and publishers were tasked with addressing errors in their citation data. This new arrangement has simplified the management of PubMed citation data significantly. In addition, it facilitates more efficient corrections of data errors. As PubMed moves forward in its twenty-first year, it is better positioned to provide timely and accurate citation data.

Acknowledgments

There were many people who contributed to the effort described in this paper. A big thank you to everyone involved. And a special thank you to the core contributors: Vladimir Korobtchenko, Sarah Weis, Grisha Starchenko, Michael Kholodov, Jared Hellman, Sharmin Hussain, Marie Collins, Deborah Ozga, J Shore, Susan Schmidt, and John Rozier.

References

1. "MEDLINE Fact Sheet." National Library of Medicine. Accessed 19 March 2017. https://www.nlm.nih.gov/pubs/factsheets/medline.html.
2. Canese K . . "PubMed® Celebrates its 10th Anniversary!" NLM Technical Bulletin, 352 (September - October 2006.): e5. Accessed 19 March 2017. https://www.nlm.nih.gov/pubs/techbull/so03/so03_oldmedline.html .
3. Demsey A , , Nahin AM , , Von Braunsberg S . . "OLDMEDLINE Citations Join PubMed®." NLM Technical Bulletin, 334 (September - October 2003.): e2. Accessed 19 March 2017. https://www.nlm.nih.gov/pubs/techbull/so03/so03_oldmedline.html .
4. "PMC Overview." National Center for Biotechnology Information. Accessed 19 March 2017. https://www.ncbi.nlm.nih.gov/pmc/about/intro/
5. Canese K . . "Book Citations Added to PubMed® and Changes to Displays." NLM Technical Bulletin, 373 (

March - April 2010.): e13. Accessed 19 March 2017. https://www.nlm.nih.gov/pubs/techbull/ma10/ma10_pm_books.html .

6. "PubMed: MEDLINE® Retrieval on the World Wide Web." National Library of Medicine. Accessed 19 March 2017. https://www.nlm.nih.gov/pubs/factsheets/pubmed.html.
7. "Citations Added to MEDLINE® by Fiscal Year." National Library of Medicine. Accessed 21 March 2017. https://www.nlm.nih.gov/bsd/stats/cit_added.html.
8. "Corrections, Retractions, Republications and Version Control." ICMJE: International Committee of Medical Journal Editors. Accessed 29 March 2017. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/corrections-and-version-control.html.
9. "Changes Coming to Author Affiliations." NLM Technical Bulletin, 394 (September - October 2013.): b4. Accessed 21 March 2017. https://www.nlm.nih.gov/pubs/techbull/so13/brief/so13_author_affiliations.html .

Bookshelf ID: NBK425541