Collaborative Reference Management System

Willem H. le Roux and Ivan D. Burke
Council for Scientific and Industrial Research

Abstract – Citation has always been a crucial part of research when it comes to accreditation for intellectual property. The maintenance of these references has mostly been the sole responsibility of the author to collect, maintain and format. There are tools to assist in this regard, both commercial and open source with varying degrees of support for user specific needs, yet it became apparent that a gap emerged between these utilities; there were no real support for collaborative reference sharing or maintenance. To address this problem a prototype collaborative reference management system was designed. This paper indicates the methodology used to implement this prototype as well as the lessons learned during the development of the system. It also identifies two possible research questions that might be worth investigating in future research.

Introduction

Research within our organisation has lately become predominantly collaborative based. This has been aided by collaboration tools such as version control systems and multimedia based tools such as wikis, web logs and communities of practice within the organisation. A need arose for a similar tool to share and maintain bibliographical data.

This paper documents the research related to bibliographical management system with a brief overview of currently available management systems. Though there are several standalone and web based personal reference management tools available (some of which will be discussed in this paper) there are no simple means of collaboration or community support for these tools

The paper also encapsulates a literature review of reference management systems as well as a review of current trend in collaborative research and tools used to facilitate this process. The reasoning behind the separate literature reports is that both these concepts have been researched at length in the information science community yet the union of these concepts hardly feature in any of them.

Lastly we state the potential for the union of these concepts as well as results achieved using prototypes we developed to test the maintenance system. The prototypes also lead to some interesting possibilities for further research

Literature study: Bibliographical reference management tools

BibTeX[1] was identified as the most common means of maintaining references in our organisation due to its open standards and separation of the data from its representation.

BibTeX databases can be created using plain text editors but this process is cumbersome and prone to errors. There is however an assortment of BibTeX software available on the market. Table 1 shows the key attributes of each.

[1] Full BibTeX specifications can be obtained at www.bibtex.org

Software	Developer	Cost[2]	Open Source	OS	Export Formats	Import Formats
2collab	Elsevier	Free	No	All	BibTeX, RIS and CSV	RIS and browser bookmarks
CiteULike	Richard Cameron	Free	No	All, web-based	BibTeX, RIS and CoinS	BibTeX
EndNote	Thomson Corp.	US$299.95	No	Windows, Mac OS	BibTeX, Medline, BiblX, Refer, RIS	BibTeX (partially), CSA, BiblX, ISI, Medline, Ovid, PubMed, RIS, SciFinder
JabRef	JabRef developers	Free	Yes, GPL License	All	BibTeX, BiblX, MODS XML	BibTeX (partially), CSA, BiblX, ISI, Medline, Ovid, PubMed, RIS, SciFinder
refDB	refDB developers	free	Yes, GPL License	All	BibTeX, BiblX, MODS XML, RIS	BibTex, Compac, ISI, BiblX, MODS XML, RIS, PubMed
RefShare[3]	Ivan Burke, CSIR	Free	Yes, still only prototype	All, web-based	BibTeX	BibTeX
refBase	Reference Base Developers	free	Yes, GPL License	All	BibTeX, BiblX, MODS XML, RIS	BibTeX, Compac, ISI, BiblX, RIS,
RefWorks	RefWorks	US$100 per year	No	All, web-based	BibTeX and RIS	BibTeX (partially), CSA, BiblX, ISI, Medline, Ovid, PubMed, RIS, SciFinder, MODS XML

Table 1 – Comparison of Reference management tools

Existing Tool evaluation

As of the writing of this paper none of the tools identified have support for collaborative reference management. (Except for RefShare, the prototype model on which this paper's research is based). It should be noted that Google Scholar does provide some form of collaborative referencing mechanisms; though they are mostly controlled by the publishers of the articles (Hence lesser known articles and most proceedings and unofficial documents are not available). The only truly open means of sharing references was to generate an output file that share support for both user's reference management systems and then import it. This is a tedious and time consuming task and has to be repeated if references change or are added

Two tools were assessed as bases for the proposed prototype system: JabRef and RefWorks.

[2] Prices as of 15 October 2008.
[3] The prototype system developed during my research

These tools were chosen seeing as they are the most widely used tools within our organisation as well as the fact that one is open source (JabRef) and the other is commercial (RefWorks) so it is a good indication of the contrast in support for each.

JabRef

JabRef[4] is an open source bibliographical management system, developed by JabRef Developers. The system is licensed under the GPL[5] (GNU general Public License) agreement and the core of the system is based on the EndNote[6], data representation model which means the data model is proprietary and only the interactions with the data model is open for peer inspection and code reuse.

The system is Java based hence it is platform independent and is supported by all operating systems. Table 1 lists the file support. JabRef only supports Harvard method of referencing internally, yet the BibTeX files, which can be generated by JabRef, can be reformatted into any referencing standard provided you have the appropriate BST[7] file. JabRef can also list Reference collection in either HTML or plain text formats.

JabRef can be integrated into Microsoft Word by installing the Bibtex4Word add-on or it can generate an Open Office spread sheet that can be imported into an Open Office Base database. It however is fully integrated to work with Emacs, VIM text editor and WinEdt.

To help users input of new files the system has connectivity to third-party publishers and data warehouse such as ArXiV, CiteSeer, IEEE Xplore and PubMed, to help ensure integrity of data as well as make it easier for users to enter new references, yet this restricts users to published work and forces them to conform to third party mark-up of Bibliography.

RefWorks

RefWorks[8] was developed in 2001 and is a business unit of the ProQuest group. RefWorks is a commercial reference management system so it's proprietary and can not directly be used for open source applications.

What is really appealing of this product is the amount of organisational support it provides, from web seminars to interactive tutorials. It is a web based application that requires no software other than the browser. Its availability and mobility is of great concern to users that do not use fixed computers to compile their research, for example students.

RefWorks has an extensive support system for both input and output file formats as was shown in Table 1. As for reference styling this tool supports most with over 1,400 styles currently available[9]

RefWorks has no direct integration with any word processing software but it does allow saving of references as a RTF (Rich Text File) file that can be read by most word processors. RefWorks is only connected to PubMed for data verification but various other publishers sites link to RefWorks and allow you to enter the reference from the publisher's side.

By comparing these software tool insight was gained into what was available with current reference management systems and what was needed to make the process more collaborative

Advances in collaboration efforts in organisations

Collaboration software such as version control systems, web logs and wikis have aided in spreading information and collaboration among authors of information. These tools all have the following traits in common:

- Accountability
- Mutual trust
- Logging of data alteration (Audit trails)

Collaboration is a group process hence it is important to first look at the group aspect of collaboration before investigating the effects of collaboration.

Spence, Muneera (2005) identified the following activities as being the basis of group interaction

[4] Jabref is available at www.jabref.com
[5] Full GPL license can be viewed at www.gnu.com/copyleft/gpl.html
[6] EndNote is a commercial reference management system developed by Thomson Corporation in 1988.
[7] BibTeX Style file
[8] Refworks is available at www.refworks.com
[9] Full list of supported files can be found at http://www.refworks.com/content/products/output_style.asp

- Look for common ground – attempt to achieve consensus regarding group member's values, perceptions and ideals. Compromise and willingness to share views and opinions is essential to ensure all views are equally represented
- Learn about others – freedom of expression and openness to consider the views and opinions of other without prejudice
- Critique the results, not the people – Criticism is a great detergent of collaboration. Judge the results of collaboration based on lessons learned and try to avoid assigning blame due to unfavourable results
- Give and get respect – Be responsible when collaborating. Remember that your actions reflect on the group and similarly theirs reflect on you. (a further look into the types and means of acquiring respect is discussed later in this paper)
- Proceed slowly – Understanding of collaborative reasoning and consensus is important. Hence it is best to tackle singular concepts at a time as far as possible. This allows users to focus one concept at a time and reduce uncertainty.
- Be explicit and clear – Pay special attention to convey the explicit knowledge not just the tacit knowledge. Ensure each member grasps your opinion fully to ensure your views are not misconstrued
- Consider the five "Cs" of effective communication – Clarity, Completeness, Conciseness, Correctness and concreteness.

According to Spence et al's (2005) research, communication consists out of fifty-two percent is based on body language, thirty-seven percent on tone of voice and only eleven percent is based on the actual words that are communicated. He recommends that the listener focus his mind on the speaker, use body language to indicate interest in respect towards the speaker and visually reflect on the words of the user. This will help facilitate a sense of mutual respect between speaker and listener.

A reoccurring issue in all collaborative tools is the premise of mutual respect and trust. It is a critical aspect of collaboration. Respect can be given or received in one of two ways. Either by giving all new contributors the benefit of the doubt and respecting them fully until the violate this respect or by granting respect and trust based on prior actions, hence no respect given until the individual has proven themselves worthy of respect.

This brings up the issue of data integrity and collaboration. Can a system that is maintained by various users be integrally correct? Users are reliant on the proper conduct of fellow collaborators to try and achieve integrity. Wierzbicki (1982) has presented numerous papers on fairness and trust on network resource allocation and fair trade based on reputation models. Similar tactics are applicable on collaboration.

Consensus is the greatest barrier in any collaboration process and should be the main focus when trying to work collaboratively. Consensus is hampered by cultural (organisational and personal) boundaries. The following are examples of cultural boundaries that hamper collaboration

- Leadership culture – Employees are taught to following the orders of superiors and team leaders and not the advice of the most qualified individual. Following leadership is not collaboration it is submission
- Fear of the unknown. – People are unwilling to learn from other people they might not know or trust.
- Knowledge is power – Employees feel that by sharing their knowledge they are losing power over the information and hence losing value
- "Not invented here" – A common problem is that people do not wish to use the products or advice of rivals or external community as it is seen as unknown and not applicable to their situation.

Ultimately a measure of effectiveness and performance analyses is needed to assess the value of the collaboration efforts of the organisation.

A study conducted by Briggs, de Vreede and Dean (2007) lead to the identification of six common collaboration patterns.

- Generate – This entails the process of developing new concepts based on prior concepts or as a result of new knowledge gained.
- Clarification – The attempt to achieve a unified meaning or understanding of concepts to insure consensus about the meaning of concepts.
- Reduce – Reducing the set of generated concepts to a more specified and deterministic set of concepts that are relevant to the current situation
- Organize – Try and correlate relationships and dependencies among the concepts that were identified.
- Evaluate – Determine the worth of the concept as it is related to the current problem.
- Build Consensus – Strive to achieve consensus regarding the value of the concept relating to the problem in order to reduce disagreement about actions needed to achieve goals

Den Hengst et al (2007) investigated the reusability of collaboration efforts within an organisation and identified several constructs of reusable collaboration (three main constructs and ten sub constructs)

Reusability

Portability – Numerous collaboration utilities are available yet it is essential that collaboration be portable between these utilities to enhance reuse of collaboration efforts
Adaptability – Collaboration utilities need to be agile enough to adapt to changes in collaboration process with minimal reconfiguration or restructuring of tool.
Specification – The specification of collaboration utility should focus on the way collaboration is achieved and not on how the collaboration utility works internally.

Predictability

Difference in input – The group composition is dynamic hence the utility requires some measures to allow for consistent results and support regardless of inputs given
Difference in output – Similarly the results of these utilities may vary but the process to achieve these results needs to be constant
Reliability – This implies that if given the same input the utility should generally reproduce the same results
Robustness – But seeing as practitioners view of concepts are ever changing and their needs continue to evolve the utility needs to be robust enough to handle these changes.

Transferability

Perceptual load – The utility needs to be user friendly and easy to use to ensure use and reuse of utility.
Conceptual load – This is the amount of understanding needed to effectively use the utility to its full potential. Lowering the amount of understanding needed to use the utility will result in greater reuse and popularity of utility
Access load – The amount of access needed to use the utility. The utility will not be greatly reused or mastered if access to its features is cumbersome and inefficient.

Prototype

Introduction to prototype system

Based on the abovementioned research we developed a tool to evaluate our findings. The tool was a simple web based reference sharing tool that allowed users to add collections, references and groupings of references and collections. The prototype tool is currently deployed at Council for Scientific and Industrial Research's Mathematical and Computational Modelling and research group.

The tool is just a prototype at present and is continually being added to, based on new requirements and user needs as they arise. The current prototype is the third generation of the system. This version has been stable for the last two months and this is an account of its design and information learned from its deployment

Requirements of system

A tool was needed that:

- Could import data from existing reference database systems with minimal effort from user.
- Could ease the data entry process of reference data
- Could facilitate collaboration by making references easily accessible/ modifiable and shareable
- The system would be as unrestrictive as possible regarding tools required for use and user's computer skills or knowledge of reference management systems
- Aims to be open source compatible and compliant

Union of Collaboration and Reference Management

As of yet none of the tools investigated showed signs of collaboration. Sharing was achievable but cumbersome (through sharing of reference output files).

Other attempts to manage the efforts of collaborative document editing and reporting have however been done. Henningsen (2007) wrote a report on a variety of collaborative tools that are used within his organisation to assist in collaborative research.

He goes into detail of the conventional means of collaborating in sequential document editing via email and how it is not feasible or effective if documents are being altered at the same time.

Henningsen runs a Subversion[10] system within his department to assist in collaborative reporting. Subversion is an open source version control system that allows users to work on the same document at the same time and after users is done with their modifications they can submit it to the system the system will try and resolve all changes in data that do not overlap. Overlapping changes are flagged and presented to the user to decide which version of data to use. A great feature of Subversion is that users can log the reasoning behind the changes and this can help later users decide whether or not to overwrite the information.

This again brings up the issue of trust: how does one prevent users from altering your data? This tends to not be a problem with organisational Subversion seeing as employees are supposed to harbour mutual respect and trust. The same can not be said for anonymous open source usage (as was one of the goals of the prototype).

Another commonly used means of trying to share and authenticate reference data is to establish bibliographical wikis. This option is widely used in Information Science systems and Library Science.

The popularity and ease of use of wikis have made it a preferred tool for librarians. The only critique for this system is that wikis are not designed with the aim of bibliographical referencing which may lead to implementation issues such as restricting the wiki to only use a specific style of referencing or not having enough support staff to ensure data integrity.

Implementation of the Tool

The tool currently allows BibTeX references to be entered into the system seeing as it was identified as being the most commonly used representation format for all the existing reference database tools.. One key issue identified with the system requirement is the ease with which new records can be added. This issue was alleviated by the use of Asynchronous JavaScript and XML's (AJAX) auto-complete feature. This feature allows users to provide users a suggestive popup box that provides users with suggestions based on prior entered data in database as well as data that was entered in other fields of the current reference. Filtering occurs on both these sets of data to help suggest the values to users. Further work on filtering could lead to better suggestion results because over filtering can lead to empty sets of suggestions due to minor differences in data. The current implementation only filters within a record based on four main data fields: Author, Book Title, Journal and Publisher seeing as this selection of filters generated the greatest ratio of useful suggestions without over filtering the result set

Originally Soundex libraries was used to help reduce the common spelling mistakes by using phonetically index scoring to assist with word matches and increase the ease of use. Though this feature was later discarded seeing as it was predominantly language specific and yielded more noise data when compared to the AJAX auto-complete/suggestion filter. View Appendix A for test results

There are facilities in the prototype to generate new collections or import BibTeX files into collections that already exist. It is also possible to add: users, user collections or single references from other users to your collection by means of search and brows facilities that were incorporated into the system. This was of use to people that had no former knowledge of reference management systems

The prototype allows users to modify and maintain each others references without restriction. This is simply because the system is a closed system at present used among colleagues within a confined space and with organisational means of communicating their alterations to the references. This is an issue that needs to be addressed in an open system. Currently the system logs all alterations and who made them and has an administrative tool that can undo changes to the database. But for open systems a mechanism of trust and user validation is needed to insure data integrity is maintainable collaboratively.

A simple solution to this could be deploying a version control system that can handle simultaneous data alteration and updates. A good example of such a system is Subversion. These tools usually provide the users an option to add commentary before committing data to the system.

Discussion boards and wikis allow users to discuss their alteration and voice their grievances with regard to the information content. This could be the better option seeing as this would allow user to discuss the actual content of the reference not just the alteration there of.

Potential for enhancement of prototype

[10] Subversion is available from http://subversion.tigris.org/

Reference validation – The current data entry approach can be assisted by synchronizing the system with references in existing research databases like ACM[11], LISA[12] or Science Direct[13] and hence increasing the size of the known references allowing for more meaningful suggestions as well as providing a means of validating data input

Entering existing reference collection – As previously stated the prototype system is able to accept BibTeX formatted data as input to reference collections. An issue identified by the prototype using this means of data entry is that frequent mismatches and false positives of duplicate data occur.

The system currently logs all algorithm based logic that is applied during duplicate detection, the issue is how to handle these detections. Notification and manual correction proved tedious and repetitive for large collections and false negatives tend to slip pass algorithms that are currently used. But by narrowing match criteria the number of false positives greatly increases. The typical root cause of this problem is due to papers that form unions with each other and hence their data fields overlap to such an extent that it seems to be a duplication of data or where data is so sparse that no real judgement on data can occur. (For the sake of this argument erroneous data will be ignored seeing as even with testing of fields for possible candidates matches can't be assured, hence the need for a third party service to validate input)

Example:

[14]@INPROCEEDINGS (sqlrevoke,
AUTHOR = {van Staden, Wynand and Martin S Olivier},
TITLE = {{SQL}'s Revoke with a View on Privacy},
BOOKTITLE = {Proceedings of SAICSIT 2007 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists},
EDITOR = {Lynette Barnard and Reinhardt A Botha},
ADDRESS = {Fish River, South Africa},
MONTH = {October},
YEAR = {2007},
PAGES = {181--188},
NOTE = {(Published electronically)}
)

[15]@inproceedings{1292512,
author = {Wynand JC van Staden and Martin S Olivier},
title = {SQL's revoke with a view on privacy},
booktitle = {SAICSIT '07: Proceedings of the 2007 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries},
year = {2007},
isbn = {978-1-59593-775-9},
pages = {181--188},
location = {Port Elizabeth, South Africa},
doi = {http://doi.acm.org/10.1145/1292491.1292512},
publisher = {ACM},
address = {New York, NY, USA},
}

Here are two examples of BibTeX data. Both these mark-ups describe the same reference (the first was taken from the author's webpage the latter from the ACM archive) yet most of the data contained in the mark-up differs. It would be nearly impossible to determine these two records are related without cognitive (human) intervention. This example is a simplified example, were both parties had seemingly trustworthy knowledge the document being referenced. This problem escalates if data is entered partially or erroneously. This leads way to a possible new research topic: How to determine data relation and dependency?

A variation of this problem is the union or overlap of references. For example: if two references from the same publication is logged in the system there needs to be some method of linking these references so that if correlative data is update in one reference it should preferably be propagated to related references. Otherwise data will start to diverge and multiple variations of the same reference would be logged in database with no means of a global correction of data. It is a key concern that data integrity should be achievable and maintainable. But in order to do this meta-data about data is required or some form of flexible relational database

[11] Association for Computing Machinery – www.**acm**.org
[12] Library and Information Science Abstracts – http://www.csa.com/factsheets/supplements/lisa.php
[13]Science Direct – www.**sciencedirect**.com
[14] Recovered from Author's webpage mo.co.za
[15] Recovered from Publishers website www.acm.org

design (a possible Object Orientated Database that uses machine learning to try and identify correlation of references)

Tracking of changes to references and locking of references**.** – As a collaborative tool it stands to reason that various users can share, update and modify each others resources. There arises an issue of trust and mutual respect which needs to be addressed to insure data integrity is maintained. The prototype is a closed system that is only used by a select few individuals that work in a fixed location and the system is immune to outsider data corruption, all changes to data is logged and easily reversible (it is even open for discussion by means of groupware software within the department).

This might be the case for the prototype but in a real world solution data integrity is a real issue with numerous malicious attacks on web based data as well as no clear means of communication between users of the system to discuss their reasoning for altering data or verify correctness of data.

One possible interim solution for this could be to validate data via a third party authority. But as seen previously even trusted third parties can disagree regarding the validity of data. Hence a mechanism is required

Analysis of prototype based on Den Hengst et al's (2007) reusability constraints.

Reusability	
Portability	The prototype accepts BibTeX input and generates BibTeX output. By doing this RefShare aims to be as portable as possible seeing as BibTeX is the greatest common denominator of reference management support tools.
Adaptability	Refshare's current implementation uses n-tier and MVC (Model View Control) to achieve separation of concern so it is fairly simple to change implementation but as of yet it still requires actual coding so it is not very adaptive
Specification	Specification of system focused on the way references were to be shared not how this process should be implemented
Predictability	
Different input	This is one of the main issues of the system. How to mitigate a variety of input sources and standards. The main concern is that users have different views on what data should be capture as well as the meaning behind the data (meta data). Research needs to be done on how to form consensus and maintain data integrity of system.
Different output	Current system has yielded no results in this regard
Reliability	Data integrity is currently not validated hence results are unreliable
Robustness	System is not really robust due to non-specified framework of mitigation of reference alteration
Transferability	
Conceptual	The implementation is plain text based and requires no prior knowledge of BibTeX or any other bibliographical storage medium.
Perceptual	Ease of use and simplicity was emphasised during the development of the prototype
Access	Currently the prototype is only available within a closed space and is not widely accessible but seeing as this is only a prototype it is reasonable for it to not be widely accessible

Conclusion and further research opportunities

Reference management and collaboration tools have become a critical part of collaborative reporting and research.

The prototype system has identified several opportunities for improvement but judging by positive feedback it is a feasible tool for future use and implementation. Mainly this paper has set the groundwork for further research. Two main research topics that arose from this research and prototype testing:

- How to insure data integrity of resources as well as how to track and consolidate changes to the references in a collaborative environment?
- How to determine whether two references are the same or share an overlapping trait

It might also be worth investigating a tool that could be used for collaborative reporting and research that encapsulates the entire research process (fact finding, testing, discussing and reporting of findings) rather than just the bibliographical of the research

References

[1] Bennis, W. G., & Biederman, P. W. (1997), Organizing genius: the secrets of creative collaboration. Don Mills, Canada: Addeison-Wesley

[2] Briggs, R.O.,& de Vreede G., & Dean, D.L, "Collaboration patterns", (2007) Proceedings of the 38th Hawaii International Conference on System Sciences, Hawaii, IEEE Computer Society Press, under review

[3] Brown, J. S., & Collins, A. (1989). Situated cognition and the culture of learning (vol. 18 No.1) Massachusetts, America: American Educational Research Foundation

[4] De Vreede, G.J. (1996). "Participative modelling for understanding: facilitating organizational change with GSS", Proceedings from "29th Hawaiian International Conference on System Sciences", Hawaii: IEEE Computer Society Press

[5] Den Hengst, M, et al, (2007) Assessing the Quality of Collaborative Processes, Proceedings from "29th Hawaiian International Conference on System Sciences", Hawaii: IEEE Computer Society Press

[6] Elsvier. (2008). Librarian factsheet. Retrieved November 27, 2008, from 2collab Online: http://info.sciencedirect.com/docs/2collab/2collab_lib_factsheet_web.pdf

[7] Henningsen, A. (2007). Tools for Collaborative Writing of Scientific LaTeX Documents. The PracTEX Journal, 3 (1),

[8] Houston Public Library. (2001). How to Use Soundex/Miracode for United States Census Retrieved November 5, 2008, from Houston Public Library, Online: http://www.hpl.lib.tx.us/clayton/soundex.html

[9] Katzenbach, Jon R., and Douglas K. Smith. (2003). The Wisdom of Teams. New York, NY: HarperCollins,

[10] Olivier, M. (1997). {Proceedings of SAICSIT 2007 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists.Available: mo.ac.za. Last accessed 21 July 2008.

[11] Patashnik, & Lamport, L. (1997, 2006). BibTeX.org - Your BibTeX resource. Retrieved 21 November 2008, 2008, from www.bibtex.org

[12] Spence, Muneera U., (2005). Graphic Design Collaborative Processes: a Course in Collaboration. Oregon State University: Philadelphia, Pennsylvania

[13] Thomson Corporation. (2006, 2008). About Us. Retrieved November 27, 2008, from EndNote Online: http://www.endnote.com

[14] Thomson Corporation. (2006, 2008). Product Information. Retrieved November 6, 2008, from Procite Online: http://www.procite.com/pcinfo.asp

[15] Vogel, D.R. and G.J. de Vreede, (1999), "The organizational application of GSS: a comparison across time and national boundaries", Proceedings of the 10th EuroGDSS Workshop, Delft, the Netherlands, Faculty of Technology, Policy and Management

[16] Wierzbicki A. P., (1982) "A mathematical basis for satisfying decision making", Math. Modell., vol. 3, pp. 391–405

[17] Wikipedia contributors. (2007, 2008). Comparison of reference management software. Retrieved 10 November 2008, 2008, from http://en.wikipedia.org/wiki/Comparison_of_reference_management_software#cite_note-Reformat_RTF-10

[18] Wikipedia contributors. (2003, 2008). Collaboration. Retrieved 14 November 2008, 2008, from http://en.wikipedia.org/wiki/Collaboration.

Appendix A
Comparison between Soundex Libraries and Auto-complete.

Test parameters

- Collection of 758 references was used. (Total references in system as of 3 November 2008).
- Both techniques were used to find references that were authored by Herman le Roux. (Only the word "Le Roux" was used as filter term in results)
- Misspelled words are omitted by this test seeing as auto-complete will in all case of a misspelled word result in zero hits. (due to the nature of the algorithm)

Soundex Results:

	Suggestions	Hit Count
L	98	17
Le	98	17
Le R	41	17
Le Ro	41	17
Le Rou	41	17
Le Roux	17	17

Auto-complete results:

	Suggestions	Hit Count
L	98	17
Le	46	17
Le R	33	17
Le Ro	21	17
Le Rou	17	17
Le Roux	17	17

Comparison:

Results analysis

In general the Soundex libraries return lager sets of suggestions that lead to less filtering of data, which would hamper the ease of data entry. Yet it has a probability of achieving hits when words are misspelled, which auto-complete does not. A combination of the two algorithms would lead to the Soundex results (since in all cases Soundex will deliver the same or more results than auto-complete would). This is not the desired result because of the amount of noise data Soundex enters into the suggestion pool, we propose a switch mechanism between the two methods, which implies auto-complete is used until no results are found for its approach (usually when word is misspelled) and then the algorithm should switch to Soundex. This seems to yields the best results of algorithms that were tested and it would be easy to chain other algorithms in this fashion.

Biographies:

Ivan Daniel Burke received his BIT degree at University of Pretoria in 2008. He is currently busy with his Msc (CS) at the University of Pretoria while working full time for the Defence Peace Safety and Security Department at the Council for Scientific and Industrial Research. His interests are trust and fairness in online based applications.
Email: iburke@csir.co.za
Defence Peace Safety and Security (CCIW)
Council for Scientific and Industrial Research, South Africa

HERMAN LE ROUX has been with the South African Council for Scientific and Industrial Research since April 1998 and is at present a Principal Engineer in the Mathematical and Computational Modelling Research Group. He is involved in Modelling and Simulation-based Acquisition Decision Support, specifically for the South African National Defence Force. Interests include command & control information systems, information fusion, biometrics, artificial intelligence and software engineering. Le Roux completed a Masters Degree in Computer Engineering at the University of Pretoria in 1999 and is currently pursuing a PhD in Command and Control Modelling.
Email: WHleRoux@CSIR.co.za
Defence Peace Safety and Security (CCIW)
Council for Scientific and Industrial Research, South Africa