Comment on the proposed conservation of usage of *Testudo gigantea* Schweigger, 1812 (currently *Geochelone* (*Aldabrachelys*) *gigantea*) (Reptilia, Testudines)
(Case 3463; see BZN **66**: 34–50, 80–87, 169–186, 274–290, 352–357; **67**: 71–90, 170–178)

Tara Lawrence

Foundation for Ecological Research Advocacy and Learning, Puducherry, India
(e-mail: tara@feralindia.org)

Neil Pelkey

Juniata College, Huntingdon, PA, and Foundation for Ecological Research Advocacy and Learning, Puducherry, India (e-mail: pelkey@juniata.edu)

Sara Soares

Universidade de Lisboa, Faculdade de Ciências, Lisboa Portugal
(e-mail: sara.pc.soares@gmail.com)

'Googleology': powerful tool or unreliable evidence?

Introduction

In recent years the internet has become a much used and very important tool for diverse forms of research. The various 'search engines' that allow users to execute searches on words or expressions and have results in a matter of seconds are especially useful and of these the Google search engine is probably the most popular (see Nielsen's Net Ratings at http://blog.nielsen.com/nielsenwire/online_mobile/top-u-s-search-sites-for-may-2010/ for U.S.A. based results). These technologies are not only relatively young but they are evolving very rapidly, so it is a challenge even for specialists in information technology to keep up to date with these very powerful tools. We hope that this article will help taxonomists to employ these tools correctly.

It is quite common to use a tool without reading the manual. Many of us have suffered the consequences of such expedience and realise that it is usually unwise. It is especially unwise when using that tool for scientific measurement. Google is no exception. A recent comment by Bour et al. (2010, BZN **67**: 73–77) used the results of Google searches as evidence to support their arguments about the relative frequency of certain generic and specific names used for the Aldabra tortoise. We will show that this approach is an unacceptable way to use internet search engines for claims of notoriety or popularity of usage. Of particular importance, authors using such an approach with the Google search engine should: (1) consult the Google manual; (2) consult the literature on using search engine hits as a representation of notability; and (3) perform post-collection data validation and verification.

Step 1: Read the Manual

The Google (2010a) documentation states: 'Google's calculation of the total number of search results is an estimate. We understand that a ballpark figure is valuable, and by providing an estimate rather than an exact account, we can return quality search

results faster. In addition, when you click on the next page of search results, the total number of search results can change. In this case, we realise that some of the query results are duplicates, and collapse those duplicates so that you can find the specific result you're looking for more easily. Collapsing the duplicates decreases the estimated number of results, as well as the overall number of results pages' (Google, 2010a).

How much does this 'collapsing' impact the 17,600 records that Bour et al. (2010, BZN **67**: 73–77) reported for *Dipsochelys dussumieri* (Gray, 1831)? We begin with a total number of 17,100 (slightly fewer than the 17,600 reported in the Bour et al. paper). We then click forward one page to look at a few of the results and the number drops immediately to 3180. We then go to the last page of the search results to find out that after collapsing duplicates there are 547 retrievable pages. The Bour et al. estimate using the initial Google estimate of 17,600 was off by 17,000.

When the Google documentation states, 'We're aware that we sometimes return erroneous estimates for the number of results that return for a query, and we're working to improve these estimates' (Google, 2010b) it should be heeded.

Step 2: Scholarship

While at first glance the table of figures in Bour et al. (2010, BZN **67**: 76), with purported quantification of usage of both generic and specific names, appears thorough, the authors failed to do a follow-up investigation on the results of their Google searches: they evidently did not query many of the 17,600 hits that they reported. Had they done so, they would have found that the vast majority are websites and web addresses, with a very small proportion of 'hard copy' publications, whether published as 'grey literature' or in peer-reviewed scientific journals. A follow-up, or literature search, quickly points out some of the pitfalls of the Google 'of about' statistic. Kilgarriff (2007) documents the substantial differences between actual 'hits' (i.e. individual records from an internet search) and real pages, but he also points out a variety of other potholes. Particularly, he explains that hits do not imply anything about the quality of those records. That is, a blog entry and a peer-reviewed publication count exactly the same in this sort of search. Uyar (2009) compares three different search engines (Google, Yahoo and Bing) across single word and multiple word accuracies and finds each of them wanting both in terms of providing pages which are not truly related to the search terms and also missing pages that are in fact related: 'The percentages of accurate hit count estimations are reduced almost by half when going from single word to two word query tests in all three search engines. With the increase in the number of query words, the error in estimation increases and the number of accurate estimations decreases.' (Uyar, 2009).

Bour et al. (2010, p. 76) noted that the Google search results were unstable, but they did not investigate further. Instead, they attributed the instability to some real phenomenon concerning a nomenclatural debate about the name of the Aldabra tortoise. However, the instability in the Google results is evidence that the results that they reported bear little relationship to the notability of the competing names for the Aldabra tortoise. Furthermore, a comparison of the use of other terms could have pointed out the folly of the approach. For example, Google searches on the terms ['venomous snakes'] and ['poisonous snakes'] provide an interesting comparison. ['Venomous snakes'] returns 250,000 'results', with a retrievable number of 817.

['Poisonous snakes'] also gets about 250,000 'results', but with 833 retrievable. The numerical superiority of the second expression is unlikely to convince many herpetologists to change their lexicon however. Even Wikipedia (2010a) does not recommend the use of Google's results count as a measure of notability.

Note that Google uses a convention of square brackets to punctuate search terms in their documentation. That is, the brackets help differentiate ["word1 word2"] meaning the exact phrase "word1 word2", and [word1 word2] meaning both words appear in the document. We only use the ["Genus species"] for all searches reported in this study because using [Genus species] can lead to results where the generic term refers to the tortoise and the specific term refers to an altogether different animal. The specific search [gigantea] by itself estimates 782,000 results and [dussumieri] returns an estimated 261,000 results.

Step 3: Validate the data

Not checking errors in one's data is rarely wise, albeit that many of us have fallen prey to it from time to time. Yet, in the case of the numbers presented by Bour et al. (2010), confirmations bias seems to have led to faith in data unseen and data unknown. A simple check of a few of the lower PageRanked websites (see below for an explanation of PageRanked) would have shown that the results were not all directly relevant to usage of the nomenclature under consideration. For example, towards the end of the list, in the 400s, we find a pet food store, a reptile care forum, and translations of the Wikipedia page on the tortoise (the entire list of 547 records has been deposited with the ICZN Secretariat). While the name *Dipsochelys dussumieri* occurs – inconspicuously – in a short, obscure list on several of these websites, none of them is likely to have much relevance to anyone looking for the scientific name of the Aldabra tortoise. Furthermore none of these sites have the appearance of being an authoritative source of information on scientific names.

A few examples of the most irrelevant results ostensibly for *D. dussumieri* are:

(1) Pet food store http://www.reptile-food.ch/Galerie-Reptilien-in-freier-Natur/. (The name *Dipsochelys dussumieri* appears in an inconspicuous list).

(2) Reptile forum http://www.reptileforums.co.uk/forums/shelled-turtles-tortoise/320545-substrate.html. The usage in the case of this forum is merely a tag in the forum topic. This forum includes the erudite posting about white stuff in tortoise urine. (The name *Dipsochelys dussumieri* appears in an inconspicuous list).

(3) A blog entry. This blog entry is about cats and roses – hardly a use of the term under discussion, let alone a scientific reference to the Aldabra tortoise. http://rvoulgari.blogspot.com/2008_08_01_archive.html (no apparent mention of a scientific name for the tortoise).

(4) A picture of a snake. http://www.clin-dieu.be/Galeries/animaux/animaux.php?zoom=1&d=6&page=1&nb_img=7&break=&picture_id=177&sub_category_id=1 (no apparent mention of any scientific name for the Aldabra tortoise).

Within the 632 records for the search ['*Geochelone gigantea*'] there are similarly irrelevant results, including 5 postage stamps, 5 library search engines and 2 pornographic sites. The following are a few examples of the most irrelevant results from a search on ['*Geochelone gigantea*'].

(1) "Aldabrasköldpaddan (Geochelone gigantea) lever på Aldabra i Seychellerna och" and "Aldabranjättiläiskilpikonna (**Geochelone gigantea**) on erityisesti Aldabran

	GENUS				
Species	***Testudo***	***Geochelone***	***Aldabrachelys***	***Dipsochelys***	**Species total all genera**
gigantea	**177**	**883**	**74**	**10**	**1144**
elephantina	**45**	**0**	**39**	**44**	**128**
dussumieri	**19**	**6**	**6**	**132**	**163**
Genus total all species	**241**	**889**	**119**	**186**	**1435**

Table 1. Number of publications that include any of the 3 species names in combination with any of the 4 generic names above.

atollilla Seychelleillä tavattava yksi maailman suurimmista kilpikonnista'' are exact Swedish and Finnish translations of the English language Wikipedia page; they are not separate, unique uses of the name.

(2) www.smacksy.com/2009_10_01_archive.html. The site is a personal blog dedicated to funny stories about a child named Bob who wanted to know if two turtles were kissing.

(3) http://www.pets-classifieds.co.uk/sed/33112.html. The site seems a good place to find a squirrel or a Jack Russell terrier, but has little relevance to the use of names for the Aldabra tortoise.

The above examples reflect three issues with the standard Google search as a measure of usage:

(1) Sites that have no mention anywhere of any scientific name of the Aldabra tortoise, and are thus completely irrelevant to the list.
(2) Sites that might have a scientific name of the Aldabra tortoise, but it is inconspicuous, unlikely to have much relevance to people looking for the 'correct' name, and certainly not a professional/authoritative site.
(3) Exact copies of information on other pages.

A verification of each of the 527 results for *D. dussumieri* and 632 results for *G. gigantea* provided further illumination of the issues with the general Google search approach (See Fig. 1). The most common use of *dussumieri* is in wikis, with 81 results. Wikis are group edited, online documents. These may be managed by large or small groups of people. The most famous is the Wikipedia site which is one of the most widely used online encyclopedias (Leuf & Cunningham, 2001). Many (40) of the 81 wiki sites for *dussumieri* were the English language Wikipedia accounts or direct translations into other languages. Journals, books, and magazine articles followed with 73 (10 of these were duplicates, but linked in different sources so we leave them in as evidence of the potential issues involved in the Bour et al. approach). *G. gigantea* had nearly twice the number of journal, book and magazine articles with 128. Overall, including many clearly irrelevant hits for both of the two scientific names under discussion, there were 18% more hits for *G. gigantea* than for *D. dussumieri*.

The table on page 76 in Bour et al. (2010) is misleading. It lacks an understanding of the tool used, simple scholarship, and basic data validation. The 'occurrence' figures therein bear little relationship to actual documents, unique references, or in some cases references to the Aldabra tortoise at all. In this light, the table is highly

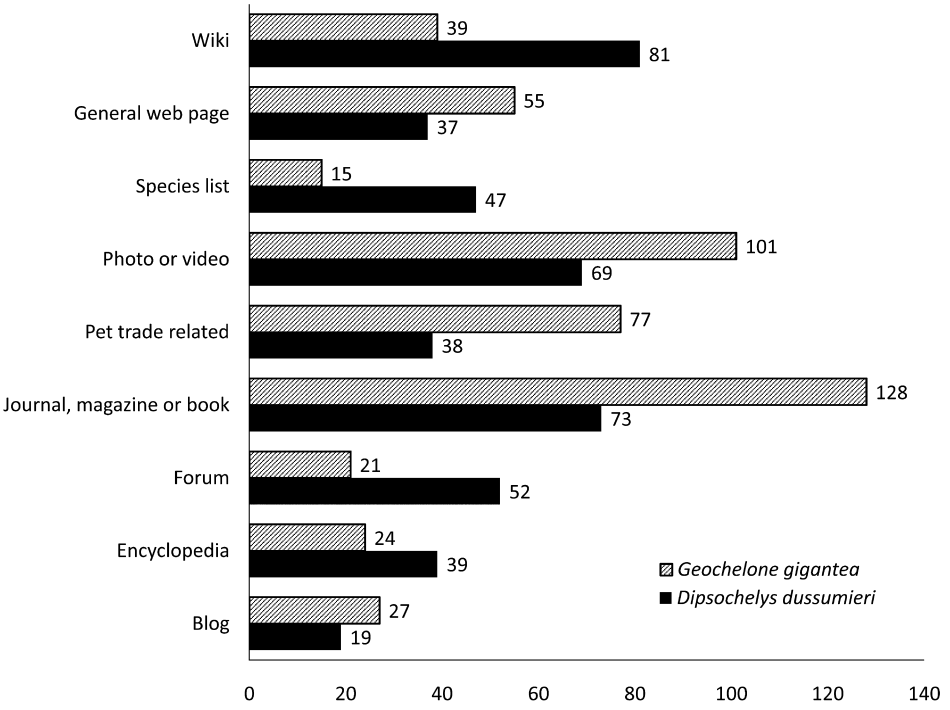

Fig. 1. Frequencies of 'hits' for the searches ['*Geochelone gigantea*'] and ['*Dipsochelys dussumieri*'] grouped into general categories (the top nine are presented here): overall total 'hits' for *G. gigantea*=632, overall total 'hits' for *D. dussumieri*=527 (for 26 May, 2010).

misleading because the figures presented do not represent what they are claimed to represent.

Bour et al.'s (2010) results should be viewed as inaccurate and their method avoided. Few biological researchers know the inner workings of Google searches. The code is, in fact, not published. There are only a few articles and Google webmaster manual pages that give a hint of the inner workings. In the same vein, most of us do not know how rack and pinion steering or disk brakes work, but we depend on them to get to work every day when using an automobile. Google and other search engines have become, to a great extent, like an appliance. We know how to use its basic features and we depend on it, but the finer details are a mystery. The following sections will demystify Google a little, as well as present a better method for compiling the sources on an issue using the internet.

Why Google doesn't give every page on every search

Google's purpose in design was, and still is, to be very fast while overcoming the spam and junk pages that plagued earlier search engines (Page et al., 1998). Google founder, Larry Page, began with a citation index concept that assumed that if a lot of people on the web thought something was important, the searcher would too. The Google search algorithm, called 'PageRank', is similar in principle to the 'H citation index' (Hirsch, 2005). The more times a page is linked by other pages, the higher the

PageRank (i.e. page rank). The higher the page rank of the linking page, the higher the weight. PageRanks go from 0–10, with 10 being the most important.

Google obtains these ranks by 'crawling' the web. That is, it downloads and parses each page that it determines as relevant (or that the owner/operators submit to the indexing engine via Google's webmasters site (Google, 2010c). Those pages are parsed for all the meaningful words, and an index is created of where in the document each word exists. That document is given a document ID number, compressed, and stored. The index data is then added to the main index which is distributed to local search computers around the world that perform the actual searches.

Google does not index every web page on the net. In fact you can block your site from being indexed using a couple simple techniques. At the same time, a website owner/operator can purposefully link their page to a certain word or expression, even though the website has nothing at all to do with the term; in this way, by selecting popular terms, an owner can increase exposure to their website and thereby increase sales and revenues. There are also locations where the top trends for a day are listed, such as the Google trends site (Google, 2010d). A web site operator can look at the hot trend words and add them to the site or buy AdWords reflecting hot topics.

When a search term is entered by an internet user, Google searches for that word in the index. It computes both relevance and page rank statistics, and then sorts the results. Google returns the top results according to weighting by relevance and page rank and then proceeds to lower relevance page rank items. When the search *'Geochelone gigantea'* is performed, then the retrievable hits (those with a high enough rank and relevance) are displayed in a sequential series of pages. Another statistic that says 'results 1–10 of about xxx,xxx' also appears on the top right of the first of these pages. This led to 'about 17,100 results' when the Google search [*Geochelone gigantea*] was performed on May 22nd, 2010. There were 632 retrievable results from the 17,100. The most relevant sites or most popular sites were near the top of the listing. The results that contained many of the journal articles appeared near the middle of the listings. This result is not unique however. If the exclusion terms for wiki, forum, blog, YouTube, etc are added, i.e. the search is on: 'Geochelone gigantea' -forum -wiki -blog -youtube -photo -blogspot, then Google returns an estimate of 11,000 pages. Many might think this would be a subset of the previous search with the pages containing the negated terms removed, but it is not. Of these, only 730 are retrievable, but the results contain pages not in the first search. Hence the results of the second search were not a subset of the first search, but a different set with different relevance and rank statistics. While this might seem logically inconsistent, it is a reflection of Google's trade-off between, speed, relevance, and popularity.

There are strategies and activities a webmaster can employ to improve the page rank. This is known as search engine optimisation or SEO. One easy and direct way to do this is to edit a Wikipedia page and put links to your site(s) in there. Since Wikipedia is the most popular online encyclopedia (Alexa, 2010), your page rank would probably increase. For example, the Aldabra Giant tortoise page at Wikipedia (Wikipedia, 2010) has two links to islandbiodiversity.com and one to arkive.org. Arkive.org in turn links to islandbiodiversity.com. This link-back and link clustering can help raise a website's page rank and thus increase its chance of being seen in a Google search.

Google is constantly adding new features to make sure that page owners do not play tricks with Google to falsely elevate their rank. The web entrepreneurs are constantly trying to outwit Google's attempts, however. Hence there is a bit of Alice in Wonderland's Red-Queen who required 'all the running you can do, to keep in the same place' in the ever-changing Google technology (Carol & Haughton, 2003 for the literary reference; Van Valen, 1977 for an evolutionary reference).

Another approach to getting one's website to the top of Google's page is through the use of AdWords. That is, advertisers and any other interested party pay Google to put their link at or near the top when certain words are searched. If someone clicks on the sponsored link, the advertiser pays Google a pre-negotiated rate per click. There is no limit to the number of organisations that can purchase the same AdWord, but that organisation or person paying the most for the search word gets to be at the top of the display. If one does a search on 'tortoise', Wikipedia comes up first, but just to the right in the sponsored links section is a 'sponsored link' for Galapagos.org and ask.com.

One of us (N.P.) purchased the AdWords 'geochelone' and 'dipsochelys' to get into the sponsored links space on Google. A tight maximum daily cap of $2.00 was set so this sponsored link will not come up with every search, but it could if we wanted to pay the cost. So for 0.75 U.S. dollars, neilpelkey.net is on par with Wikipedia with the conspicuously displayed notice 'Study in the Andamans', even though neilpelkey.net has a PageRank of just 0.

Thus while Google is impressive technology that can produce rapid and highly relevant web searches, it puts Wikipedia, online archives, blogs, and web forums at a higher rank than published literature. It is also fairly time-consuming to check each of the links that are fetched. For example, we spent an estimated 26 hours doing web searches to provide the basic information reported here.

'Scientific Webology': A better approach

We suggest that a better approach for determining the relative frequency of use of different scientific names is a dedicated search of the published literature, legitimate web exchanges, and web data sources. While Google Scholar might seem like a viable alternative to the normal Google, it is far from complete. We therefore set about to search specialised bibliographic web sources including SCIRUS, PubMed, and Oxford Journals. We also used proprietary sources including EBSCO Premier (Taylor & Francis, etc.), Science Direct (Wiley Interscience), BioOne and Jstor. While this would not find every publication, it should start to approach a reliable comparison of publications or electronic usage.

It is important to point out that monographs and books are less likely to be archived on the web than shorter journal articles and reports. Also, newer articles are more likely to be available on the web than older articles. Some articles which have not been digitised may have no web reference at all. Hence, for the case in hand, some older records that used any scientific names that included the species *gigantea*, which has been in use since 1812, would be invisible to web searches. The earliest electronic reference we found was to Lartet et al. (1851), but clearly the percentage of early journals and other sources currently digitised is a small proportion (less than 5%; see Kelly, 2006). Furthermore grey literature such as correspondence and newsletters

from previous decades will never be digitised while more current grey literature such as Wikipedia and other user-driven encyclopedias will appear atop the Google search results.

The steps used here to provide a more accurate estimate of usage of the generic and specific terms were to:

(1) have a single researcher search the commercial sites Science Direct, EBSCO Academic Search Premier, Jstor, and BioOne;

(2) have two different researchers do exactly the same searches independently using the publicly available search engines: SCIRUS, Oxford, PubMed, and Google Scholar;

(3) repeat steps (1) and (2) for the four generic, three specific, and twelve binomial names that are shown in the table in Bour et al. (2010);

(4) download all citations located in all of these searches into the online Zotero bibliographic data tool (Zotero, 2010);

(5) inspect each record, cleaning up the citations and removing duplicates from each generic, specific, binomial combination;

(6) upload the cleaned-up records to the online bibliographic service Cite-U-Like, attaching tags for genus, species, and the combinations thereof; and

(7) compile the results.

This approach provides a publicly accessible bibliography with electronic links to all the citations electronically available. It will also provide, available on request, a database of all of the citations located in this study, which can be used for further analysis. It should be noted that the database still includes some duplicates. This is because some articles list more than one generic and/or specific name and/or binomial combination, which is a function of the search process by combinations. There are also some links that have become 'dead' since the original search. We invite people to upload legitimate citations that we may have missed that have scientific or academic use of any of the scientific names under discussion; people are also invited to comment on links to citations that are broken.

Results

The results of the above approach are presented in Fig. 1 and are available at http://www.citeulike.org. The citations can all be seen by signing up for citeulike and joining the group 'dussumieri vs. gigantea'. The citeulike group data presented here is for May 26, 2010.

It is clear that the scientific and academic use of *gigantea* is an order of magnitude higher than that of either *dussumieri* or *elephantina*. *Geochelone* has five times the usage of *Dipsochelys* and four times that of *Testudo*. There might be other arguments for nomenclature based on other criteria, but the common usage measured by a more reliable electronic search leans strongly in favor of *gigantea*. Even in the flawed approach used in Bour et al., *dussumieri* produced half the use of *gigantea* in regard to 'hard copy' publications.

Acknowledgements

Thanks to Jack Frazier, Janaki Lenin, Anand Pillai, Uwe Fritz, Jeanne Mortimer and two anonymous reviewers for invaluable comments and editorial suggestions.

References

Alexa. 2010. Alexa Web Information Company 3 Month Traffic Ranking for Wikipedia.org. http://www.alexa.com/siteinfo/wikipedia.org

Bour, R., Pritchard, P.C.H. & Iverson, J.B. 2010. Comments on the proposed conservation of usage of *Testudo gigantea* Schweigger, 1812 (currently *Geochelone* (*Aldabrachelys*) *gigantea*) (Reptilia, Testudines). *Bulletin of Zoological Research*, **67**(1): 73–77.

Carroll, L. & Haughton, H. 2003. *Alice's adventures in Wonderland and 'Through the looking-glass and what Alice found there'. Lewis Carroll. Ed. and with an introd. and notes by Hugh Haughton*. Penguin classics. Penguin, London.

Google. 2010a. How does Google calculate the number of results? – Webmaster Tools Help. http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=70920.

Google. 2010b. Google unique page count issues. N.D. http://code.google.com./apis/searchappliance/documentation/62/xml_reference.html#appendix_num_results.

Google. 2010c. Google webmaster tools. https://www.google.com/webmasters/tools/home?hl=en

Google. 2010d. Google Trends**.** http://www.google.com/trends/hottrends

Hirsch, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(46): 16569–16572.

Kelly, K. 2006. Scan This Book! The New York Times, May 14, Magazine. http://www.nytimes.com/2006/05/14/magazine/14publishing.html?_r=2&oref=slogin&pagewanted=all.

Kilgarriff, A. 2007. Googleology is bad science. *Computational Linguistics*, **33**(1): 147–151.

Lartet, E., Noulet, J.B. & Dupuy, D. 1851. *Notice sur la colline de Sansan, suivie d'une récapitulation des diverses espèces d'animaux vertébrés fossiles, trouvés soit à Sansan, soit dans d'autres gisements du terrain tertiaire du miocène dans le bassin sous-pyrénéen.* impr. J.A. Portes.

Leuf, B. & Cunningham,W. 2001. *The Wiki Way : quick collaboration on the Web*. Addison-Wesley, Boston.

Page, L., Brin, S., Motwani, R. & Winograd, T. 1998. The PageRank citation ranking: Bringing order to the web. (Technical Report). Stanford University.

Uyar, A. 2009. Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, **35**(4): 469–480.

Van Valen, L. 1977. 'The Red Queen'. *American Naturalist*, **111**(980): 809–810.

Wikipedia. 2010a. Wikipedia: search engine test – Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Google_tests

Wikipedia. 2010b. Aldabra Giant Tortoise – Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Aldabra_Giant_Tortoise&oldid=358610481

Zotero. 2010. Zotero: See it. Save it. Sort it. Search it. Cite it. http://www.zotero.org/