atoms

MDPI

Article

Referencing Sources of Molecular Spectroscopic Data in the Era of Data Science: Application to the HITRAN and AMBDAS Databases

Frances M. Skinner [1,2,*], Iouli E. Gordon [1,*], Christian Hill [3,*], Robert J. Hargreaves [1], Kelly E. Lockhart [4] and Laurence S. Rothman [1]

[1] Atomic and Molecular Physics, Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA 02138, USA; robert.hargreaves@cfa.harvard.edu (R.J.H.); lrothman@cfa.harvard.edu (L.S.R.)

[2] Undergraduate Chemistry Department, University of Massachusetts Lowell, Lowell, MA 01854, USA

[3] Nuclear Data Section, International Atomic Energy Agency, Vienna International Centre, PO Box 100, A-1400 Vienna, Austria

[4] The SAO/NASA Astrophysics Data System (ADS), Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA 02138, USA; kelly.lockhart@cfa.harvard.edu

* Correspondence: frances.skinner@cfa.harvard.edu (F.M.S.); igordon@cfa.harvard.edu (I.E.G.); Ch.Hill@iaea.org (C.H.)

Received: 27 February 2020; Accepted: 25 April 2020; Published: 30 April 2020

Abstract: The application described has been designed to create bibliographic entries in large databases with diverse sources automatically, which reduces both the frequency of mistakes and the workload for the administrators. This new system uniquely identifies each reference from its digital object identifier (DOI) and retrieves the corresponding bibliographic information from any of several online services, including the SAO/NASA Astrophysics Data Systems (ADS) and CrossRef APIs. Once parsed into a relational database, the software is able to produce bibliographies in any of several formats, including HTML and BibTeX, for use on websites or printed articles. The application is provided free-of-charge for general use by any scientific database. The power of this application is demonstrated when used to populate reference data for the HITRAN and AMBDAS databases as test cases. HITRAN contains data that is provided by researchers and collaborators throughout the spectroscopic community. These contributors are accredited for their contributions through the bibliography produced alongside the data returned by an online search in HITRAN. Prior to the work presented here, HITRAN and AMBDAS created these bibliographies manually, which is a tedious, time-consuming and error-prone process. The complete code for the new referencing system can be found on the HITRAN*online* GitHub website.

Keywords: HITRAN; database; bibliography; citation; reference

1. Introduction

Knowledge of spectroscopic parameters for transitions between energy levels in atoms and molecules is essential for interpreting and modeling the interaction of radiation (light) with different media. In order to aid researchers, spectroscopic parameters are being compiled into reference databases. In particular, the HITRAN (high-resolution transmission) molecular spectroscopic database [1] is a compilation of spectroscopic parameters used to simulate and analyze the transmission and emission of light in gaseous media, with an emphasis on planetary atmospheres. For half a century HITRAN has been considered to be an international standard which provides one recommended value per parameter for millions of transitions for different molecules. HITRAN employs both experimental and theoretical data which are gathered and processed from articles, books, proceedings, databases, theses, reports, presentations, unpublished data, papers in-preparation and private communications. Commencing with the HITRAN1986 edition [2], HITRAN started to provide reference mapping for the line positions, transition intensities, and broadening coefficients due to the pressure of air. Starting with the HITRAN2004 edition [3], the majority of parameters in HITRAN have complete reference attributions. The current edition of HITRAN [1] contains references for dozens of parameters per transition. It is imperative that all of these contributions to HITRAN receive acknowledgement through proper referencing to their cited material in the HITRAN database. This gives users an option to read more about how the parameters were determined and also acknowledges contributing papers and enables their citation by the users of the database. Furthermore, it assists the managers of the database in maintaining and updating complex segments of the database.

While HITRAN is unique in providing references for the majority of the spectral parameters in every transition, it is important to note that other reference molecular spectroscopic databases exist, and they adopt different approaches to citing original references. One such database is GEISA [4] which has many of the same parameters as in HITRAN. GEISA does not provide references to the individual parameters and usually provides a reference index for the entire transition; however mapping for this index is available only to the managers of that database for maintenance purposes. It is also important to mention two databases (NASA's Jet Propulsion Laboratory (JPL) microwave catalogue [5] and the Cologne Database for Molecular Spectroscopy (CDMS) [6]) that concentrate mostly on spectroscopic parameters in the microwave region for molecules of astronomical interest. While individual parameters or transitions do not have references in these databases, a supplementary bibliography per line list is provided. This bibliographic information provides a list of references to the reported data that were used in the databases directly or as input into the global model used for calculation. A similar approach has been adapted by the databases of extensive ab initio line lists, including ExoMol [7] and TheoReTs [8].

In this paper we describe a new, automated referencing system to provide consistent, accurate and detailed bibliographies to every source of data on the website. Starting from the HITRAN2016 edition [1], the data is being distributed through HITRAN*online* (https://hitran.org) which is built on a relational database model described in Hill et al. [9]. The relational database approach removes the constraints of fixed-width text fields for the storage of parameters and allows an arbitrary number of parameters to be stored and retrieved for each transition, along with their uncertainties and bibliographic references. Each data set returned by a search is accompanied by a bibliography. Each data file provides citations, hyperlinks, and notes to the original data sources to make it easier for users to credit data providers. Bibliographies can be exported in several formats, including HTML, plain text and BibTeX.

All of the information on the contributions to HITRAN has previously been entered manually into the database; this method is error-prone and time-consuming. The purpose of this work is to create a convenient bibliographic system, to enable contributors' work to be easily cited. Users utilizing this system need only enter a single line of information into the program, in order to obtain the complete bibliography entry for the paper they wish to cite. This system was designed to prevent common mistakes and ensure faster updates to the references system in HITRAN as well as to the Atomic and Molecular Bibliographic Data System (AMBDAS) database. We have previously reviewed existing data practices in molecular spectroscopy [10] and have identified that there is room for improvement which can be facilitated with specialized tools. The goal of this work is to encourage an environment that promotes data sharing provenance and good practice amongst researchers and databases.

The AMBDAS database is a collection of references to articles concerning collisional and spectroscopic processes in plasmas and plasma–material interaction data, with a particular focus on their application to nuclear fusion energy research. The database website, accessible from https://amdis.iaea.org/databases/, provides an interface for querying by collision species, process category and publication metadata. The bibliographic entries are maintained using the Python software described in this article, through established collaborations with the National Institute for Fusion Science (NIFS) in Japan, the National Fusion Research Institute (NFRI) in South Korea and the National Institute of Standards and Technology (NIST) in the United States of America, as well as through ad hoc arrangements with individual consultants. The use of an intuitive, automated administration interface for importing data is an important way in which errors, ambiguities and duplications are minimized within the AMBDAS database.

The software tools presented here are described as applied to the HITRAN and AMBDAS databases, but are self-contained and can be applied to other database services. The complete code for the new referencing system is provided on an open-source and cost-free basis under version 2 of the Apache Software Foundation License [11], along with detailed instructions and resources for customizing the output formats of citations. The objective is that with enough availability and use, databases and researchers alike will be encouraged to share information between databases, in turn making their work more accessible to users. Enhancement of the HITRAN reference infrastructure will directly benefit the Virtual Atomic and Molecular Data Centre (VAMDC) [12] and their recommended citation practices [13,14]. VAMDC provides the infrastructure for dissemination of atomic and molecular data from different databases from the same platform. For instance, one can simultaneously access the data from the HITRAN, JPL and CDMS databases.

This paper will not describe in detail the HITRAN data itself, searching methods, how to access data, graphing data nor other technical accessibility information. For details on this information, please refer to the papers describing the quadrennial HITRAN editions (the most recent one is HITRAN2016 [1]) and the Hill et al. [9] article describing HITRAN*online*. The HITRAN*online* website has been available at https://hitran.org since May 2015. Registration, at https://hitran.org/register/ (and also linked from the home page), is free and requires the user to provide only a name and email address.

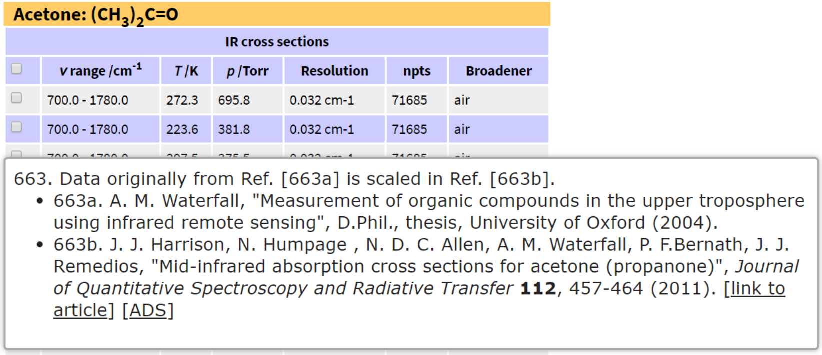
Current Status of References in HITRAN

Changes to the referencing system in HITRAN are useful only if users are aware of how to access this information when using HITRAN's data access capabilities. Therefore, this section is dedicated to providing a detailed review on how to retrieve bibliographies from HITRAN. There are several sections in HITRAN where the user may search for, graph, preview or download data. The corresponding bibliographies for these data are made available in several different ways which depend on the section the user is using at the time, as well as what specific information they are retrieving.

First of all, in each bibliography the user will see a number assigned to every reference. This number is a unique "global" identifying integer ID, which is referred to in relevant data files and is recorded for user accessibility and administrator storing purposes. Users of the legacy HITRAN 160-character .par format will find this "global" integer identification system convenient; the "per-molecule" identifying integers are retained and cross-referenced with their global equivalents when this default output format (.par) is selected. Alternatively, a custom output format may be created and used as described in the article by Hill et al. [9].

Some data in HITRAN are calculated from *multiple* references and sources; to provide full credit to all contributors HITRAN nests multiple references under a single bibliographic entry. The main bibliography where multiple references are nested, is technically a "note", while the nested references are complete bibliographies stored in the database. Therefore, any note can be created and assigned to data that is being referenced; the note will then pull the complete bibliographies of the papers that the data set is generated from. An example of this technique can be seen in Figure 1.

HITRAN has several major sections that provide different types of spectroscopic data, with the traditional section being the line-by-line or molecular transition section. In the line-by-line section, when a user accesses their desired data, they will see a query-results web-page that will contain a "downloads" table. In this "downloads" table, there are two bibliography files giving sources and notes relating to the returned data. One bibliography is an HTML file with links to the cited articles at their publisher websites and on the ADS database [15] (Figure 2). The other bibliography file is a .bib file containing these references in BibTeX format that enables the inclusion into a LaTeX document (Figure 3). If there are fewer than 1000 transitions returned by the query made by the user, then those transitions are listed in an HTML table on the same query results web page (Figure 4). Hovering the mouse cursor over each parameter in this table brings up a bibliography entry for that parameter which contains links to the article and any relevant notes on the reference.



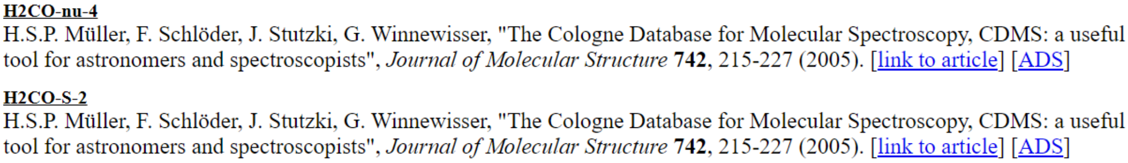
Acetone: $(CH_3)_2C=O$

IR cross sections						
□	ν range /cm^{-1}	T/K	p/Torr	Resolution	npts	Broadener
□	700.0 - 1780.0	272.3	695.8	0.032 cm-1	71685	air
□	700.0 - 1780.0	223.6	381.8	0.032 cm-1	71685	air

663. Data originally from Ref. [663a] is scaled in Ref. [663b].

- 663a. A. M. Waterfall, "Measurement of organic compounds in the upper troposphere using infrared remote sensing", D.Phil., thesis, University of Oxford (2004).
- 663b. J. J. Harrison, N. Humpage , N. D. C. Allen, A. M. Waterfall, P. F.Bernath, J. J. Remedios, "Mid-infrared absorption cross sections for acetone (propanone)", *Journal of Quantitative Spectroscopy and Radiative Transfer* **112**, 457-464 (2011). [link to article] [ADS]

Figure 1. The bibliography for a given temperature-pressure (T-p) absorption cross section of acetone. In this example, the cursor is over the second entry of the table. The number 663 to the left of the bibliography is the corresponding "global" ID integer, and the two links at the end will take the user to the full-text article on the publisher's website or ADS, respectively. There are two references displayed in this bibliography for one line of data; these two are labeled separately 663a and 663b.



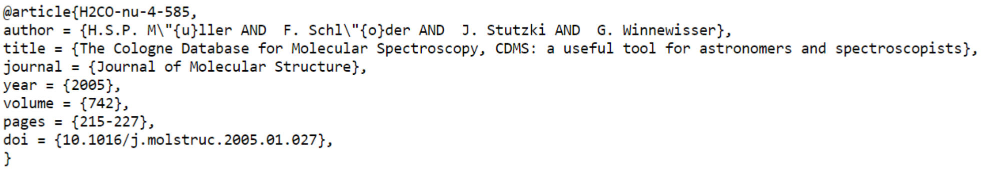
H2CO-nu-4

H.S.P. Müller, F. Schlöder, J. Stutzki, G. Winnewisser, "The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists", *Journal of Molecular Structure* **742**, 215-227 (2005). [link to article] [ADS]

H2CO-S-2

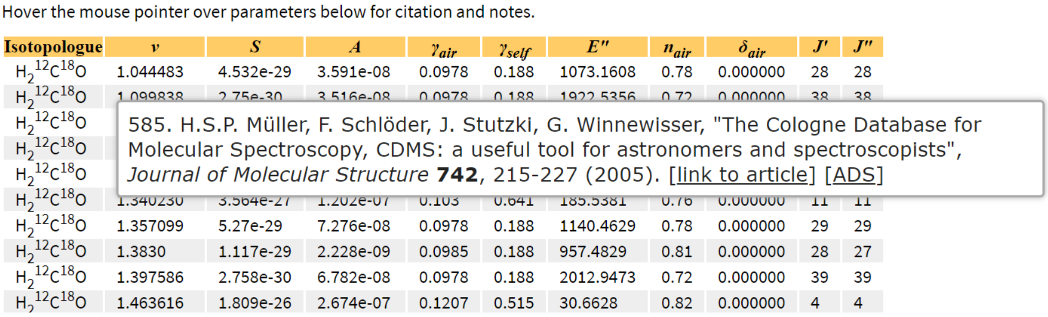
H.S.P. Müller, F. Schlöder, J. Stutzki, G. Winnewisser, "The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists", *Journal of Molecular Structure* **742**, 215-227 (2005). [link to article] [ADS]

Figure 2. An excerpt of the bibliography for the isotopologue $H_2^{12}C^{18}O$ of formaldehyde in the line-by-line section of HITRAN*online*. Every reference includes two hyperlinks at the end of each line that will take the user to the full-text article on the publisher's website or ADS, respectively.



```
@article{H2CO-nu-4-585,
author = {H.S.P. M\"{u}ller AND  F. Schl\"{o}der AND  J. Stutzki AND  G. Winnewisser},
title = {The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists},
journal = {Journal of Molecular Structure},
year = {2005},
volume = {742},
pages = {215-227},
doi = {10.1016/j.molstruc.2005.01.027},
}
```

Figure 3. An excerpt of the BibTeX formatted bibliography from HITRAN*online* for use with LaTeX. This bibliography is for the isotopologue $H_2^{12}C^{18}O$ of formaldehyde in the line-by-line section.



Hover the mouse pointer over parameters below for citation and notes.

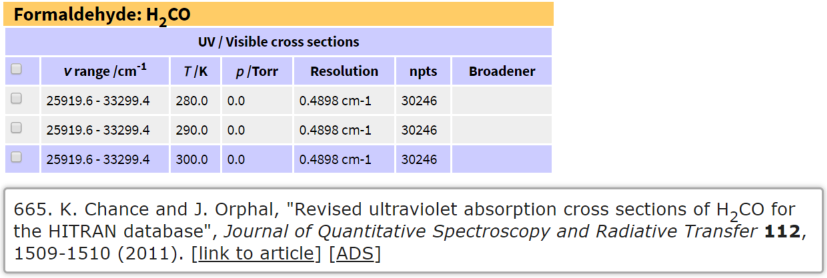
Isotopologue	ν	S	A	γ_{air}	γ_{self}	E''	n_{air}	δ_{air}	J'	J''
$H_2{}^{12}C^{18}O$	1.044483	4.532e-29	3.591e-08	0.0978	0.188	1073.1608	0.78	0.000000	28	28
$H_2{}^{12}C^{18}O$	1.099838	2.75e-30	3.516e-08	0.0978	0.188	1922.5356	0.72	0.000000	38	38
$H_2{}^{12}C^{18}O$										
$H_2{}^{12}C^{18}O$										
$H_2{}^{12}C^{18}O$										
$H_2{}^{12}C^{18}O$	1.340230	3.564e-27	1.202e-07	0.103	0.641	185.5381	0.76	0.000000	11	11
$H_2{}^{12}C^{18}O$	1.357099	5.27e-29	7.276e-08	0.0978	0.188	1140.4629	0.78	0.000000	29	29
$H_2{}^{12}C^{18}O$	1.3830	1.117e-29	2.228e-09	0.0985	0.188	957.4829	0.81	0.000000	28	27
$H_2{}^{12}C^{18}O$	1.397586	2.758e-30	6.782e-08	0.0978	0.188	2012.9473	0.72	0.000000	39	39
$H_2{}^{12}C^{18}O$	1.463616	1.809e-26	2.674e-07	0.1207	0.515	30.6628	0.82	0.000000	4	4

585. H.S.P. Müller, F. Schlöder, J. Stutzki, G. Winnewisser, "The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists", *Journal of Molecular Structure* **742**, 215-227 (2005). [link to article] [ADS]

Figure 4. A sample of ten transitions of the $H_2^{12}C^{18}O$ isotopologue from the line-by-line section of HITRAN*online*. This HTML table is displayed when there are fewer than 1 000 transitions returned by the query. A bibliography pop-up will appear for each parameter when the user hovers their cursor over the corresponding data. In this example, the cursor is over the wavenumber (ν) of the first row.

In the absorption cross section part of HITRAN*online* there is a complete list of references contained in the supplemental folder (provided at the top left of the window). In this folder the referenced sources can be found listed in HTML, Excel and plain text formats. If a user wishes to view the full bibliography for a particular absorption cross section, they click on their desired molecule and hover their mouse cursor over any of the rows of data listed for that particular molecule. An example is displayed for an absorption cross section of formaldehyde (Figure 5). After selecting one or more data sets, the user will be prompted to the final screen which then presents a page with links to the data files and a complete bibliography.

In the collision induced absorption section of HITRAN*online*, the corresponding references are given in the reference document PDF which is provided at the top of the web-page. In the aerosol properties section, the references for the utilized sources can be obtained through a PDF document. In the HITEMP section, users are asked to cite the original sources of data by using the assigned reference codes for each line transition. These reference codes are consistent with those in HITRAN line-by-line. In the supplemental section there is a list of references at the bottom of the window and in each subsection (Line-Mixing, Total Internal Partition Sums, Supplemental Absorption Cross Sections and Radioactive Isotopologues). There exists an instruction manual containing the references used for the data provided. Overall these bibliographies are displayed so that users will understand where the data came from and to be able to use this information for their records and further research. HITRAN also provides the HITRAN Application Programming Interface (HAPI) [16] that allows downloading the spectroscopic data from HITRAN and carry out sophisticated calculations using predefined or custom functions. At the moment HAPI does not enable downloading reference data but work is underway to provide this option in the near future.



Formaldehyde: H_2CO

UV / Visible cross sections

□	ν range /cm^{-1}	T /K	p /Torr	Resolution	npts	Broadener
□	25919.6 - 33299.4	280.0	0.0	0.4898 cm-1	30246	
□	25919.6 - 33299.4	290.0	0.0	0.4898 cm-1	30246	
□	25919.6 - 33299.4	300.0	0.0	0.4898 cm-1	30246	

665. K. Chance and J. Orphal, "Revised ultraviolet absorption cross sections of H_2CO for the HITRAN database", *Journal of Quantitative Spectroscopy and Radiative Transfer* **112**, 1509-1510 (2011). [link to article] [ADS]

Figure 5. The bibliography pop-up for a given temperature-pressure *(T-p)* absorption cross section of formaldehyde. In this example, the cursor is over the third entry of the table. The number 665 to the left of the bibliography is the corresponding "global" ID integer, and the two hyperlinks at the end will take the user to the full-text article on the publisher's website or ADS, respectively.

2. Results

2.1. Digital Object Identifier

A digital object identifier (DOI) is a string of numbers, letters and symbols used to permanently identify an article or document and link to it to its online original source; a DOI is assigned to almost every work that is published in the modern age [17]. Even proceedings and people can have their own DOI to electronically link public work to the sources. This DOI system is traceable and permanent, which is why HITRAN and AMBDAS chose to use the unique DOI for citing and referencing sources in their respective databases. The International DOI Foundation designates DOI's, which act as a unique identifier for content online and in print. Once assigned, DOI's are unchanging and therefore provide a permanent link to the location of individual works. Digital objects may change physical locations, but the DOI assigned to that object will never change. Therefore by using the DOI the object is always accessible to the user.

The DOI for an article can be found at the top or bottom corners of the published paper, or in the hyperlink to the paper. The new automated method implemented by the software described in this article enables the administrator to enter only the DOI for published work and in return retrieve the bibliography of the publication.

2.2. Automatic Referencing System

This section describes the process of querying and retrieving bibliographies and references with the new automated referencing system. All code for the automatic referencing system was written in the Python programming language through the web-based, interactive computing notebook environment, Jupyter Notebook and is also implemented in a Django application available at https://github.com/hitranonline/pyref. Interested users are encouraged to test the referencing system on this Jupyter notebook, which provides detailed instructions. Users who are interested in implementing the code through the provided Django application can find further details in the `setup.py`, `settings.py` and `README` files.

Several libraries are referenced, along with the necessary packages which must be installed and imported by the users. The referencing system was developed for the ease of use of the administrators of the HITRAN and AMBDAS database, as well as for the reliability of adding accurate information for the users and contributors of HITRAN alike.

The automatic referencing system provides three different output formats for every reference generated. Several formats are necessary so that users have multiple options when viewing and accessing the bibliographies in HITRAN. The format outputs for references are as follows:

- HTML
- JSON (Text)
- BibTeX

Every referenced material in HITRAN also has the option of adding a detailed note to the bibliography. This detailed note option is consistently used for describing where data was taken from in the article, what data was not used and any other information necessary for understanding and using the referenced information. This same note option is included in the new automatic referencing system in HITRAN. Hyperlinks are included for all references as well, so that the users will have access to the full-text of the paper as well as the ADS link to the paper. The ADS link is a hyperlink to the ADS database, which provides bibliographic information to a majority of astronomical researchers worldwide. ADS provides several unique systems; the user has the option to view author network visualizations, paper citations, paper downloads, access citations in multiple formats, and account users have the option to develop a private library for their records. Thus, HITRAN has endeavored to include links to the ADS database for articles as well as the DOI links to published work.

The new referencing system is detailed in six clear steps, and a visualization is provided in Figure 6. However, if the administrators are referencing a source that exists in the ADS database, then they only need to follow three easy steps in order to cite the source properly in all three output formats and include relevant notes. The six steps are explained in further detail in the Section 4 and they are listed as follows:

1. Retrieve the DOI of the paper that is being cited. Enter the DOI in the prompt provided.
2. This step will populate the corresponding ADS bibcode for the paper. If no bibcode is generated, skip to step 4
3. Use the retrieved bibcode from step 2 for the Section 4.2. The output is customizable; all output formats are possible.
4. Use the DOI to search in Section 4.3. The output will be the bibliography as a JSON output.
5. The citation from the Section 4.3 will be automatically formatted in HTML in the following Section 4.4.

6. The initial DOI is automatically populated in the final Section 4.5 to retrieve the BibTeX citation for the paper.

Compared with manual entry, there is much less scope for human error in this new referencing method. The only part on the user side is to include any necessary notes required for citing the source properly. Users of the HITRAN database will see contributors' names and links to their publications on every window where their data is displayed, and all of this information will be accurate and up to date. This system of displaying contributors' work has been the conventional method used by HITRAN since 2015. Prior to this referencing system, there existed a static file containing all of the references in HITRAN at the time and to what molecules or data these references were referred to.

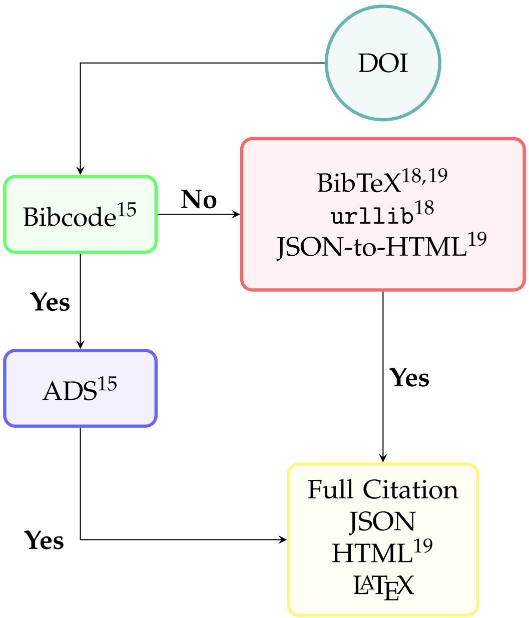


Figure 6. A diagram of the steps required to retrieve a reference from an article in multiple formats. The superscripts above the method names and outputs refer to the references for those codes.

Henceforth, HITRAN will continue the good-practices it has always used, for providing references of data shown in the database. This new automatic method for adding sources into HITRAN and AMBDAS will be the next stepping stone towards modernization of each database.

3. Discussion

HITRAN gets its data from contributors who publish their results, present their data at a conference or communicate privately with the HITRAN team. Adding these contributors' references in HITRAN is currently all done by hand. Therefore, this project created a new automated method for adding references in HITRAN, using only a DOI. This new automated system will be easier for the administrators/users to maintain, and data replacement will also be simplified as well as users' access to references. Most importantly, the contributors to HITRAN and AMBDAS will receive proper credit and representation. This change will ideally set an example to the research community and encourage a productive data sharing environment amongst researchers.

Prior to creating the new automated method, references in the HITRAN database were updated manually, either by transcribing reference metadata or by cutting-and-pasting from websites, PDF files and so on. Every reference needed to be processed painstakingly by hand, including: correlating the bibliography information to the main article or resource, adding the DOI information on the article, checking spelling, adding separate plain text, LaTeX and HTML markups, etc. About fourteen hundred references were corrected with up-to-date information, and their outputs were tested and double checked before updating the database. In addition, all changes are recorded for administrator record-keeping, and this further increases the time and attention required by the administrator. The new automated method of simply retrieving references with a single click of a button will ensure faster results, create a smaller time requirement and ensure minimal mistakes.

4. Methods

4.1. Bibcode Method

In every method, including this method, several modules will need to be imported or installed. The modules that are used include `requests` and `JSON`. The `json` module of the Python standard library [1] is necessary for obtaining JSON outputs from bibliography retrieval methods designed in the new automated referencing system. The `requests` module allows the user to send HTTP/1.1 requests, without the need for manually adding query strings to URLs, or to form-encode POST data[2].

In this method the administrator or user is required only to enter the DOI in the prompt provided. The user will then be shown the bibcode generated from the search through the ADS database. The bibcode shown will be populated throughout the entire Section 4.2 unless changed, allowing the user to only run their desired fields to retrieve custom formatted citations. The DOI entered in this initial prompt is also populated for the Section 4.3 and the Section 4.5 for ease of use.

A bibcode is what the NASA Astrophysics Data System uses to identify literature in their database. The bibcode is a 19 digit identifier and takes the format YYYYJJJJJVVVVMPPPPA where[3]:

- YYYY: Year of publication
- JJJJJ: A standard abbreviation for the journal (e.g., ApJ, AJ, MNRAS, Sci, PASP, etc.). A list of abbreviations is available.
- VVVV: The volume number (for a serial) or an abbreviation that specifies what type of publication it is (e.g., conf for conference proceedings, meet for Meeting proceedings, book for a book, coll for colloquium proceedings, proc for any other type of proceedings).
- M: Qualifier for publication:

  E: Electronic Abstract (usually a counter, not a page number)

  L: Letter

  P: Pink page

  Q-Z: Unduplicating character for identical codes
- PPPP: Page number. Note that for page numbers greater than 9999, the page number is continued in the M column.
- A: The first letter of the last name of the first author.

The fields that are empty in a bibcode are replaced with periods (.) so that the code is always 19 characters long. As an example, the bibcode "2017JQSRT.203....3G" corresponds to the Journal of Quantitative Spectroscopy and Radiative Transfer (JQSRT), year 2017, volume 203, first page 3 and the first letter of the last name of the first author is G.

If a Bibcode was not generated in the Bibcode method, then that would mean the paper the administrator or user is looking for is not in the ADS database. This will require the user to skip the following Section 4.2 and move on to the Sections 4.3–4.5. These three methods are fairly straight forward. The user will only need to run these coded cells for them to work properly since the

[1] JSON, **2019**, A Lightweight Data Interchange Format Inspired by JavaScript Object Literal Syntax. https://docs.python.org/3/library/json.html

[2] Reitz, K., **2019**, Apache Software License (Apache 2.0), https://pypi.org/project/requests/

[3] NASA/ADS help page at https://adsabs.github.io/help/actions/bibcode.

necessary information is already automatically populated for the user, so no manual entering is required. However, these last three methods provide less customization options compared to the Section 4.2.

4.2. ADS Method

Use of the ADS API requires that the developer has read and agrees to the NASA/ADS terms and conditions which can be found at https://adsabs.github.io/help/terms/. In order to use the ADS method, in addition to the bibcode from the initial Section 4.1 step, the user will need a valid token in order to use ADS's API. A token can be issued when the user registers for an account on NASA/ADS at https://ui.adsabs.harvard.edu/ and is required for every fetch of a bibliographic entry while using the ADS method. This method allows the user to customize the bibliography output in various ways. The set customization, already programmed for the HITRAN and AMBDAS administrators, is widely used. The structure of the reference output is as follows and in the following order:

- notes are entered first
- the authors are listed as F. I. Surname, with commas in between each author
- the title of paper is in quotations
- the journal is italicized
- the volume number is in bold
- the pages of the article are written as first page–last page
- the year of the article is written in parentheses
- hyperlinks are finally included at the end, one for the DOI hyperlink and one for the ADS hyperlink

ADS has made it available to include more details and information in bibliographies when using their database. A user has access to information such as: the abstract, copyright, citation count, author affiliation, keywords, publication category and the arXiv e-print number of the article, etc. The user also has more output options such as: EndNote, ProCite, RIS (Refman), RefWorks, MEDLARS, AASTeX, Icarus, MNRAS, Solar Physics (SoPh), DC (Dublin Core) XML, REF-XML, REFABS-XML, VOTables and RSS. See https://adsabs.github.io/help/actions/export for more details and information.

Most importantly, HITRAN and AMBDAS chose to utilize ADS because of efficiency. The administrator has the option to search more than one bibcode at a time, allowing for multiple reference retrievals in any of the three designated formats desired. All the administrator will need to do is enter their personal token, and their bibcode(s). There are three formats that HITRAN uses for its reference methods: BibTeX, JSON and HTML. Therefore there are three main search parameters for users to use when creating bibliographies; all three of these are already custom formatted.

4.3. urllib Method

`urllib` is a package that collects several modules for working with URLs. The two sub-modules used in this method are `urllib.request` and `urllib.error`. The `urllib.request` module defines functions and classes which help in opening URLs, basic and digest authentication, re-directions, cookies and more. While the `urllib.error` contains the exceptions raised by `urllib.request` [18].

This method, titled `urllib` method, provides a JSON formatted output along with hyperlinks and the source DOI. The administrator need only enter the DOI for the article or paper that is being cited into the program. Upon doing so and running the method, the user will retrieve the complete JSON bibliography of the paper. The administrator is not required to complete this method or the following methods titled Section 4.4 or Section 4.5; the Section 4.2 will provide the desired outputs that would be generated in these methods. This method and those following are made available in case the paper is not in the ADS database only, since the ADS contains items within (or relating to) the field of astrophysics.

4.4. JSON-to-HTML Method

In this method, the administrator will need to import the Python module `html` [19] only. The JSON-to-HTML method provides an HTML encoded output for the administrator while the only information it is given is the JSON encoded bibliography from the Section 4.3. The entire JSON output from the Section 4.3 will automatically be populated into the designated region in this method. Therefore, the administrator need only to run the method to retrieve the now HTML encoded bibliography. The html encoder will replace the following characters (& < " ' >) with recognized html entities (&amp; &lt; &quot; &#x27; &gt;) respectively.

4.5. BibTeX Method

In this method there are several modules and packages that the administrator will need to install or import, including: `urllib.request` with the sub-modules `quote`, `Request` and `urlopen` [18]. Users will also need to install the `BeautifulSoup` package [4], and import the Python standard library packages `re` (for regular expression matching), `logging` (for applications to configure different log handlers and a way of routing log messages to these handlers) and the `html.entities` sub-module `name2codepoint`.

This particular method, titled BibTeX method, was designed so that the user's initial DOI entered in the Section 4.1 would automatically populate. Once the administrator runs the code with the entered DOI, they will be shown a detailed bibliography, in LaTeX format, with labels to all the corresponding information retrieved. The bibliography will contain, from top down, the paper's title, list of authors, journal, volume, number, pages, year and publisher. The only requirement on the user's side of things, is to enter the DOI for the paper or article into the initial prompt in the Section 4.1 and run that program first.

It is important to note that this method is the only one that will be able to also work for other keywords when searching for bibliographies. For example, if one were to search the title of the paper, instead of the paper's DOI, then they would receive a full bibliography corresponding to the information entered into the program. However, this bibliography may point to an incorrect article or paper. Therefore, even though this is a convenient option of simply entering the title of the paper, it is recommended that users use the DOI of the article instead.

5. Conclusions

In this paper we describe a new, automated referencing system to provide consistent, accurate and detailed bibliographies to every source of data in scientific databases. The goal of this work is to encourage an environment that promotes data sharing provenance and good practice amongst researchers and databases. Overall, it is imperative that all contributions to scientific databases receive acknowledgement through proper referencing to their cited material. All code for the automatic referencing system was written in the Python programming language through the web-based, interactive computing notebook environment, Jupyter Notebook and is also implemented in a Django application available at https://github.com/hitranonline/pyref.

This work provides a convenient bibliographic system, to allow database administrators to implement bibliographies faster and without human errors into their database systems. Users utilizing this system need only enter a single line of information into the program in order to obtain the complete bibliography entry for the paper they wish to cite. This system was designed and tested for the HITRAN and AMBDAS databases, but can also be used through the raw code outlined in the Jupyter Notebook or integrated into a database through the Django application. The complete code

[4] Richardson, L., beautifulsoup4, **2019**, Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree. Covered under MIT License. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

for the new referencing system is provided through a permissive, open-source and cost-free basis under version 2 of the Apache Software Foundation License [11], along with detailed instructions and resources for customizing the output formats of citations and implementation of the Django application. Ideally, with enough availability and use of proper referencing systems, databases and researchers alike will be encouraged to share information between databases, in turn making their work more accessible to users.

Author Contributions: Conceptualization, I.E.G., L.S.R.; methodology, I.E.G., C.H., F.M.S.; software, F.M.S., R.J.H., K.E.L. and C.H.; validation, I.E.G. and F.M.S.; formal analysis, I.E.G.; investigation, F.M.S., C.H.; resources, I.E.G., F.M.S. and K.E.L.; data curation, F.M.S., C.H.; writing–original draft preparation, F.M.S.; writing–review and editing, I.E.G., C.H.; visualization, F.M.S.; supervision, I.E.G.; project administration, I.E.G.; funding acquisition: I.E.G., C.H. All authors have read and agreed to the published version of the manuscript.

Funding: FMS acknowledges the internship though the Smithsonian Astrophysical Observatory Latino Initiative Program funded by the National Science Foundation under Grant No. 1745460. The HITRAN database is supported through the NASA Aura and PDART grants NNX17AI78G and NNX16AG51G, respectively.

Acknowledgments: Joshua Karns for technical support while coding in Python. The NASA Astrophysics Data System allowed use of their API during development of this software and provided helpful personal guidance and assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ADS	NASA Astrophysics Data System
AMBDAS	Atomic and Molecular Bibliographic Data System
API	Application Programming Interface
CDMS	Cologne Database for Molecular Spectroscopy
DOI	Digital Object Identifier
GEISA	Gestion et Etude des Informations Spectroscopiques Atmosphériques
HAPI	HITRAN Application Programming Interface
HITEMP	High-Temperature Molecular Spectroscopic Database
HITRAN	High-Resolution Transmission Molecular Absorption Database
HTML	Hypertext Markup Language
JPL	Jet Propulsion Laboratory
JSON	JavaScript Object Notation
NASA	National Aeronautics and Space Administration
NFRI	National Fusion Research Institute
NIFS	National Institute for Fusion Science
NIST	National Institute of Standards and Technology
VAMDC	Virtual Atomic and Molecular Data Centre

References

1. Gordon, I.; Rothman, L.; Hill, C.; Kochanov, R.; Tan, Y.; Bernath, P.; Birk, M.; Boudon, V.; Campargue, A.; Chance, K.; et al. The HITRAN2016 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.* **2017**, *203*, 3–69. [CrossRef]

2. Rothman, L.S.; Gamache, R.R.; Goldman, A.; Brown, L.R.; Toth, R.A.; Pickett, H.M.; Poynter, R.L.; Flaud, J.M.; Camy-Peyret, C.; Barbe, A.; et al. The HITRAN database: 1986 edition. *Appl. Opt.* **1987**, *26*, 4058. [CrossRef] [PubMed]
3. Rothman, L.; Jacquemart, D.; Barbe, A.; Chris Benner, D.; Birk, M.; Brown, L.; Carleer, M.; Chackerian, C.; Chance, K.; Coudert, L.; et al. The HITRAN 2004 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.* **2005**, *96*, 139–204. [CrossRef]
4. Jacquinet-Husson, N.; Armante, R.; Scott, N.; Chédin, A.; Crépeau, L.; Boutammine, C.; Bouhdaoui, A.; Crevoisier, C.; Capelle, V.; Boonne, C.; et al. The 2015 edition of the GEISA spectroscopic database. *J. Mol. Spectrosc.* **2016**, *327*, 31–72. [CrossRef]
5. Pickett, H.; Poynter, R.; Cohen, E.; Delitsky, M.; Pearson, J.; Müller, H. Submillimeter, Millimeter, and Microwave Spectral Line Catalog. *J. Quant. Spectrosc. Radiat. Transf.* **1998**, *60*, 883–890. [CrossRef]
6. Endres, C.P.; Schlemmer, S.; Schilke, P.; Stutzki, J.; Müller, H.S. The Cologne Database for Molecular Spectroscopy, CDMS, in the Virtual Atomic and Molecular Data Centre, VAMDC. *J. Mol. Spectrosc.* **2016**, *327*, 95–104. [CrossRef]
7. Tennyson, J.; Yurchenko, S.N.; Al-Refaie, A.F.; Barton, E.J.; Chubb, K.L.; Coles, P.A.; Diamantopoulou, S.; Gorman, M.N.; Hill, C.; Lam, A.Z.; et al. The ExoMol database: Molecular line lists for exoplanet and other hot atmospheres. *J. Mol. Spectrosc.* **2016**, *327*, 73–94. [CrossRef]
8. Rey, M.; Nikitin, A.V.; Babikov, Y.L.; Tyuterev, V.G. TheoReTS—An information system for theoretical spectra based on variational predictions from molecular potential energy and dipole moment surfaces. *J. Mol. Spectrosc.* **2016**, *327*, 138–158. [CrossRef]
9. Hill, C.; Gordon, I.E.; Kochanov, R.V.; Barrett, L.; Wilzewski, J.S.; Rothman, L.S. HITRANonline: An online interface and the flexible representation of spectroscopic data in the HITRAN database. *J. Quant. Spectrosc. Radiat. Transf.* **2016**, *177*, 4–14. [CrossRef]
10. Gordon, I.E.; Potterbusch, M.R.; Bouquin, D.; Erdmann, C.C.; Wilzewski, J.S.; Rothman, L.S. Are your spectroscopic data being used? *J. Mol. Spectrosc.* **2016**, *327*, 232–238. [CrossRef]
11. Apache Software Foundation. *Apache Licence, Version 2.0*; Apache Software Foundation: Forest Hill, MD, USA, 2020. Available online: https://www.apache.org/licenses/LICENSE-2.0 (accessed on 30 April 2020).
12. Dubernet, M.L.; Antony, B.K.; Ba, Y.A.; Babikov, Y.L.; Bartschat, K.; Boudon, V.; Braams, B.J.; Chung, H.K.; Daniel, F.; Delahaye, F.; et al. The virtual atomic and molecular data centre (VAMDC) consortium. *J. Phys. B At. Mol. Opt. Phys.* **2016**, *49*, 074003. [CrossRef]
13. Zwölf, C.M.; Moreau, N.; Dubernet, M.L. New model for datasets citation and extraction reproducibility in VAMDC. *J. Mol. Spectrosc.* **2016**, *327*, 122–137. [CrossRef]
14. Zwölf, C.M.; Moreau, N.; Ba, Y.A.; Dubernet, M.L. Implementing in the VAMDC the new paradigms for data citation from the research data alliance. *Data Sci. J.* **2019**, *18*, 1–13. [CrossRef]
15. Kurtz, M.J.; Eichhorn, G.; Accomazzi, A.; Grant, C.S.; Murray, S.S.; Watson, J.M. The NASA Astrophysics Data System: Overview. *Astron. Astrophys. Suppl. Ser.* **2000**, *143*, 41–59. [CrossRef]
16. Kochanov, R.; Gordon, I.; Rothman, L.; Wcisło, P.; Hill, C.; Wilzewski, J. HITRAN Application Programming Interface (HAPI): A comprehensive approach to working with spectroscopic data. *J. Quant. Spectrosc. Radiat. Transf.* **2016**, *177*, 15–30. [CrossRef]
17. International DOI Foundation. *The International DOI Foundation (IDF) Is a Not-For-Profit Membership Organization That Is the Governance and Management Body for the Federation of Registration Agencies Providing Digital Object Identifier (DOI) Services and Registration, and Is the Registration Authority for the ISO standard (ISO 26324) for the DOI System*; International DOI Foundation: Westwood, MA, USA, 2018. Available online: https://www.doi.org/ (accessed on 30 April 2020).
18. urllib, 2019. urllib is a package that collects several modules for working with URLs. Available online: https://docs.python.org/3/library/urllib.html (accessed on 30 April 2020).
19. html, 2019. Simple, elegant HTML, XHTML and XML generation. Available online: https://docs.python.org/3/library/html.html (accessed on 30 April 2020).