

PRIVACY-ENHANCED PUBLIC NAME-AUTHORITY SYSTEM  
FOR BUILDING RESEARCH COMMUNITIES

by

Jaehyun Paek

Submitted in partial fulfillment of the  
requirements for the degree of  
Master of Computer Science

at

Dalhousie University  
Halifax, Nova Scotia  
February 2011

© Copyright by Jaehyun Paek, 2011

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “PRIVACY-ENHANCED PUBLIC NAME-AUTHORITY SYSTEM FOR BUILDING RESEARCH COMMUNITIES” by Jaehyun Paek in partial fulfillment of the requirements for the degree of Master of Computer Science.

Dated: February 28, 2011

Supervisor:

---

Reader:

---

---

DALHOUSIE UNIVERSITY

DATE: February 28, 2011

AUTHOR: Jaehyun Paek

TITLE: PRIVACY-ENHANCED PUBLIC NAME-AUTHORITY SYSTEM  
FOR BUILDING RESEARCH COMMUNITIES

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: M.C.Sc.

CONVOCATION: May

YEAR: 2011

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

---

Signature of Author

# Table of Contents

<b>List of Figures</b> . . . . .	<b>vi</b>
<b>Abstract</b> . . . . .	<b>viii</b>
<b>List of Abbreviations and Symbols Used</b> . . . . .	<b>ix</b>
<b>Acknowledgements</b> . . . . .	<b>xii</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
<b>Chapter 2 The Federated World Directory of Mathematicians</b> . .	<b>5</b>
2.1 Background . . . . .	5
2.2 Design of Federated Searching . . . . .	7
2.3 Directories . . . . .	8
2.4 Federated searching in the FWDM . . . . .	9
2.5 An Analysis of the Strengths and Weaknesses of the FWDM . . . . .	11
<b>Chapter 3 Name Disambiguation Techniques</b> . . . . .	<b>14</b>
3.1 Clustering-algorithm based Named Disambiguation of Personal Data on the Web . . . . .	14
3.2 User-assisted Name-disambiguation of Personal Data on the Web . .	21
<b>Chapter 4 Designing a Name-Authority System</b> . . . . .	<b>24</b>
4.1 Name Disambiguation . . . . .	24
4.1.1 Name Disambiguation in a Name-Authority System . . . . .	26
4.2 Data Control . . . . .	27

4.2.1	Data Controlling in a Name-Authority System . . . . .	32
4.3	Other Design Requirements for a Name-Authority System . . . . .	33
4.4	Design Alternatives for a Name-Authority System . . . . .	33
4.4.1	Data-filtered Name-Authority System . . . . .	34
4.4.2	User-filtered Name-Authority System . . . . .	35
<b>Chapter 5</b>	<b>MathPeople - The Data-filtered Name-Authority System</b>	<b>37</b>
5.1	System Overview . . . . .	37
5.2	User Management . . . . .	39
5.3	Data Contribution Management . . . . .	40
5.3.1	Managing Contributions from Authenticated Users . . . . .	40
5.3.2	Managing Contributions from Unauthenticated Users . . . . .	43
5.4	User Interface - Search and Profile Pages in MathPeople . . . . .	44
5.5	Discussion . . . . .	46
<b>Chapter 6</b>	<b>Populi Scientiae - The User-filtered Name-Authority Sys-</b>	
	<b>tem</b> . . . . .	<b>49</b>
6.1	System Overview . . . . .	49
6.2	User Management . . . . .	50
6.2.1	Account Creation . . . . .	51
6.2.2	Trusted Users and Confidence Level . . . . .	52
6.3	Data Management . . . . .	53
6.4	User Interface - Search and Profile Pages in Populi Scientiae . . . . .	55
6.5	Discussion . . . . .	57
<b>Chapter 7</b>	<b>Conclusion and Future Works</b> . . . . .	<b>59</b>
7.1	Future Work . . . . .	60

**Bibliography . . . . . 62**

## List of Figures

Figure 2.2	List of Membership Directories included in the FWDM . . . . .	10
Figure 2.3	FWDM result page . . . . .	11
Figure 4.1	Principles of the Canadian Act . . . . .	30
Figure 4.2	Design Diagram for the Data-filtered Name-Authority System	34
Figure 4.3	Design Diagram for the User-filtered Name-Authority System	35
Figure 5.3	MathPeople: Bookmarklet . . . . .	43
Figure 5.4	Search engine translation of diacritics . . . . .	45
Figure 6.3	Populi Scientiae Sign-up Page . . . . .	52
Figure 6.4	Populi Scientiae: Successful Registration of an User . . . . .	52
Figure 6.5	Populi Scientiae: Editing Trusted user list . . . . .	53

## Abstract

Today, the Internet has become an important source of information about academic researchers and their research activities. The types of information one can obtain from the Internet include contact information, publication information, biographical information, photographs, and other miscellaneous information. Some of this information is generated by professional societies and academic institutions, while other information is generated by individuals and independent enterprises.

As the quantity of academic material on the web grows, finding and processing information about a researcher's work is increasingly difficult. For example, it is often hard to discern whether authors of different papers in tangentially-related areas are the same person, based solely on a name. Even if one can determine this information, it is often difficult to assess the accuracy of information obtained, especially if it was generated either by an individual or by community of users.

In this thesis, we propose a novel community-based and web-accessible repository of information about academic researchers and their research activities. First, we introduce a web-application called *Federated World Directory of Mathematicians*(FWDM), which retrieves personal information from a variety of disparate data-sets, and which inspired the solutions proposed in this thesis. We then propose a *public name-authority system*, as a means to provide high quality disambiguated information on researchers. The proposed system helps to ensure the quality of information by obtaining only the information approved by the research community. We introduce and describe two approaches to the design of public name-authority systems - the data-filtered and the user-filtered name authority systems - in order to explore their benefits and drawbacks.



## List of Abbreviations and Symbols Used

<b>A/CDC</b>	The agglomerative/conglomerative double clustering
<b>ACM</b>	The Association of Computing Machine
<b>AMATYC</b>	The American Mathematical Association of Two-Year Colleges
<b>AMS</b>	The American Mathematical Society
<b>AP</b>	The dataset augmentation process
<b>AWM</b>	The Association for Women in Mathematics
<b>baseNP</b>	The base Noun-Phrase
<b>CDE</b>	The International Mathematical Union's Commission on Development and Exchange
<b>CEIC</b>	The International Mathematical Union's Committee on Electronic Information Communication
<b>CML</b>	Combined Membership List of various American and Canadian mathematical societies maintained by American Mathematical Society
<b>CMS</b>	The Canadian Mathematical Society
<b>df</b>	Document frequency
<b>dfidf</b>	Multiplication of document frequency and inverse document frequency factor
<b>EWDM</b>	The World Directory of Mathematicians

<b>FWDM</b>	The Federated World Directory of Mathematicians
<b>HAP</b>	The host-based dataset augmentation process
<b>HD</b>	The host-based detection of network motif
<b>ICM</b>	The International Congress of Mathematicians
<b>ICMI</b>	The International Mathematical Union's Commission on Development and Exchange
<b>idf</b>	Inverse document frequency
<b>IMU</b>	The International Mathematical Union
<b>LEAF</b>	The Linking and Exploring Authority Files
<b>MAA</b>	The Mathematical Association of America
<b>mi</b>	Mutual information
<b>MSN</b>	The webportal provided by Microsoft
<b>NCM</b>	The National Committee for Mathematics
<b>NER</b>	The Name Entity Recognition
<b>OECD</b>	The Organization for Economic Co-operation and Development
<b>PD</b>	The page-based detection of network motif

<b>PIPEDA</b>	The Personal Information Protection and Electronic Documents Act
<b>PNR</b>	The popular node removal in a network
<b>sf</b>	Server frequency
<b>sfidf</b>	Multiplication of server frequency and inverse document frequency
<b>SIAM</b>	The Society of Industrial and Applied Mathematics
<b>SMF</b>	The Mathematical Society of France
<b>TD-IDF</b>	Term frequency-inverse document frequency
<b>WDM</b>	The World Directory of Mathematicians

## Acknowledgements

First of all, I would like to thank all members of my supervising committee. I would like to thank Dr. Andrew Rau-Chaplin, my supervisor, for his guidance in completing this thesis. Also, I would like to thank Dr. Christian Blouin and Dr. Alex Brodsky for kindly volunteering to be the readers for the defense of my thesis.

Second, I would like to thank the following people for their guidance in the implementation of two prototype systems proposed in this thesis. I would like to thank Dr. Jonathan Borwein and Dr. James Pitman of University of California, Berkeley for their directions throughout my research. I would like to thank Dr. Michael Deturbide, the Associate Dean at Dalhousie Law School and Dr. John McHugh for their advice during the design phase of the systems. As well, I would like to thank Mason Macklem, Scott Wilson, and Hadley Wickham for their invaluable assistance in designing and implementing these systems. In addition, I would like thank Mason for his invaluable feedback.

Finally, I would like to thank Rob Hutten, Andrew Siffert, and all the friends at FlagstoneRE for their support.

# Chapter 1

## Introduction

Today, the Internet is an important source of information on academic researchers and their research activities. The types of information one can obtain from the Internet includes contact, publications, biographies, photographs, and other miscellaneous information. Some of these resources are generated by professional societies and academic institutions, while others are generated by individuals and independent businesses.

As the quantity of academic materials available on the web grows, identifying materials related to a given author has become a major challenge. It is often hard to discern whether authors of different papers are the same person based solely on their names. One type of difficulty in identifying an author arises from the author sharing their names with several other researchers in the same area of research. Difficulty may also arise from an author publishing papers under different names. An author may change his or her surname due to a change in marital status. As well, an author may have listed his initials in some publications, but not in others. An author may also assume a new name when migrating to another country, or for any other personal reasons. This problem is illustrated in Figure 1.

The difficulty is compounded when an author publishes papers in a number of different fields of research. In general, even if an author has been authoritatively identified on a website or a group of websites, it may not be trivial to figure out whether the author is the same person as one listed on different websites. For example, it may be difficult to determine whether a “John Smith” authoritatively identified

as a specific individual within a bibliographical database such as MathSciNet is the same “John Smith” listed in a public listing of researchers such as the American Mathematical Society’s Combined Membership List (CML). We refer to this problem as the *name disambiguation problem* [2, 14, 21, 11, 12, 23, 10, 15].

A second major challenge arising from the increase of web based data sources is how to ensure control of the personal information remains with the subject of the information. As anyone can post materials on the Internet, it is easy for individuals to disseminate inaccurate or defamatory materials. Such information can portray an incorrect view of someone’s research, and could potentially harm his/her reputation. In general, it may be difficult to discern the accuracy of material on the web without personal knowledge of the field, either because the material contain enough accurate information to appear authoritative, or because the inaccurate information is confirmed by a number of web pages on different websites. This is becoming a particularly serious problem as the number of community-based resources such as Wikipedia increases because they give a measure of credibility to incorrect information. We refer to this as the *data control problem* [4, 13, 18, 24, 17, 9, 1].

Current research efforts aimed at addressing the name-disambiguation problem can fit largely in two camps: solutions based on clustering algorithms, and solutions based on user-driven or user-assisted name disambiguation. Name-disambiguation systems based on clustering algorithms attempt to aggregate personal information from web resources based on some common links that are assumed to exist among different sources of personal information. User-assisted name-disambiguation incorporates input from user in order to aggregate data, the results of which are used as either the sole source of information or as a supplement to data obtained through clustering algorithm.

The benefit of using clustering algorithm for name-disambiguation system is that it can aggregate large data sets quickly and in an automated fashion. As well, since it

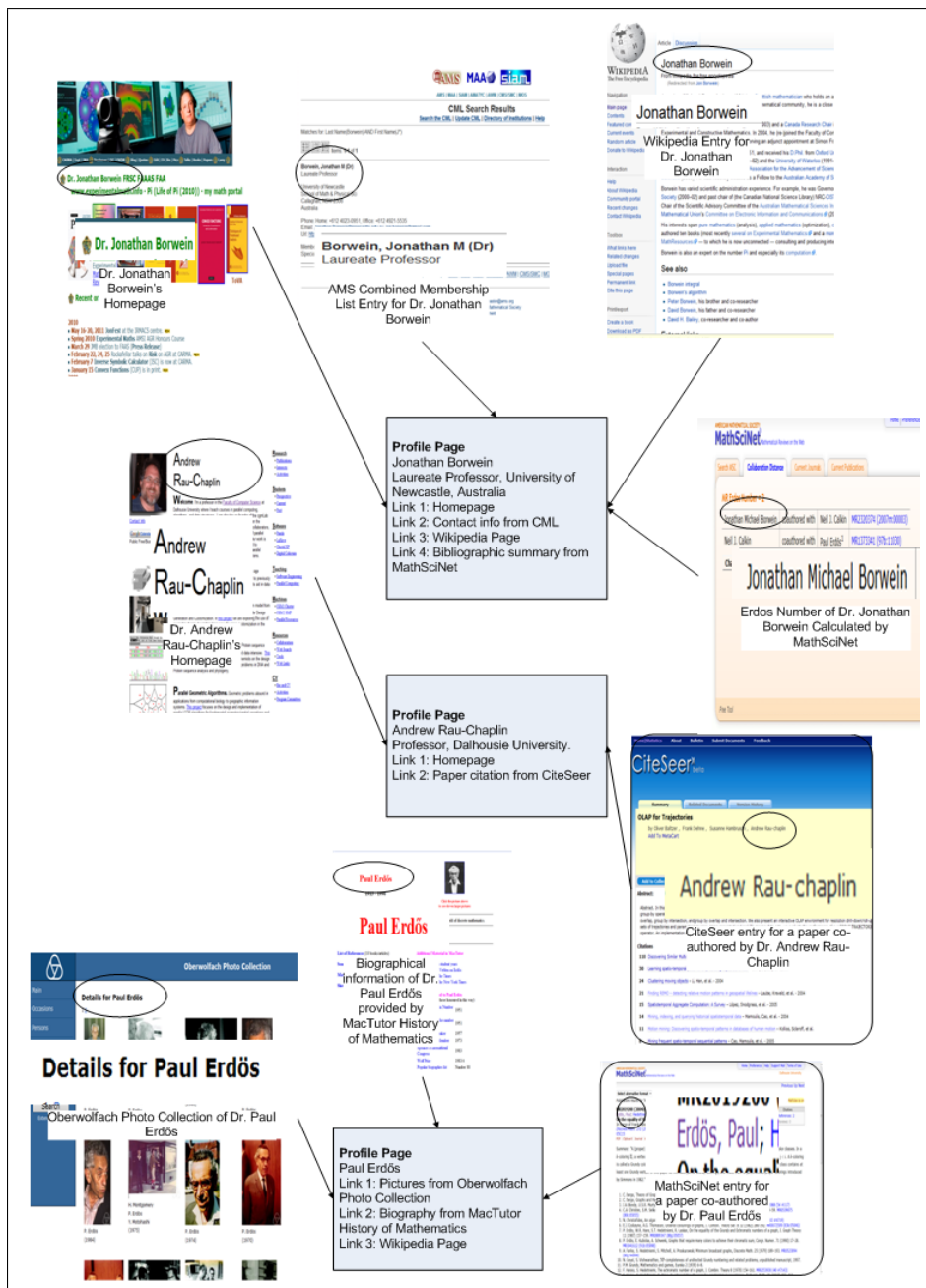


Figure 1.1: An example of name disambiguation of information on researchers through profile pages

is based on a deterministic algorithm, the way the aggregation is done is predictable and consistent. However, the problem with the clustering algorithm approach is that it relies on the existence of the common linkage among different sources of data, which often does not hold true in highly heterogeneous environments such as the web.

The benefit of a user-driven name-disambiguation system is in the fact that people can make connections between information coming from different sources using the context of the data and personal knowledge. One major challenge in using user input for aggregation is in determining which users should be permitted to contribute to aggregation. If the system allows any user to contribute to the system (i.e. *an open system*), it is difficult to ascertain the accuracy of the data; however, if the system only allows contributions from users who have obtained valid credentials from the system (i.e. *a closed system*), it would be difficult to maintain large and continually growing data sets due to the inherent limit on the number of people who can contribute.

Our previous effort to address the name-disambiguation problem and the data control problem in the context of information about academic researchers and their research activities resulted in the Federated World Directory of Mathematicians <sup>1</sup>. The Federated World Directory of Mathematicians was implemented using a real-time search paradigm called the Federated Searching with a Simple name-matching algorithm for name disambiguation. Although this research effort resulted in a system that addressed the data control problem, it did not adequately address the name disambiguation problem and exhibited performance issues.

In this thesis, we present a novel user-driven name-authority system to address both the name disambiguation and the data control problems. The name-authority system attempts to combine the benefits of both open and closed systems by allowing any user to contribute data to the system, but only presenting data to the public if it is approved by an authenticated user. The way a name-authority system addresses

---

<sup>1</sup>See <http://www.mathunion.org/fwdm/index.shtml>



name-disambiguation problem is through the construction of a profile page. A profile page is a web page representing authoritative information on a researcher. It contains references to web resources that provide information on either on a researcher or previous research efforts by a researcher. In addition, it may contain additional information to easily help identify the researcher discussed by a profile page, as illustrated in Figures 1 and 1.

To better explore the design trade-offs inherent in current name authority systems, we implemented two very different designs. In the first approach, called the *data-filtered name-authority system*, data is submitted to the system by both approved and unapproved users, which are then filtered and aggregated by the approved users before being published. In the second approach, called the *user-filtered name-authority system*, instead of allowing any user to contribute data, it requires the user desiring to contribute data to acquire approval from the users who have been authenticated by the system.

The remainder of thesis is composed of the following chapters. In Chapter 2, we present an in-depth study of the design and implementation of the Federated World Directory of Mathematicians, a federated search engine that provides contact information of mathematicians around the world, resulting from our previous research, in order to motivate the solution proposed in this thesis. In Chapter 3, we explore various solutions currently proposed in the literature to address the name-disambiguation problem. In Chapter 4, we outline our design proposal for the name-authority system in light of the name-disambiguation and the data-control problems. In Chapters 5 and 6, we examine the two prototype systems that implement our design proposals for name-authority systems. Finally, in Chapter 7, we summarize our work and outline future research directions.

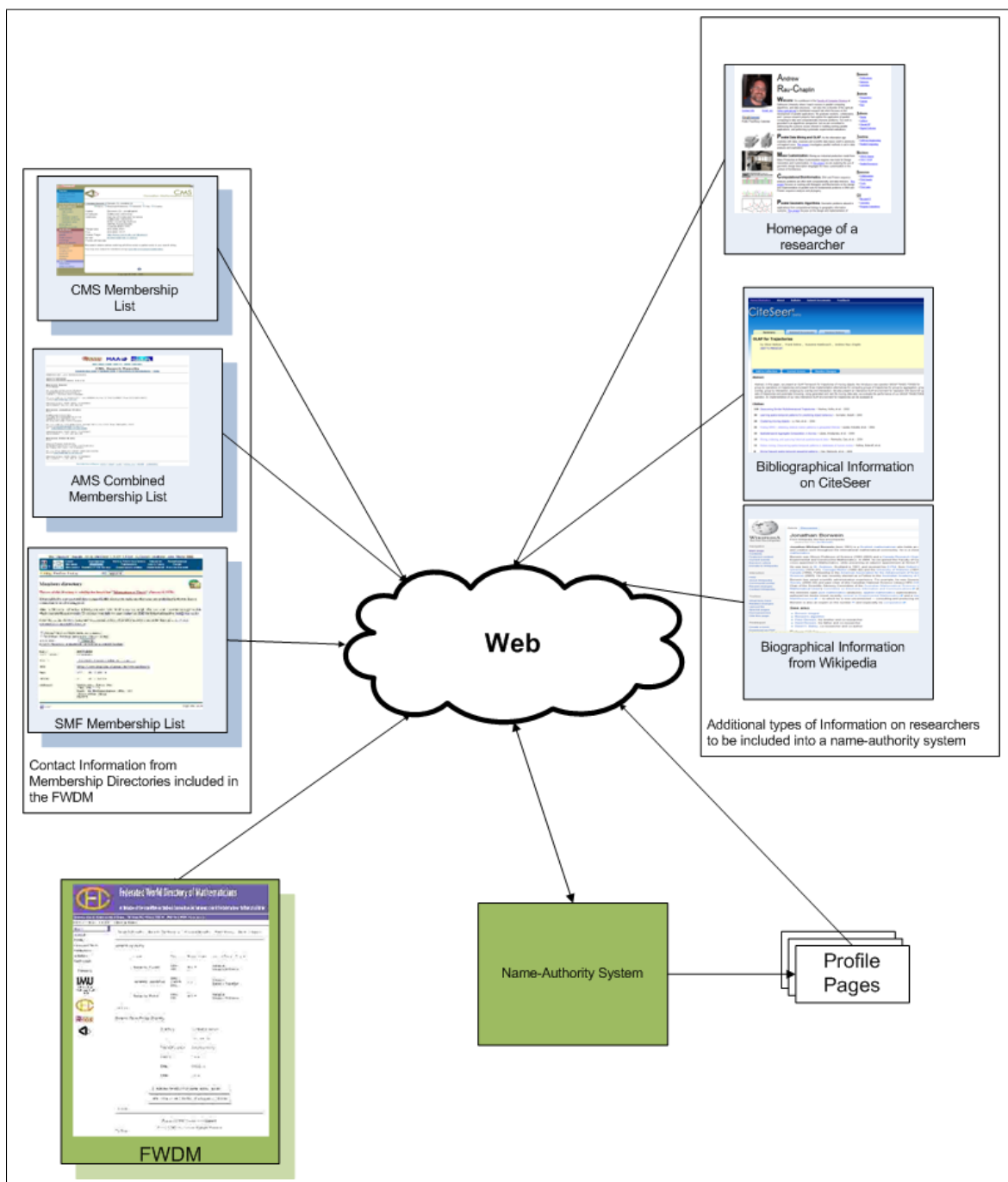


Figure 1.2: This diagram provides some example data sources for the FWDM and a Name-Authority System demonstrates how these two systems interact with these data sources. The FWDM can only retrieve data from the membership lists, which are exemplified by the data sources drawn with the shadow, but a Name-Authority System can retrieve data from any of the data sources shown in this diagram

## Chapter 2

### The Federated World Directory of Mathematicians

Throughout this thesis we use the international Mathematical community as a case study for our work on name disambiguation and data control techniques. Much of the research in this thesis was motivated by our earlier work in building the Federated Word Directory of Mathematicians (FWDM) <sup>1</sup>. The FWDM project, which was conducted for the International Mathematic Union’s Committee on Electronic Information Communication, tackles the problem of how to create researcher profile pages that combine data from multiple catalogues of researchers. In this chapter, we describe the FWDM in detail with a focus on federated searching and highlight the strengths and weaknesses of the approach taken in order to motivate the design of name-authority system, which is the main topic of this thesis.

In Section 2.1 we provide historical background on the FWDM; in Section 2.2, we describe the design of a generic federated searching; in Section 2.3, we review the online directories of various mathematical societies that were considered for the inclusion in the initial release of the FWDM; in Section 2.4, we describe the design methodology of the FWDM; finally, in Section 2.5, we provide our analysis of the lessons learnt in the design and the implementation of the FWDM.

---

<sup>1</sup>The FWDM project was supervised by Jonathan Borwein under the auspices of the International Mathematical Union’s Committee on Electronic Information Communication (CEIC). Jaehyun Paek acted as the leader developer/designer for the FWDM. The Technical Supervisor and co-developer of the FWDM was Mason Macklem. The inclusion of all of the national mathematical societies would not have been possible without the assistance of Peter Michor, Carlos Perpétuo, Eugenio Rocha, José Francisco Rodrigues, and Gerald Teschl, as well as all of the past contributions from the IMU and the CEIC.

## 2.1 Background

The International Mathematical Union (IMU) is a non-governmental and non-profit scientific organization whose mission is to promote international cooperation in mathematics. The membership of the IMU consists not of individual mathematicians, but of national mathematical societies or national academies of science, currently representing 66 nations that are mandated to uphold standards of mathematical research. As an umbrella body of various mathematical societies, the IMU fulfills its mandate through wide range of responsibilities, which include providing help to improve mathematical education in developing countries and sponsoring lectures and international meetings. To help meet its responsibilities, the IMU established a number of commissions. These commissions include the Commission on Development and Exchange (CDE), the International Commission on Mathematical Instruction (ICMI) and the Commission on Electronic Information and Communication (CEIC).

The most prominent of the activities organized by the IMU is the quadrennially-held International Congress of Mathematicians (ICM). The activities at the IMU include the presentations of the recent research on outstanding mathematical problems and the awarding of the Fields Medals and the Nevanlinna Prize. Another important responsibility of the IMU is the publication of the World Directory of Mathematicians (WDM), whose publication coincides with the ICM. The WDM, which aims to provide the contact and the affiliation information on all active research mathematicians throughout the world in a single hardcopy directory, is compiled through the collaboration between the IMU and the American Mathematical Society (AMS).

In 1998, the IMU, recognizing enormous costs and efforts required to update and publish the WDM every four year, asked the CEIC to investigate the feasibility of replacing the WDM with the Electronic World Directory of Mathematicians (EWDM), a version of the WDM that would be maintained digitally through a centralized

database and that can be searched and accessed over the internet. In the WDM, collecting of information on individual mathematicians for inclusion in the WDM is not performed directly by the IMU, but by each of the member societies. However, because these data are collected only to be published in the hard-copy version of the WDM, the CEIC concluded that it would require the IMU to obtain explicit permissions from each mathematician in order to satisfy the privacy laws of some countries, and given the logistics require to obtain these permissions, the CEIC concluded that it would not be feasible for the IMU to implement the EWDM as a replacement for the WDM.

As a way to present information contained in the WDM without obtaining explicit permissions, the CEIC later recommended exploring implementation of a federated search engine called the *Federated World Directory of Mathematicians*(FWDM). A federated searching is a search paradigm where the search engine does not maintain a centralized database of information, but retrieves information from heterogeneous datasets. In the FWDM, each member nation was to develop a publicly-accessible search engine that would search its membership database, whose information would be collected according to the national and the local privacy laws. These search engines, in turn, would be searched by a single user interface and results would be merged into a single set of search results. In 2004, the IMU endorsed moving ahead with a federated search protocol. The result of this endorsement was the implementation of the Federated World Directory of Mathematicians.

## **2.2 Design of Federated Searching**

At the heart of our design for the FWDM was a federated search engine. A federated search engine provides a consolidated search experience over multiple data sources by providing the end users with a single search interface.

In federated searching, the user query is executed in following six phases, which

is described in Figure ??.

1. The federated search engine collects the user query through the query interface.
2. The user query is parsed into a format that is recognizable to each search interface for a data source.
3. The federated search engine submits the parsed query to each of the search interfaces.
4. The federated search engine collects the search results from each of the data sources.
5. The federated search engine merges all the search results into a single combined result set using the disambiguation module.
6. The federated search engine presents the combined results to the user through the user interface.

The key challenge in federated searching is how to combine the results (step 5). In case of the FWDM, each search result consists of a list of individuals with their contact information. The merge step must solve the name disambiguation problem in order to group results pertaining to a single individual together.

### **2.3 Directories**

We have examined membership directories of various Mathematical societies for the inclusion into the FWDM as data sources. The suitability of the inclusion into the FWDM was determined by the following criteria:

1. The membership directory is either maintained or authorized by the IMU or one of its member societies.

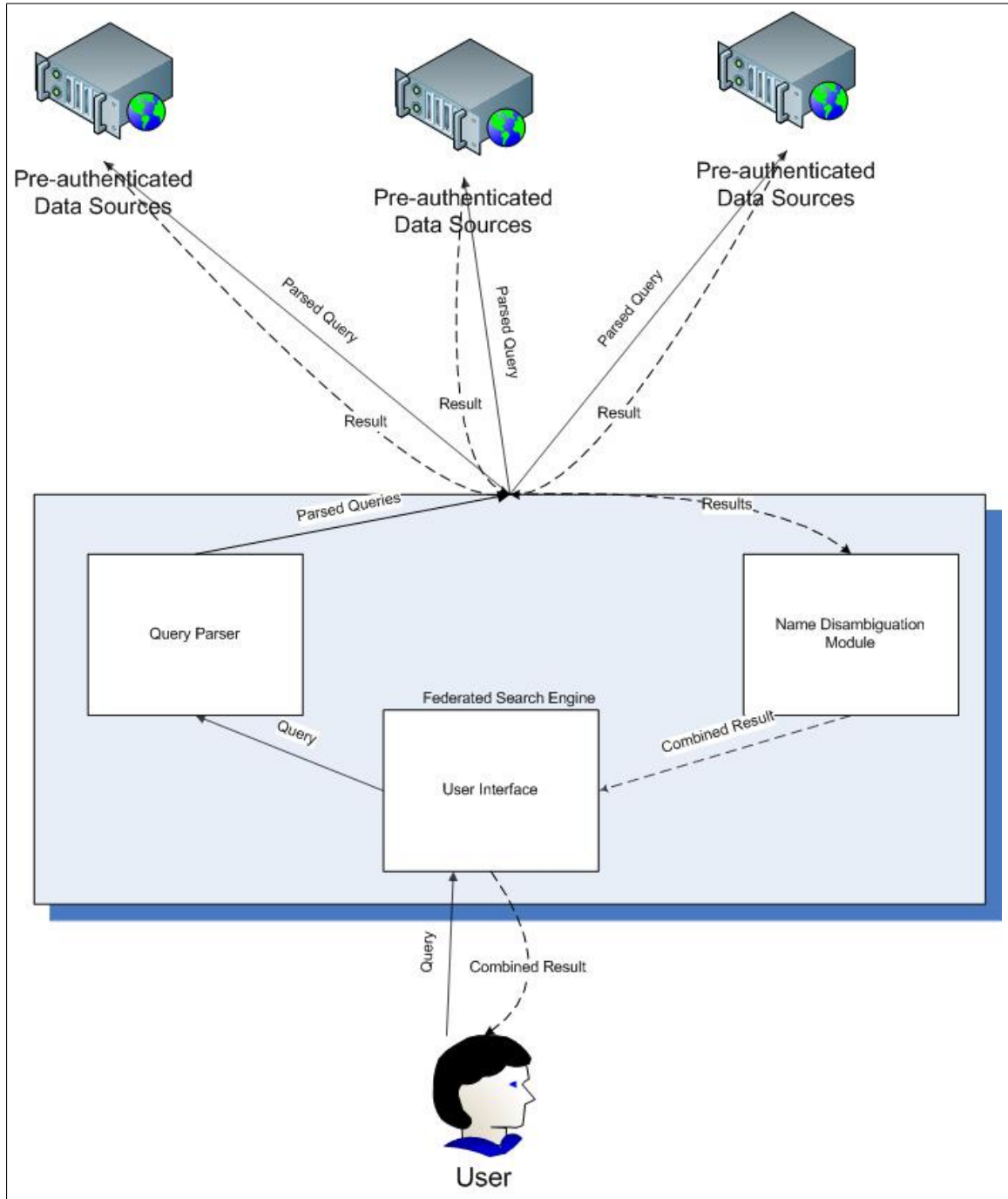


Figure 2.1: This diagram shows the query execution process in the search engine.

2. The membership directory is organized in such a way that it should be easy to find contact information for a researcher given the researcher's name. Most commonly, this can be achieved through a search interface based on the GET HTTP method that allows user to search for a contact information using the name of the Mathematician; however, it could also be achieved in the membership directory that does not provide a search interface if the listing of the single web page with page organized in such a way that the matching name can be easily found.
3. The membership directory should provide some contact information for each of the member listed. The contact information can be mailing address or phone number, but it could also be the name of the institution the member currently works at.
4. The output page from the membership directory must have clear delimitation among the returned entries, as well as each piece of information contained within an entry.
5. The membership directory should provide a link to a web page or a section of a web page where the information pertaining to the Mathematician searched can easily be found.

Our investigation revealed that many of the member societies maintained a membership directory meeting the criteria outlined above. The FWDM prototype included the membership directories shown in Figure 2.2.

#### **2.4 Federated searching in the FWDM**

The FWDM provides three different query interfaces to the users: a *Basic Search interface*, which consists of the first and last name input forms; a *Standard Search*



<b>The Combined Membership List (CML)</b>	The sole online repository containing the membership listings of the American Mathematical Society (AMS), the Mathematical Association of America (MAA), the Society of Industrial and Applied Mathematics (SIAM), the American Mathematical Association of Two-Year Colleges (AMATYC), and the Association for Women in Mathematics (AWM). In addition, it contains the membership listing of the Canadian Mathematical Society (CMS).
<b>The Membership Directory for the Canadian Mathematical Society (CMS)</b>	a directory of Canadian Mathematicians maintained by the CMS .
<b>PERSONA MATHEMATICA</b>	An online directory designed and driven by the Math-Net group of the Mathematical Institute / University of Cologne. This search engine provides the membership listings from more than 1000 mathematical websites in Germany and Austria, including the membership listing for the Deutsche Mathematiker-Vereinigung, the German National Mathematical Society.
<b>The Membership Directory for the Mathematical Society of France (SMF)</b>	a directory of French Mathematicians maintained by the Mathematical Society of France.
<b>The Membership Directory for the National Committee for Mathematics (NCM)</b>	an online listing of Australian mathematicians collected in 2001 for the inclusion in the WDM.
<b>The Electronic World Directory of Mathematicians (EWDM)</b>	An open online directory of mathematicians around the world maintained by the IMU.

Figure 2.2: This table provides a brief description on each of the membership directories included in the FWDM.

*interface*, which contains input forms for fields common to all membership directory search interfaces; and an *Advanced Search interface*, which contains input forms for all the fields that are available in any member society database. The three user interfaces provides interface for the most common search pattern through the Basic Search Interface, while the Standard and the Advanced Search Interface allow more refined search as needed

The search in the FWDM search engine is executed in following order:

1. Parse the search parameters entered by the user into a query string that conforms to the format required by each of the membership directory search engines.
2. Submit the query string to each of the search engines.
3. Collect the result pages from each of the membership directories. Parse the results to identify and separate each individual entry, and create a list of results for the given membership directories.
4. Once all membership directories have returned their results, combine the individual listings into a single list, removing duplicate entries.

The name disambiguation in the FWDM is a simple personal name matching. In the FWDM, if a personal name extracted matches a name found in another membership directory, these two results are combined. The assumption used in this name disambiguation is that a membership directory does not contain duplicate entries. The extraction of the name terminates once the desired number of names have been found, or if all the names found in a directory have been extracted.

In Figure 2.3, we show how the FWDM disambiguates contact information. To combine the contact information, we had used the name. Although matching email or matching home page might give greater confidence result, the fact that people maintain more than one email account or home page (as is the case in the example),

or that there might be difference in spelling (as is the case in the email), would have resulted information on same person not being merged. In addition, a number of entries were found to not contain this information. However, merge on name is inaccurate especially when dealing with common names (such as John Smith). For the prototype, it was determined that being loose and presenting a more abbreviated result page is more useful than being strict and having many un-disambiguated results.


Once the results have been consolidated through the name disambiguation process, the FWDM presents the result in paginated web pages. The result page of the FWDM is divided into following four sections:

- Search Results.
- CML Only Search Results.
- Search Results by Society.
- Google Search Results.
- Query Form

## **2.5 An Analysis of the Strengths and Weaknesses of the FWDM**

A key strength of the FWDM is the effective management of the personal information. As the information is obtained from membership directories of the national mathematical societies, there is reasonable expectation that each piece of contact information is obtained directly from the researchers with consent and that the personal information is released in accordance with the privacy laws of the country of the researcher's residence. Also, because the FWDM obtains the data in real-time, the information presented is guaranteed to be the most recent data the membership directory contains.

The FWDM also exhibited a number of significant weaknesses. These include



# Federated World Directory of Mathematicians

An Initiative of the Committee on Electronic Information and Communication of the International Mathematical Union

[International Mathematical Union](#) | [IMU on the Web](#) | [WDML](#) | [EWDM](#) | [FWDM](#) | [Contact](#)

[D-Drive Home](#) > [FWDM](#) > [Query Form](#)

---

[Home](#)

[General](#)

[People](#)





[News and Alerts](#)

[Publications](#)

[Activities](#)

[Further Info](#)

Partners:

[Search Results](#)   [Results by Society](#)   [Google Results](#)   [Your Query](#)   [Search Again](#)

---

*Search Results*

Name	Society	Home Pages	Link to Google Search
<b>Borwein, David</b>	CMS CML	NONE	<a href="#">Google Scholar</a>
<b>Borwein, Jonathan</b>	CMS EWDM CML	<a href="#">1</a> <a href="#">2</a>	<a href="#">Google Scholar</a>
<b>Borwein, Peter</b>	CMS CML	NONE	<a href="#">Google Scholar</a>

[To Top](#)

---

*Search Results by Society*

Directory	Number of results
Total:	3 names found
Total(CML-only):	0 names found
<b>EWDM:</b>	1 Results
<b>CML:</b>	3 Results
<b>CMS:</b>	3 Results

Directories timed out or contained no results  
**SMF / Math-Net / OeMG / Portuguese / NCMS**

[To Top](#)

---

Found 139,000 results from **Google**

Found 3,740 results from **Google Scholar**

[To Top](#)

Figure 2.3: The results returned from the FWDM when searched for last name Borwein

1. performance issues,
2. crude name disambiguation, and
3. inability to function with heterogeneous data.

A problem with the real-time searching of online data-sources is that the execution time is bound by the speed of the network connection and the execution time of the data source search interface, which can be unacceptably slow. Basically, it is a distributed system whose performance is bound by the performance of its slowest component. One way to address this problem is to maintain a centralized database where each constituent data source can easily contribute most up-to-date data. The challenge in maintaining a system based on a centralized database is how to provide same level of personal information control that is provided in the FWDM.

For name-disambiguation the FWDM uses a simple name matching algorithm, which is crude and prone to error. The FWDM disambiguates based on the first and last name of the researcher, ignoring middle initials. Such a disambiguation algorithm can result in high probability of match in case of uncommon names such as “Jonathan Borwein,” but it is unreliable when the data sources contain a large number of common names, such as “John Smith.” An alternative strategy might be to use an alternative field or combination of fields for disambiguation. For instance, two entries sharing common email address almost certainly pertain to same individual. However, as noted in Section 2.4, because, other than names, any of the membership directories can reliably assure the availability of other types of information, such method would result in a large number of false negatives. Clearly a much more sophisticated automatic name disambiguation method might be designed, but given the diversity of data, it seems likely that any automated method will be imperfect. This suggests it may be worthwhile to consider user-assisted methods.

Another limitation of the FWDM is that its data sources are mainly constrained to

contact information provided by membership lists of various Mathematical societies. Although it provides a single source for a rich set of highly trusted contact information, the usefulness of the FWDM is limited due to lack of other types of information on researchers, such as a bibliographic listing of past publication, an online encyclopedia containing biographical information on famous researchers, a home page, a collection of photographs, and a web directory containing recordings of lectures. One way the FWDM tried to alleviate this problem is through providing links to Google and Google Scholar search result pages. However, as these search engines do not provide disambiguated results, it puts onus on the users to discern which information belongs to the researcher in question.

One way to address both the name disambiguation problem and the inclusion of heterogeneous data is through user assistance. Because users can make imply connection between pieces of data that algorithms cannot recognize, the user contributions can greatly enhance the quality of the name disambiguation. The key challenge when incorporating user assistance is in determining who can contribute different types of input.

## Chapter 3

### Name Disambiguation Techniques

In this chapter, we provide overview of various techniques proposed in the literature to address the name-disambiguation problem. In Section 3.1, we present different cluster-algorithm based solutions that have been proposed by the research community. In Section 3.2, we examine various user-assisted name-disambiguation solutions proposed for web repositories of personal information.

#### 3.1 Clustering-algorithm based Named Disambiguation of Personal Data on the Web

Several researchers have investigated various clustering methods, the methods for grouping together entities having similar properties, to implement algorithm for disambiguating personal data made available through the web resources [2, 14, 21, 11, 12, 23]. The use of clustering algorithm to disambiguate personal entities stems from previous researches on disambiguating author entities from citations, such as [10] and [15]. The previous focus on author disambiguation was due to the wide availability of resources providing standard citation information, such as CiteSeer <sup>1</sup>. Although many of these resources does not employ sophisticated disambiguation algorithms suggested in these researches (for example, CiteSeer only provide name disambiguation based on simple name matching), a few digital libraries and online abstract and review services provide automated name disambiguation through clustering algorithm or manual matching with varying degree of accuracies. For example, when searching

---

<sup>1</sup><http://citeseer.ist.psu.edu/>

for an author on Web of Science <sup>2</sup>, it gave us an incomplete listing of publications, and a search for an author in Scopus <sup>3</sup>, an online abstract and citation database, returned us duplicate listings for the author we searched for. However, MathSciNet <sup>4</sup>, an online review service for the mathematical publications that employs mixture of automated and manual name disambiguation, gave us highly accurate and complete results for our author search.

The proposed disambiguation algorithms for information found in more heterogeneous web resources assume that various web resources containing information about a certain person have some common properties among them and that the web resources can be algorithmically clustered based on these properties. The common properties can be generalized into three categories:

1. the textual information provided by the web resources,
2. the biographical information presented by the web resources, and
3. the web links contained in the web resources.

The disambiguation algorithms based on textual information assume that documents about a person either contain some common words or have similar grammatical structures. The disambiguation algorithms based on the biographical information assume that all documents containing personal data provide common biographical information (such as personal name, date of birth, and work history) about the subject and that the biographical information can be identified through some pattern matching algorithms. Finally, the disambiguation algorithms based on the web links assume that different web resources pertain to same data subject are connected through cross-referencing of web links.

---

<sup>2</sup><http://www.isiwebofknowledge.com/>

<sup>3</sup><http://www.scopus.com/scopus/home.url>

<sup>4</sup><http://www.ams.org/mathscinet/>



In [2], Bekkerman and McCallum propose a two-step clustering approach that utilises both the textual similarity and the web links interconnectedness. For textual-based clustering, the proposed approach employed an agglomerative/conglomerative double clustering (A/CDC). The agglomerative clustering, or bottom-up clustering, is a clustering algorithm where every element forms its own cluster at the start and these clusters are repeatedly merged until some termination condition is met. The conglomerative clustering, or top-down clustering, takes an opposite approach to agglomerative clustering. The conglomerative clustering starts with all elements in a single cluster and recursively divides clusters until some termination condition is met. The clustering algorithm proposed by Bekkerman and McCallum use agglomerative clustering to cluster document and conglomerative clustering to cluster all words contained in the documents. In this clustering algorithm, the clusters of words are first divided using conglomerative clustering, then based on the resulting clusters, the documents are merged using agglomerative clustering. This process is repeated until there are only three document clusters, at which point one of the clusters is chosen based on inter-connectedness of the web links. To evaluate the performance of the proposed algorithm, Bekkerman and McCallum gathered a data set of 1085 web pages containing information about 187 individuals. These web pages were collected by querying for 12 names against Google and manually filtering out junk web pages. For the comparison, they chose a greedy agglomerative clustering based on textual information, a clustering based on interconnectedness of web links, and a textual-clustering using A/CDC. The test results showed that the proposed algorithm outperformed all of the clustering algorithms evaluated in terms of F-measure and recall, while the precision of the proposed algorithm was comparable to, if not better than, that of the other algorithm. In addition, they provided the number of correct and incorrect matches the proposed algorithm made during the evaluation. The proposed algorithm made correct association for 313 pages, while failing to cluster 107 pages

and making incorrect association for 47 pages.

In [14], Mann and Yarowsky focus their research efforts on evaluating importance of discriminating biographical features when applying clustering algorithm for name disambiguation. In order to do so, Mann and Yarowsky examined four different algorithms:

1. base model clustering based on proper nouns,
2. clustering based on relevant words,
3. clustering based on biographical features, and
4. clustering based on extended biographical features.

A clustering algorithm based on cosine similarity of proper nouns was chosen as the base model. For second set of algorithms, Mann and Yarowsky examined relevant word extraction based on mutual information (mi) and term frequency-inverse document frequency (TD-IDF) weighing algorithm. For third algorithm, Mann and Yarowsky employed a clustering based on words that were considered to represent biographical features. In this approach, the biographical features are identified as proper nouns that match predefined syntactical pattern. For instance, an occurrence of such pattern as a proper noun followed by four-digit number, e.g. Mozart, 1756, is considered to be a biographical feature indicating birth year of person. For the last algorithm, the biographical feature extraction is modified to give more weights to certain words. For example, if the string pattern 1756 has been identified as birth year by a number of web resources, any subsequent appearance of this string is given higher probability of being a biographical feature. These algorithms were tested against the results from Google search of eight different names (Haifa Al-Faisal, William Blake, Tom Cruise, Woody Harrelson, Hermann Hesse, Wolfgang Amadeus Mozart, Anna

Shusterman, and Bryon Tosoff). The test showed that combining the proper noun extraction, relevant words, and biographical feature extraction fared best, while adding extended biographical features did not improve the result.

In [21], a clustering method based on the biographical and lexical features is proposed by Wan, Gao, Li, and Ding. In the proposed system, which leverages results extracted from a pre-existing search engine such as Google, five biographical features are extracted from the results of query:

- personal name,
- title,
- organization,
- email address, and
- phone number.

Although the system assumes that the query term consists of the personal name, Wan, Gao, Li and Ding propose an algorithm for extracting names in case the resulting document only contains the surname of the individual. In this algorithm:

1. the groups of words that contain the queried terms and have the first character of each word capitalized are extracted;
2. these groups are scored based on heuristics and frequency of their appearance;
3. the most commonly referred groups are chosen as either canonical full name or a given-name/surname pair.

The other four biographical features are extracted using simple pattern matching. The clustering is done using agglomerative clustering, which terminates when the similarity between the documents in a cluster fall below some threshold. The similarity among documents is calculated based on the frequency of:

- lexical features, such as title words, meta-words, and words in the text,
- linguistic features extracted using in-house base Noun-Phrase (baseNP) extractor and Name Entity Recognition (NER), and
- biographical features.

For evaluation, Wan, et. chose a simple lexical clustering as base model and compared its performance against a clustering algorithm based on lexical and linguistic features and a clustering algorithm that combines the result from lexical and linguistic feature extraction with the result from personal information extraction. These algorithms were tested against the results from the queries of the 200 names that were most frequently searched using the MSN. For each query, the top 100 web pages that were retrieved successfully by MSN were collected. The result showed that adding personal information improved the result by about 7% compared to simple lexical clustering, whereas just adding linguistic feature only improved the result by less than half that amount.

In [11], Harada, Sato, and Kazama propose another method based on personal name matching. In this method, the results from a web search are first filtered based on the relevance score returned from the search engine. Then, words that are presume to reference an entity (i.e. a personal name) was extracted from each document. These references are grouped based on the similarity of spelling. Finally, a group is scored based on the proximity to the given personal name and on a function that measures likelihood that a group references a single individual rather than multiple individuals with similar names. For aforementioned likelihood-calculation, Harada, et al, propose four different alternative scoring functions. In the first scoring function, called document frequency (df), the score is calculated simply based on the number of pages in a group. In the second function, called server frequency (sf), the score is calculated mainly based on the number of web servers found in a group, with

an addition of a small factor of  $df$ . For the last two scoring functions,  $dfidf$  and  $sfidf$ , Harada, etc. applied inverse document frequency ( $idf$ ), which is obtained by applying a logarithmic function to the quotient of the total number of web pages being considered divided by the number of web pages in the group. The  $idf$  factor is multiplied by  $df$  and  $sf$  to get  $dfidf$  and  $sfidf$  respectively. For evaluation, Harada, et al collected 45 search terms, each of which is a name of a musical instrument, a name of sport, or a keyword related to information technology. These search terms were queried against Google to find top 10 relevant people for each search term. The test was to measure the precision of each of the four scoring function as the number of names to be merged changed. The result showed that  $sf$  and  $sfidf$  always outperformed  $df$  and  $dfidf$ , while  $dfidf$  always outperformed  $df$ . However,  $sf$  performed better than  $sfidf$  when the number of people was small, but the precision of  $sf$  degraded quicker than  $sfidf$  as the number of people increased.

In [12], similar to Harada, Sato, and Kazama's proposed system, Kalashnikov proposes another system that takes advantage of biographical features and search engine results. In this system, the clustering is applied to top  $K$  results from the search engine. The clustering is based on the personal names that were extracted using a proprietary method, hyperlink, and email addresses. The algorithm was tested against the methods proposed by Bekkerman and McCallum [2]. Their test shows that their algorithm outperforms the algorithm proposed by Bekkerman and McCallum by 9.5% in terms of F-measure.

In [23], Yang, Chiou, Lee, and Ho propose a grouping of person search results based on link connections. The proposed system clusters the results by applying five components:

1. a web search component collects the search results from the search engines;

2. a data set augmentation component augments the web page dataset by following both the incoming and the outgoing hyperlink up to l-level, denoted as augmented process (AP), and extracting hosts of those links, denoted as host-based augmented process (HAP);
3. a popular node removal (PNR) component removes several popular web sites, which may cause too much hyperlink connections between pages;
4. a network motif detection component detects a cluster of web pages with shortest distance;
5. a web page grouping component clusters the web pages into several groups and returns the grouped results to the user.

A purpose of network motif detection is to show that certain number of web pages is connected directly or through some intermediary pages. The network motif pattern type can be defined by the number of intermediary pages allowed. In this paper, all possible network motif patterns between the 2-node pattern, which allows no intermediary page, and the 5-node pattern, which allows up to 3 intermediary pages, types were examined. For determining whether there is a connection between two pages, Yang, et al. examined two different approaches. In the first approach, called page-based detection (PD), two URLs are considered equivalent if the URLs are exactly the same. In the second approach, called host-based detection (HD), two URLs are considered equivalent if they have a same domain name. For the evaluation, different combination of the network motif under different data set augmentations and different motif detection approaches were applied against the data found in [2]. The results show that a F-measure improves when more network motifs are used under the AP+PD. This is due to the fact that precision decreases when an error hyperlink is introduced to the network. The best overall results are achieved by network motifs

consisting of 4 or less nodes (2+3+4 node motifs) under the HAP dataset and the PD approach, with the F-measure of 70%. This result is similar to that of 4-node network motifs using HD approach only. The difference is in recall rate, where HAP+PD achieves the recall rate of 70%, while HD only achieves 60%. This result may indicate that using host-based detection can increase the number of entities disambiguated correctly of disambiguation without using dataset augmentation process, but the dataset augmentation may help in identifying more ambiguous entities.

### 3.2 User-assisted Name-disambiguation of Personal Data on the Web

Although, much of current research efforts have been concentrated on developing an effective clustering-algorithm for name-disambiguation in web resources, many social media web sites and online repository of academics sources have employed user contribution to either address or supplement aggregation achieved through the clustering-algorithm. These efforts resulted in development of various algorithms for incorporating user contribution, which can be broadly divided into two categories: *closed-systems* and *open-systems*.

In a closed-system, the name-disambiguation contribution can be made by a limited number of authenticated users. In our investigation, we have identified following three approaches to implementing a closed user-assisted name-disambiguation system:

- an identity-agnostic name-disambiguation by system-approved users,
- a peer-based name-disambiguation, and
- a profile-based name-disambiguation by system-approved users.

In a system that employs identity-agnostic name-disambiguation approach, any user who has been granted permission can make name disambiguation suggestion on any data. The system is identity-agnostic in the sense that it is not aware of

who has made an association between a data and a person, being only aware of the fact that is made by a user approved the system. An example of a system that employs this approach is MathSciNet [19], a sophisticated author-identification tool containing information on more than two million publications. MathSciNet employs both the name-disambiguation program and staff member contributions to determine association between the publications and the authors, where the name-disambiguation program is successful in making association in eighty percent of the time and the staff members disambiguating the remaining twenty percent.

In a peer-based name-disambiguation system, a group of authorities maintaining a set of disambiguated data, share data and disambiguation suggestion through a centralized system. An example of a peer-based name-disambiguation system is the Linking and Exploring Authority Files (LEAF) [3], which aims to develop a centralized disambiguated data source, called “Central Name Authority Files,” for data coming from different European libraries and archives.

In a profile-based name-disambiguation, each authenticated user is responsible for maintaining the aggregated data of himself/herself through a profile page. The professional societies, such as the Association for Computing Machine (ACM) and IEEE, have implemented profile-based social networking services as a way to aggregate their members’ professional information.

In an open-system, name-disambiguation can be performed by any user, whether they are authenticated or unauthenticated. Some of these systems limit the contributions to the users who have created an account, but these systems usually do not restrict who can create an account, nor do they validate the identity of the users. There are two popular models for open user-assisted name-disambiguation systems:

- the consensus-based name-disambiguation model, and
- the profile-based model.



The consensus-based model, popularized by Wikipedia, seeks to obtain contribution from as wide range of users as possible. In purest form, any user can make any disambiguation changes at any time [22]. Some examples of sites that employ the consensus-based model include LibraryThing <sup>5</sup>, and personal search engines like iSearch <sup>6</sup>, Wink <sup>7</sup>, and PeekYou <sup>8</sup>.

The profile-based open-system model is similar to the closed-system. The difference is that in the closed-system, only the authenticated user can take responsibility for a profile page. In the open system, any user can take responsibility for a profile page that does not have a user assigned to it. This strategy is popular among many social networking sites, such as MySpace <sup>9</sup> and Facebook <sup>10</sup>, and social bookmarking sites, such as Delicious <sup>11</sup>, StumbleUpon <sup>12</sup>, and Digg <sup>13</sup>.

The main strength of open user-assisted name-disambiguation system is the potential for accelerated growth in the size of data-set and for swift correction of incorrect data. By opening the system up to general public for aid in name disambiguation, it is possible to maintain a large and growing set of data as the size of the user community grows. The main weakness of open name-disambiguation system is that the users cannot have same confidence in data quality as with closed systems. A key problem, therefore, in the design of a user-assisted name-disambiguation system is how to combine the strengths of the open and closed approaches. The goal should be to harness the power of many unauthenticated contributors while maintaining an effective review system based on authenticated users.

In this thesis, we propose a system, called *Name-Authority System*, that would

---

<sup>5</sup><http://www.librarything.com/about>

<sup>6</sup><http://www.isearch.com/>

<sup>7</sup><http://wink.com/wink/about>

<sup>8</sup><http://www.peekyou.com/>

<sup>9</sup><http://www.myspace.com>

<sup>10</sup><http://www.facebook.com/>

<sup>11</sup><http://www.delicious.com/help/learn>

<sup>12</sup><http://www.stumbleupon.com/>

<sup>13</sup><http://digg.com>

allow unauthenticated users to contribute name-disambiguation suggestions, while only presenting to end-users data that has been filtered by the authenticated users. In the remainder of this thesis, we explore the design and the implementations of two Name-Authority Systems.

## Chapter 4

### Designing a Name-Authority System

In this chapter, we outline the key requirements that we believe a name-authority system should address and describe two design alternatives. In Section 4.1, we discuss the strengths and the weaknesses of the previous efforts to address the name disambiguation problem and propose a set of design requirements. In Section 4.2, we examine current legal issues surrounding the use and transfer of personal information through the web and suggest a data control policy that we believe addresses them. In Section 4.3, we examine other design requirements for a name-authority system. Finally, in Section 4.4, we examine two design alternatives for a name-authority system proposed in this thesis.

#### 4.1 Name Disambiguation

To help motivate our selection of key design requirements for a name-disambiguation system, we will explore the strengths and weaknesses of the previously proposed automated and user assisted approaches. As discussed in the Chapter 3, previous approaches to name disambiguation have tended to be either based on cluster or user input. The use of clustering algorithm is based on the assumption that the underlying data shares some common information, such as the textual information provided by the web resources, the biographical information presented by the web resources, or the web-links contained in the web resources. The problem with these clustering algorithms is that the presence of common features used to cluster may not indicate common subject. In the case of the textual similarity and common links, it may

indicate common authorship, but it is not clear whether it would indicate the common subject as well. In the case of common biographical information, often the only common information is the individuals' names. In fact, as described in Chapter 3, there has been a disproportionate amount of research focused on extracting names from web resources at the expenses of other types of biographical information. In addition, the clustering algorithms based on textual similarity or common biographical information require the information about the individual to be textual, which excludes pages that primarily provide non-textual information, such as photographs of the researchers or video lectures.

The limited effectiveness of personal information aggregation based on clustering in heterogeneous environments such as the web has been widely recognized and leads us to explore user-assisted approaches [8, 19, 2]. For example, the Association for Computing Machinery (ACM) in its overview of the newly created Profile Pages for its members points out the need for human intervention to present accurate disambiguation of academic citations [8]. This has led to a general awareness of the importance of including user input when performing name disambiguation. Various ways of incorporating user input has been proposed including open-systems, where any user can contribute data, and closed-systems, where only approved users can contribute data.

An advantage of the open-system approach, especially when the size of the user community is large, is that it can collect the contribution from the large number of users, enabling it to scale well. The downside to using an open-system approach is that it is vulnerable to corruption of data due to anonymous malicious users contributing erroneous data, diminishing the trustworthiness of the system.

In a closed-system, we can ensure the accuracy of data, or at least the credential of the users who are responsible for the contribution, by limiting the users who can contribute data. The challenge becomes one of scale: because the number of users

who can contribute to the data is limited, it is difficult to get contribution from wide range of sources, and often therefore difficult to deal with large heterogeneous data sets such as the web.

The name-authority systems proposed in the remainder of this thesis take a modified user-assisted approach that combines the strengths of both open and closed systems. They bridge the gap between the open-system and closed-system approaches by allowing any user to input data into the system, but discriminate how that data is disseminated based on the system's knowledge of the contributor's identity. This way, the system can take advantage of the knowledge of the wider user community, while preventing indiscriminate incorporation of data which may diminish the trustworthiness of the data.

#### **4.1.1 Name Disambiguation in a Name-Authority System**

Recall that our motivation for constructing a name-authority system is to aid in the construction of profile pages. A profile page is a web page containing links to a set of web resources that have been identified to belong to an individual. Also, a profile page may contain a short description of the individual being referred to in order to help users to better identify the said individual. To construct a profile page, we believe a name-authority system should accept contributions from both unauthenticated and authenticated users, but should treat them differently in order to diminish the chance of erroneous data being associated with a profile.

In a name-authority system, we believe unauthenticated user should easily be able to contribute name disambiguation information. However, their contribution should not be automatically incorporated into public view of the profile page. Rather, the name-authority system should provide a mechanism for authenticated users to approve or reject each contribution. Also, if there is to be some delay between the

unauthenticated user's contribution of name disambiguation information and the authenticated user's approval or rejection of it, the name-authority system needs to provide access control making it available to authenticated users only. Finally, the name-authority system needs to label the source of the contribution as being an unauthenticated user.

Authenticated users should be able to both contribute name disambiguation information and correct any erroneous name disambiguation information contributed by others. Also, the name disambiguation information contributed by the authenticated users should be available to the public immediately. Finally, each profile needs to have at least one approved user who is assigned as an authority on that person. An authority on a profile page must have been authoritatively identified by the researcher who is the subject of the profile page.

## **4.2 Data Control**

Data control is the second key function of a name-authority system. In helping to construct profile pages, a name-authority system must control the use and transfer of the personal information over the web. How the data control issue is understood and addressed within the name-authority system not only has ramification in how trustworthy the data is but also has legal ramification as discussed below.

Unfortunately, there is no international consensus on what legal protection needs to be provided around the online transfer and the use of personal information [4, 13]. For instance, the US federal government does not currently provide a comprehensive law regulating use of personal information, except for highly sensitive information such as medical and financial record. In contrast, the European Union, through Directive 95/46/EC of the European Parliament [7], has created standards for the processing of personal data and the free movement of such information. This directive, commonly known as the Data Protection Directive, mandates its member

states to regulate the use of any information that can be used to identify an individual. The difference in legislation governing the use of personal information stems from differences in cultural, historical and political realities of various regions and countries [18, 24, 17, 9, 1].

Another contributing factor that accounts for the difference of laws governing personal information in digital media is the varying assessments of the impact of new communication technology, such as the Internet, on personal privacy [20, 9]. In particular, H.T. Tavani states that “the difference in these laws reflects whether the legislators view the privacy on the web necessitating new privacy policy or view it as warranting mere enhancement of the pre-existing privacy policy [20]”.

One possible consequence of the differences in laws regulating the use of personal information is that it can pose a market entry barrier to enterprises that provide personalized web services. For example, online retailers that allow users to make online payment and present personalized suggestion list, search engines that attempt to present more relevant results to their users by caching previous search results and users online activities, and social-networking sites that allow users to share personal information with families, friends, and acquaintances may have difficulty in expanding their operation to overseas market if the law governing the use of personal information in the prospective market significantly differs from the law of the home country. This is because the use of personal information is regulated by the place of residence of the *data subjects* - the people who are the subjects of the information in question, instead of the location of the *data controller* - the party responsible for maintaining the repository containing the personal information. Therefore, users residing in jurisdiction with strict regulation may only be able to access restricted version of some personalized services or may not even be able to access these services at all. Furthermore, much of the legislation also regulates and restricts the transfer of personal information to certain jurisdiction, so many of these enterprises may not even be able

to provide these services without opening an office in these jurisdictions.

To mitigate the negative economic and social effects of conflicts in these laws, as well as provide a comprehensive protection against misuse of personal information, the Organization for Economic Co-operation and Development (OECD) adopted the Guidelines on the Protection of Privacy and Trans-border Flows of Personal Data (OECD's Guidelines) in 1980 [6]. The stated goal of these guidelines was to provide a common framework for privacy laws so that the difference in privacy laws of various nations does not become an obstacle in the free flow of personal information, an obstacle that could become a detrimental to the electronic commerce, while preventing misuse of personal information as it is readily made available through digital media [5]. The guidelines established eight principles that a *data controller* should follow in order to provide sufficient protection to the privacy of *data subjects*. The eight principles are [6]:

1. *the collection limitation principle*, which stipulates that personal data should be collected lawfully and fairly, with either consent or knowledge of the data subject, and with clear limitation on the scope of data collected;
2. *the data quality principle*, which stipulates that the data controller should ensure that published data are accurate, complete, and up-to-date;
3. *the purpose specification principle*, which stipulates that the data controllers ought to notify data subjects the purposes for which data are collected at the time of the collection, and that the data controller must collect only the data that fulfil those purposes;
4. *the use limitation principle*, which stipulates that personal data should neither be disclosed, nor made available for use other than to fulfil the stated purposes;
5. *the security safeguards principle*, which stipulates that reasonable measures



must be taken to ensure data against unauthorized access, destruction, use, modification, or disclosure of data;

6. *the openness principle*, which stipulates that development process, practices, and policies with respect to personal data should be made transparent, which include providing a way to establish existence and nature of data collection, the main purpose of the data collection, and the identity of the data controller;
7. *the individual participation principle*, which stipulates that data subjects should have a way to establish whether information about themselves had been collected, and to challenge the data so that it could be erased, rectified, completed, or amended; and
8. *the accountability principle*, which stipulates that the data controller is to be accountable for violation of the privacy policy.

Although non-binding, the OECD Guidelines have been recognized by all member states of the OECD, and the privacy regulations of many nations, including the EU's Data Protective Directive and Canada's *Personal Information Protection and Electronic Documents Act*, reflects the principle outlined in the OECD Guidelines [13, 4, 1].

In Canada, the online usage of personal information, which is defined as any information that can identify a person except for information that can be found on the phone listing, such as title, name, business, address, and telephone number, is regulated by the *Personal Information Protection and Electronic Documents Act* (PIPEDA, Bill C-54), enacted in April 13, 2000 [16]. This act, which is based on the Canadian Standards Associations *Model Code for the Protection of Personal Information*, is based on ten principles, which closely reflects the eight principles outlined in the OECD Guidelines (see Figure 4.1). Also, the PIPEDA act provides number of exclusions, such as data used for journalistic, artistic, or literary purposes.

1. Accountability	An organization is responsible for personal information under its control and shall designate an individual who are accountable for the organization's compliance with the following principles.
2. Identifying Purposes	The purposes for which personal information is collected shall be identified by the organization at or before the time the information is collected.
3. Consent	The knowledge and consent of the individual are required for the collection, use or disclosure of personal information, except when inappropriate.
4. Limiting Collection	The collection of personal information shall be limited to that which is necessary for the purposes identified by the organization. Information shall be collected by fair and lawful means.
5. Limiting Use, Disclosure, and Retention	Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by the law. Personal information shall be retained only as long as necessary for fulfilment of those purposes.
6. Accuracy	Personal information shall be as accurate, complete, and up-to-date as is necessary for the purposes for which it is to be used.
7. Safeguards	Personal information shall be protected by security safeguards appropriate to the sensitivity of the information.
8. Openness	An organization shall make readily available to individuals specific information about its policies and practices relating to the management of personal information.
9. Individual Access	Upon request, an individual shall be informed of the existence, use and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.
10. Challenging Compliance	An individual shall be able to address a challenge concerning compliance with the above principles to the designated individual or individuals for the organization's compliance.

Figure 4.1: 10 Principles of Personal Information Protection and Electronic Documents Act [16]

In the EU, all member states are required to enact a legislation that would provide at least as much protection on processing of personal information as it is stipulated in the Data Protection Directive. Reflecting all principles of the OECD Guidelines, the Data Protection Directive elaborates upon it. Some of the stipulations in the Data Protection Directive include [7]:

- data controller must ensure
  - that data is processed fairly and lawfully,
  - that data is collected for explicitly stated purpose,
  - that data is accurate and current,
  - and that data are kept no longer than necessary for the stated purpose;
- data shall only be processed when
  - data subject has given unambiguous consent,
  - and processing is necessary to meet the purpose of the service;
- data controller, at the time of disclosure of data, must provide to data subject
  - the identity of data controller,
  - the purpose of processing,
  - and other information pertaining to the use of data;
- data subject has rights
  - to obtain data, or at least the confirmation as to whether data pertaining to him is being processed, in a timely manner without being charged,
  - to obtain knowledge of logic in the automated process,
  - and to erase or block inaccurate data.

- data controller must ensure that data is only released to the party authorized by the data subject and must take appropriate security measures to ensure this;
- data can transfer to data controllers residing in countries outside of the EU
  - if the country in question provides adequate level of protection on the use of personal information,
  - data subject has consented to the transfer,
  - or data needs to be transferred to meet a contractual obligation;
- and data cannot be collected if it pertains to the data subject's ethnic origin, political opinion, or religious affiliation.

The Data Protection Directive, however, does provide exemptions in extraordinary circumstances, such as matters concerning state security.[7]

One important consensus the Data Protection Directive and PIPEDA is that the measures taken to protect the personal information need to be commensurate to the sensitivity of the data [7, 13, 4]. Although, the Data Protection Directive does not make exceptions for such common information as those can be found in the phone book, the threshold for the measure expected to protect this kind of data are much lower than the threshold for the measure expected for the financial and health records.

#### **4.2.1 Data Controlling in a Name-Authority System**

Given the rise of some consensus around data protection, as evident in the Data Protection Directive, the OECD Guidelines, and the PIPEDA, we propose following requirements for the name-authority in regards to the use and transfer of the personal information:

- the data subject must grant an informed consent to release the information,

- the data subject must be able to update and challenge the information,
- the data subject must be able to delete a piece of information or the whole profile from the system, and
- the type of information accrued by the system should be restricted to personal name(s), links to other web pages providing information about the researcher or researcher's activity, publication history, biographical and contact information, and some other non-sensitive piece of information that can help in identifying a researcher uniquely, such as a profile picture.

### 4.3 Other Design Requirements for a Name-Authority System

In addition to aforementioned requirements, we believe that the name-authority system should conform to following requirements:

- the system should implement the search interface, allowing users to search for information on a researcher by name,
- the system should provide unique link to each profile page,
- the system may provide the data in alternative format, such as XML, so that it can be used by the third-party, and
- the system should have some way to ensure either no malicious web page can be linked through the system, or to flag suspected malicious sites.

The unique link and the alternative format are required for the interoperability. Since the name-authority system leverages off of the data from third-party, having the data in accessible format could encourage other parties to make the data more easily accessible.

#### 4.4 Design Alternatives for a Name-Authority System

At the heart of implementing a name-authority system lay two questions: who does the system authenticate, and what data control privileges does the system grant to each group of authenticated users? A name-authority system can authenticate either groups of researchers, similar to the FWDM's authentication of Mathematical societies, or individual researchers, as is the case in the profile-based user-assisted name disambiguation system. For data control, the question is whether the authenticated users should review every contributions made by unauthenticated users, or whether to manage user contributions by assigning different privileges based on the access level of a user.

In this thesis, we explore two alternatives approaches to these questions. In the first alternative, called a *Data-Filtered Name-Authority System*, we are mainly concerned with authenticating authorities representing a group of researchers who, then, are responsible for reviewing every contribution. In the second alternative, called a *user-filtered name-authority system*, we authenticate individual researchers, who in turn assign other users with the privilege to contribute data within strict limitation.

##### 4.4.1 Data-filtered Name-Authority System

Similar to the FWDM, the purpose of the data-filtered name-authority system is to design a system that would facilitate the management of information by authenticated users, where each authenticate user represents a group of researchers. In the proposed design, which is shown in Figure 4.2, the data contribution and the name-disambiguation are open to any user, but these contributions are examined by the authenticated users before being made public. This is achieved through a design consisting of three modules:

- access control module,

- name-disambiguation module, and
- search user-interface.

The access control module, which is responsible for managing users, fulfils two purposes. First, it distinguishes between the authenticated and unauthenticated users. Second, it discerns the roles and responsibilities of each authenticated user. Both of these purposes are accomplished during the log-in process. When the user submits the log-in request, the access control module checks the information provided by the user against the database and, if the user exists, returns the role assigned to the user, in some form of web credential, back to the user.

*The name-disambiguation module* is where the users submit the input data and the name-disambiguation suggestions. Through the name-disambiguation module, the data-filtered name-authority system exposes the data it currently holds and the researcher each data is associated with. If a user notices a missing data or incorrect association, the user submits his/her contribution to the name-disambiguation module either with the credential, in case of an authenticated user, or without. The name-disambiguation module, then, submits this contribution, along with the credential, to the database.

The information about each researcher is exposed to the wider public through *the search user-interface*. Based on a query submitted by a user, this module searches for researchers matching the information provided in the query. After successful search, the search user-interface returns the researchers and the information associated with each researcher in a list of results. To ensure that only the information approved by the authenticated users is returned to the user, the search user-interface checks whether each piece of information associated with a researcher has proper credential before returning it to the user in the result list.

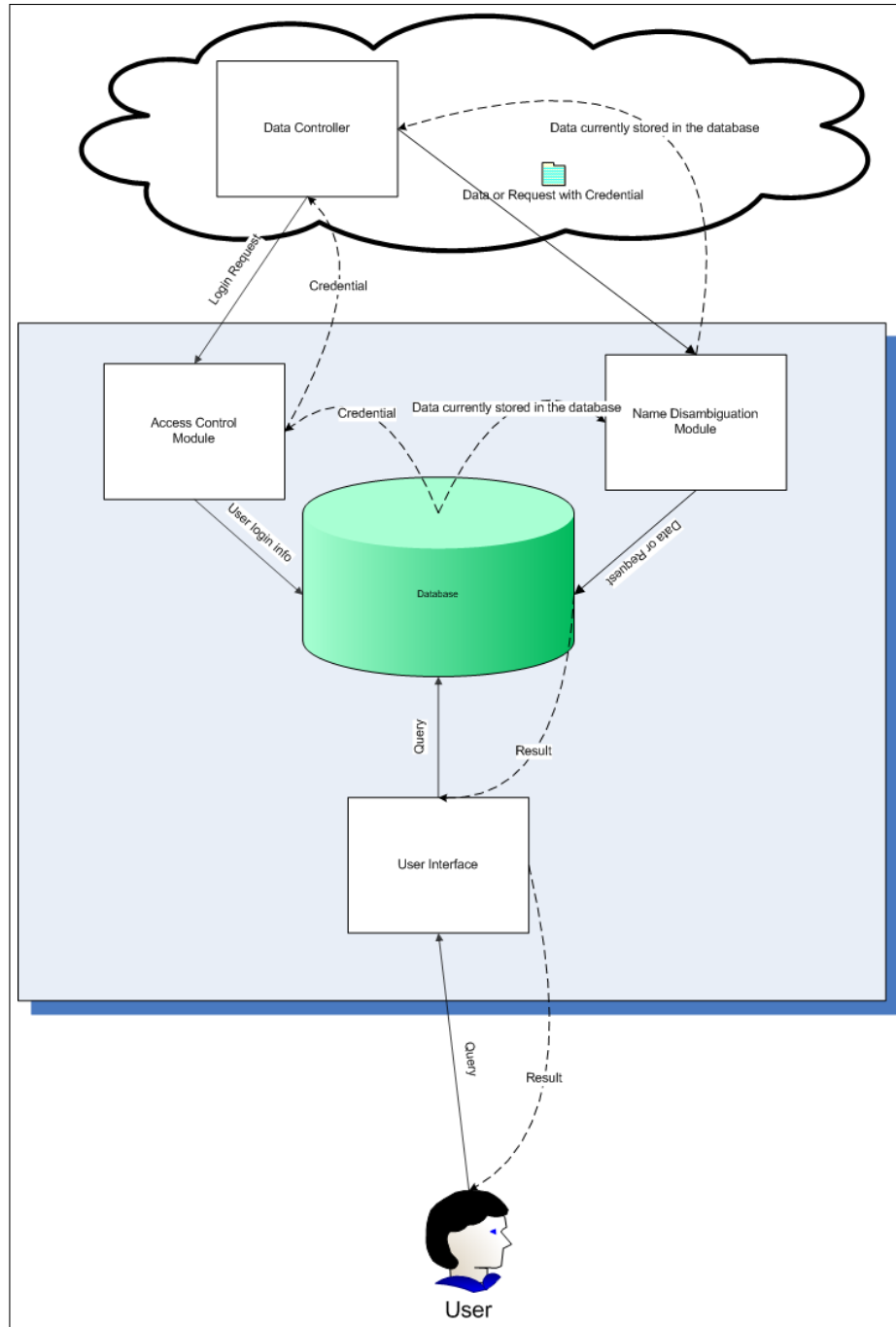


Figure 4.2: This diagram demonstrates the basic design of the data-filtered name-authority system, along with the data flows within the system.



#### 4.4.2 User-filtered Name-Authority System

For the user-filtered name-authority system, our design is based on the idea that the decision to accept the contribution from a user is made based on the trust level the researcher (who is the subject of the contribution) has with the user making of the contribution. In this system (see Figure 4.3 for the design diagram), each user is assigned with a trust level against each profile by either the system or by the researcher who owns the profile. The user can contribute data directly against the profile if his/her trust level is sufficient.

In the user-filtered name-authority system, *the access control module* determines the acceptability of the contribution of a user. The access control module in the user-filtered name-authority system, unlike in the data-filtered name-authority system, assigns different credentials to a user for different profiles. In this design, a user who wishes to contribute some information regarding a researcher is asked for his/her credential for the profile in question. If the given credential is sufficiently high, the user can submit his contribution directly to the profile. Otherwise, the contribution gets rejected. Because each contribution is directly assigned to a profile, a user-filtered name-authority system does not require a name-disambiguation module.

The search user-interface for the user-filtered name-authority system behaves similarly to the one in the data-filtered name-authority system. One difference is that the search user-interface for the user-filtered name-authority system does not require filtering the user contribution based on the credential as the contribution with insufficient credential gets rejected before it can be stored into the database.

In the following chapters, we describe two prototype systems that demonstrate these two design alternatives. The two prototype systems are designed to provide information on researchers in mathematical science fields, such as pure and applied mathematics, statistics, computer science and operations research. The types of

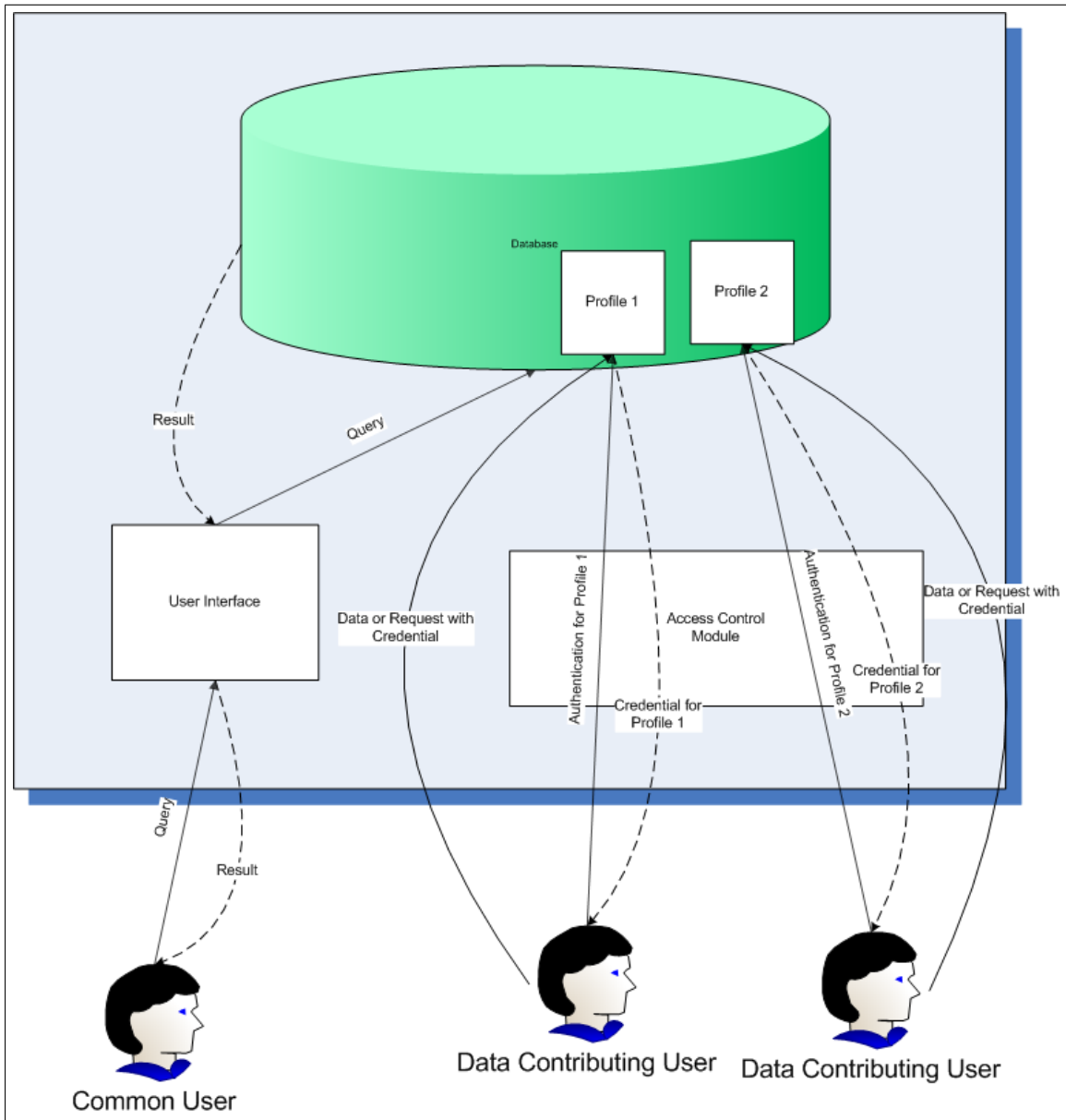


Figure 4.3: This diagram demonstrates the basic design of the user-filtered name-authority system, along with the data flows within the system.

information these systems aim to provide include:

- contact information of the researcher,
- biographic information about the researcher,
- bibliographic information from the researcher's publication history, and
- information on the researcher's current and past researches.

The prototype for the first system, called *MathPeople*, was implemented by Jaehyun Paek. The primary supervisors for this project were Dr. Jonathan Borwein and Dr. James Pitman. Also, Dr. Mason Macklem and Dr. Hadley Wickham provided technical supervision. It utilizes a design based on the data-filtered name-authority system approach, where personal data are collected through proxy data controllers. In the second system, called *Populi Scientiae*, a prototype designed and developed by Jaehyun Paek, was initially focused on serving members of the Faculty of Computer Science at Dalhousie University, we introduce a user-filtered name-authority system approach.

## Chapter 5

### MathPeople - The Data-filtered Name-Authority System

In this chapter, we provide in-depth discussion on implementation of MathPeople, a prototype data-filtered name-authority system. The primary goal of MathPeople is to extend the functionality of the FWDM to provide, in addition to contact information, references to other types of web resources such as web pages providing biographical or bibliographical information about a researcher while providing same level of data control. Instances of MathPeople were hosted on servers maintained by D-Drive, an experimentation facility at Dalhousie University that was directed by Dr. Jonathan Borwein until its closure in 2009 and by the Institute of Mathematical Statistics from 2007 to 2010. During its deployment, these instances of MathPeople maintained over 140,000 records from more than twenty different sources.

This chapter is organized into the following sections. In Section 5.1, we provide an overview on the design of MathPeople. In Section 5.2, we examine the different types of users defined by MathPeople and how they are managed. In Section 5.3, we examine how MathPeople manages contributions from both authenticated and unauthenticated users. In Section 5.4, we examine the user interfaces for MathPeople's search page and profile pages. Finally, in Section 5.5, we discuss the strengths and weaknesses of MathPeople implementation.

## 5.1 System Overview

As discussed in Section 4.4.1, the data-filtered name-authority system leverages the authenticated users to filter user contributions based on the content of the contribution. In MathPeople, this filtering is achieved by distinguishing the data that is exclusively managed by the authenticated users and those that are managed by both the authenticated users and the unauthenticated users, as shown in Figure 5.1. The database for MathPeople is based on following nine entities:

- Person.
- Record.
- Suggested\_Record.
- Suggested\_Relationship.
- User.
- Role.
- Right.

Central to the MathPeople design is the Person entity. The *Person* entity represents the researchers in MathPeople. The main attributes of the Person entity are the name of the researcher and the person ID, an integer key assigned by MathPeople. All the contributions from both the authenticated and the unauthenticated users are associated with this entity; through this association, the information about a researcher is exposed to the end-users.

For the data that is managed exclusively by the authenticated users, MathPeople supports the *Record* entity. A record encapsulates a web resource that provides information about a researcher. An example of a web resource encapsulated by a

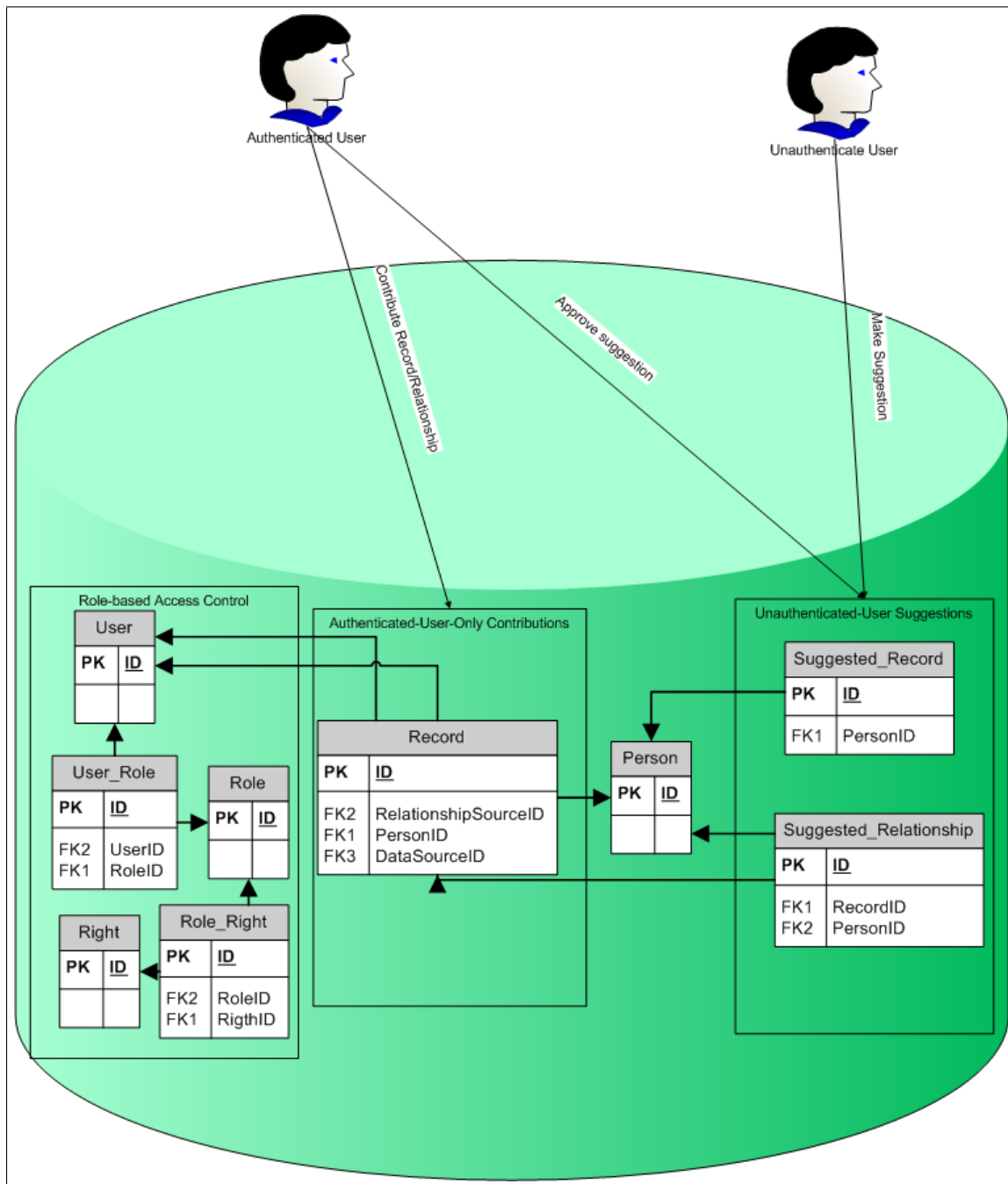


Figure 5.1: The overview diagram of how the user's contribution interacts with the database. The ER-diagram for the database only shows the key entities in MathPeople and their relationship with each other through the primary keys and foreign keys.

record is a web page containing publication history of a researcher. However, as an authenticated user can add a row into the Record table, while another can make the association between this record and a person, the Record table maintains two foreign keys to the User table: the DataSourceID key for the user who entered or updated the data, and the RelationshipSourceID key for the user who made association to the person that the record currently refers to.

Mirroring the schema for the authenticated user contributions, MathPeople supports the *Suggested Records* entity for the data contributed by the unauthenticated users. In addition to the Suggested Records entity, MathPeople supports the Suggested Relations entity for the name-disambiguation suggestion made by the unauthenticated users. As a suggestion made by an unauthenticated user becomes publicly available only when it is approved by an authenticated user, MathPeople allows the authenticated users to interact with both the Suggested Records and the Suggested Relations in order to approve them. The management of user contribution is discussed in further details in Section 5.3.

For controlling the access control of the authenticated users, MathPeople supports role-based access control through following entities: *User*, representing individual authenticated user, *Role*, representing a given responsibility for a given user and *Right*, representing a type of tasks and user can perform given a role. In MathPeople, each user object, which contains the username and encrypted password as attributes, is assigned to one or more roles and each role has some rights dictating what contribution the user can make. The *user\_roles* and the *role\_rights* entities represents the relationship among the Users, the Roles, and the Rights entities. Further discussion on how MathPeople manages users is provided in Section 5.2

## 5.2 User Management

For the role-based access control in MathPeople, we have defined three different roles for the authorized users. These are:

- the data controller,
- the editor and
- the administrator.

A *data controller* is the primary interface between the researchers and MathPeople. In many cases, a data controller will represent a body representing a group of researchers, such as a professional society or a faculty. Each data controller is responsible for managing personal information on a group of researchers, including obtaining consent from each researcher on the use of his/her personal and research information by MathPeople and adding and updating the information on each researcher. For interaction with MathPeople, a data controller may directly upload contribution using MathPeople webpage or upload contribution by setting up a web application that will automatically interact with MathPeople. MathPeople grants four different rights to these users, which are:

- insertion/modification of records,
- insertion/modification of association between records and people,
- approval of suggested records and suggested relationships, and
- setting of the canonical names for researchers.

An *editor* is a member of MathPeople responsible for assuring that the records are associated to correct researchers and for approving suggestions. As with the data controllers, MathPeople grants editors the right to



- insert/modify the association between records and people,
- approve suggested relationships and suggested records, and
- set the canonical names for researchers.

However, an editor does not have the right to insert or modify records.

To manage the authenticated users, MathPeople defines another set of users: *an administrator*. The responsibility of the administrator is to:

- Add a new data controller or editor.
- Remove a data controller or an editor.

### 5.3 Data Contribution Management

As discussed Section 5.1, the information on researchers is represented by the Record and the Suggested\_Record entities. To ensure that the information disclosed through MathPeople is of low-sensitivity in nature, the information regarding a researcher contained within a record is limited to the researcher's name and the link to webpage that contains information about the researcher. Also, only the record, which are generated by the authenticated users, are exposed to the end-users through a different process than the suggested records, which are generated by the unauthenticated users.

#### 5.3.1 Managing Contributions from Authenticated Users

In MathPeople, only the data controllers can add, delete, or update a record. MathPeople expects all records to contain following three mandatory fields of information:

- the name of the researcher it is associated with,
- the URL of the record's webpage, and
- the data source identifier.

As a researcher can be associated with multiple names, MathPeople assigns one of the names as being “canonical.” This name is the name that is first appears on both the search result page and the profile page of MathPeople. MathPeople may set one of the names found in records associated with a researcher as a canonical, initially the name attached to the first record that gets associated to the researcher, or a name supplied by a data controller or an editor. All names from the records that are not set as canonical are assigned as aliases. These names, although not returned on the result page, are listed on the profile page, helping user to identify the researcher.

To ensure the quality of the data, the webpage linked by a record in MathPeople is limited to the webpages from the well-known data sources, such as MathSciNet, a high-quality bibliographic information database maintained by the AMS, and Wikipedia. The data source identifier is a unique key to easily identify the webpage and its source associated with a record. The data source identifier consists of the data source alias, a short-hand name given to a data source when it is first registered, and the ID that the data source has assigned to a given link. If a record contains the data source identifier that already exists in MathPeople, this is identified as the update of the previous record and update the data accordingly.

In addition to the mandatory fields, a record in MathPeople may contain a short description on the webpage and the name-disambiguation information. For name-disambiguation, the record may contain

- the data source identifier for some record that refers to same person, or
- the ID MathPeople assigned to the person the data source refers to.

If the record does not contain any name-disambiguation information and the matching data source identifier is not found in MathPeople, the person for whom the record is associated does not yet exist or the person for whom the data is associated is unknown to the data controller. The latter case is possible due to the fact that

a researcher may be managed by multiple data controllers. In this case, MathPeople creates a new row in the person table and assigns the record to the new row. The assumption is that, through the name-disambiguation process, the duplicate person entry will be removed quickly.

MathPeople implements two different web-interfaces to allow data controllers to manipulate records. In each of the interface, the maintainer of a data source specifies the task to be performed, whether it is to add or update, or to delete records. In the first interface, MathPeople allows the upload of an individual record through a web form that contains an input field for each of the five fields that constitutes a record. In the second interface, MathPeople allows the user to upload a XML or CSV file containing multiple records. Although both of these interfaces have been designed with a user being an actual person, as oppose to a web-application, it is easy to implement a script that would automatically upload records. A sample Perl-script demonstrating this has been provided through these interfaces.

The data controllers and the editors can disambiguate the records through the result page of MathPeople search page after logging into the system. The name-disambiguation interface allows the authenticated users to disambiguate data in one of three ways:

- through merging multiple person entities,
- through separating a record from a person entity, which results in generation of a new person entity, or
- through assigning multiple records to a person.

The name-disambiguation web-interface for MathPeople consists of a searchable list of researchers, shown in Figure ???. This interface lists all the people matching the search term and the records associated with each of the user. For each person and each record, the interface provides checkbox next to it. For merging people and

The screenshot shows the Mathpeople interface. At the top left is the title "Mathpeople" and three silhouettes of people. Below the title is a navigation bar with "« Previous 1 2 3 Next »". The main content area lists several person entries, each with a checkbox and a list of associated records:

- Pitman, Damien
  - Pitman, Damien ([ArXiv: Pitman-D\\*](#))
- Pitman, J. B
  - Pitman, J. B ([CIS:89602](#))
- Pitman, Eric Bruce
  - Pitman, Eric Bruce ([MR:242247](#))
  - Pitman, Eric Bruce ([MG:38639](#))
- Pitman, Edith J
  - Pitman, Jane ([MR:199765](#))
  - Pitman, E. Jane ([MG:42102](#))
  - Pitman, Edith J ([AusMath.P](#))
  - Pitman, Jane ([Erdos:](#))

At the bottom of the list is another navigation bar with "« Previous 1 2 3 Next »" and two buttons: "Merge" and "Split". On the right side, there is a search box with the text "Pitman" and a "Search" button. Below the search box is a grey box with the text: "Search only in first name 'firstname:jim' or last name 'lastname:wickham'. Find misspellings with 'pitnam-', or use wildcards 'smit\*', 'pitman j\*'. See [Lucene syntax reference](#) for more options."

At the bottom left of the page is the copyright notice: "© IMS. [Log out](#)".

Figure 5.2: The interface for merging person entities in MathPeople. To merge entities, the editor mark the check boxes and press “Merge” button

records, the user will check the checkbox next to the people and records to be merged and click the *Merge* button at the bottom. For splitting a record from a person entity, the user will check the records to be split and click the *Split* button.

When multiple people objects are merged, the merged entities are assigned the ID of the smallest ID among the people being merged. The person ID for other people objects are then aliased to the merged person ID. This is so that previous association with the person does not get lost as person ID is one piece of the name disambiguation information that a record may contain.

### 5.3.2 Managing Contributions from Unauthenticated Users

As discussed in Section 5.1, MathPeople supports the contribution from the unauthenticated users through suggested records and suggested relationships. For the prototype, suggested records have been limited to the homepages of the researchers. The suggestions made by the unauthenticated users are stored in the database, but are hidden from the end-users pending the approval for data controllers, which is achieved through the user interface listing suggested records and suggested relationship similar to the name-disambiguation user interface.

For record suggestion, MathPeople provides input form for the homepage suggestion through the profile page. Through this input form, the user simply enters the homepage URL for the given researcher.

MathPeople also provides a web-browser bookmarklet for record suggestion (see Figure 5.3). This is a small application the user can embed into his/her web browser. When visiting a researcher's homepage, user will open the bookmarklet, which will populate the URL field for the homepage as well as providing small search engine for researchers in MathPeople. If the user already knows the MathPeople person ID for the researcher, user will simply enter the ID. Otherwise, the user will search for the researcher and select the appropriate researcher. Once the user submits the URL and the ID, the bookmarklet will inform the user that the suggestion has been added to MathPeople.

Like the data controllers and the editors, the unauthenticated users can make relationship suggestion through MathPeople's search result page. MathPeople distinguishes the suggested relationship from the name-disambiguation made by the data controllers and the editors by checking whether the user has logged-in to the system and if so, whether the user has either the data controller or the editor role. If either of the condition has not been met, the relationship changes are stored as suggestions and



Figure 5.3: A user suggesting the homepage for Dr. Jonathan Borwein using bookmarklet.

are not reflected in the search results until the approval from either a data controller or an editor.

These suggestions serve two purposes. First, the suggestion helps identify the incorrect association that has been missed by the editors and the data controllers, as well as supplements the information regarding the researcher. Second, because the suggestions are immediately available to data controllers, these can be incorporated into the database of data controller's system, if it exists, encouraging greater participation from them.

#### 5.4 User Interface - Search and Profile Pages in MathPeople

As discussed in Section 5.3, MathPeople supports number of the user interfaces for data management. In additions to these, MathPeople provides a search engine and profile page for presenting information about researchers to the end-users.

MathPeople's search engine incorporates a full-text search engine. To implement

the search engine, we used the `Acts_as_solr` plugin <sup>1</sup>, a Ruby on Rails plug-in that provides full-text search engine capability to a database-backed Ruby on Rails projects. A user does not have to enter the full name of a researcher with a full-text search engine. Instead, the user can query for a researcher using a part of the researcher's name. MathPeople's search engine is case-insensitive and each word in a query constitutes a search term.

A user can construct a search term using any of five different search techniques. One technique is the exact word search. An example of the exact word searching would be when a user searches for all Smiths in the system by querying Smith. Another technique is the wildcard character search. A useful case for wildcard character search is when a user knows the given name of the researcher the user is looking for is either Jon or John. In this case, the user can enter `Jo*n` as a query term. The third technique is the field-specific search. If a user wants to find all researchers whose first name is James, the user can submit `firstname:james` to MathPeople's search engine. The fourth technique is fuzzy search. When a user don't remember any of the names of the researcher, but remember what it sounded like, the user can use the fuzzy search. The fuzzy search is performed when tilde is appended to a search term. The last technique is the Boolean search. A user can perform ANDed search by inserting word "AND" or "+" sign between search terms. To perform ORed search, and user can insert word OR between search terms. When a user prepend a search term with word "NOT" or "~" sign, MathPeople will return all the researchers whose name doesn't contain the search term. A user can also group search terms using parenthesis.

One challenge of implementing a search engine that searches for people's name is the character encoding issue. For researchers originating from non-English speaking countries, their names may contain diacritical characters, such as the umlaut. To address this issue, MathPeople added two features. In MathPeople, all data, including

---

<sup>1</sup><http://acts-as-solr.rubyforge.org/>

Diacritics	Non-diacritical translation made by MathPeople/Populi Scientiae
À, Á, Â, Ã, Ä	A
Å	AE, A
È, É, Ê, Ë	E
Ö	OE, O
Ò, Ó, Ô, Õ, Ø	O
Û, Ú, Û	U
Ü	UE, U
Æ	AE
ß	ss
Ñ	NY
ç	c
Ý	Y


Figure 5.4: If a name of a researcher contains a character on the left column, a user can search for the researcher by entering the character on the corresponding right column in place of the character. The translation is based on <http://ar.sky.ru/2.html>.

names of the researchers, are stored as UTF-8 characters in the database. This not only allows diacritical and non-roman alphabetical character to be displayed correctly on the web browser, it also allows users to enter diacritical or non-roman alphabetical character, as a part of a search term. Another feature is non-diacritical character translations of diacritical characters. Many diacritical characters can be translated into non-diacritical characters, but this relationship is not necessarily one-to-one. For example, ö is commonly translated as “oe” but it could also be translated into an “o.” To address this, MathPeople stores translated versions of names, along with original names, for all names with diacritical characters. Figure 5.4 lists the non-diacritical characters that each diacritical character is translated into.

MathPeople presents results in a paginated list (see Figure ??). Each entry in the result list contains the canonical name as a header and a sub-list. The sub-list displays all the records associated with the entry as a pair of data source identifier and the name associated with the record. The canonical name links to the profile page.



# Mathpeople



---

« Previous **1** 2 Next »

**Borwein, Peter**

- [Borwein, P \(CIS:43711\)](#)
- [Borwein, P. B \(CIS:45509\)](#)
- [Borwein, Peter B \(MR:39835\)](#)
- [Borwein, Peter \(ArXiv:Peter-Borwein-P\\*\)](#)
- [Borwein, Peter B \(Erdos:\)](#)
- [Borwein, Peter Benjamin \(MG:17637\)](#)
- [Borwein, Peter \(WikipediaMath:Peter\\_Borwein\)](#)

**Borwein, David**

- [Borwein, D \(CIS:45510\)](#)
- [Borwein, David \(MR:39825\)](#)
- [Borwein, David \(ArXiv:Peter-Borwein-D\\*\)](#)
- [Borwein, David \(Erdos:\)](#)
- [Borwein, David \(MG:42224\)](#)
- [Borwein, David \(WikipediaMath:David\\_Borwein\)](#)

« Previous **1** 2 Next »

**Merge**

**Search**

Borwein

Search only in first name  
"firstname:jim" or last name  
"lastname:wickham". Find  
misspellings with "pitnam-", or  
use wildcards "smit\*", "pitman  
j\*". See [Lucene syntax reference](#)  
for more options.

---

© IMS. [Log in](#)

Figure 5.5: MathPeople’s result page for the query “Borwein”

The profile page of MathPeople, an example of which is shown in Figure ?? consists of three sections. These are the name section, the record section, and the sidebar. In the name section, the canonical name is printed on top. All other names found from the chosen researcher's records are listed in a line under the canonical name.


The record section contains a list of data source aliases. The data sources listed in this section are those sources that contain records about this researcher. Each data source alias is listed with record IDs, where each record ID is the link to the webpage found with the record. A record description may be printed under the line containing data source alias.

The sidebar contains the links to Google and Google Scholar. These links returns the result page from searching the canonical name of the researcher.

Also on the side bar is the option to view the profile in XML or RSS format. These formats are added to facilitate interoperability among websites run by data sources. A direct link to this profile page can be made using data source identifier. By default, the last component of the profile page URL is the person ID assigned by MathPeople. However, a link to profile can be generated by replacing person ID in the profile URL with lookup/data source alias/record ID. In combination of the link by data source identifier and XML or RSS view, any data source can extract data from MathPeople and use it to disambiguate their data.

## 5.5 Discussion

As is the case with the FWDM, the primary advantage of MathPeople is that it provides effective data control. Because it relies on data controllers, who are authorized by individual researchers, as a primary provider of information, we are reasonably assured that data provided by these users is accurate and that it is provided with consent from individual researchers. Furthermore, because all of the researchers represented in MathPeople are associated with at least one data controller, each researcher

**Mathpeople**


---

## Borwein, Peter

([Borwein, P.](#), [Borwein, P. B.](#), [Borwein, Peter](#), [Borwein, Peter B.](#), [Borwein, Peter Benjamin](#))

**Records**

- CIS: [43711](#), [45509](#)
- Erdos:
  - Erdos Number 2
- WikipediaMath: [Peter\\_Borwein](#)
  - (Canada, 1953 - )
- MG: [17637](#)
  - 862:1979 Thesis:Rational Approximations
- ArXiv: [Borwein-P\\*](#)
- MR: [39835](#)

(Other formats: [xml](#), [rss](#))

**Picture**

[View Pictures](#)

[Add Picture](#)

**Search**

[Google](#)

[Google scholar](#)

**Modifications**

This is version 0.

---

© IMS. [Log in](#)

Figure 5.6: The Peter Borwein's profile page in MathPeople

can update inaccurate data through a data controller.

MathPeople, in addition to fulfilling data control requirement, has a number of advantages over the personal search engines discussed in Chapter 3. By separating disambiguation module from data acquisition and maintenance module, MathPeople allows for the bulk uploading of data. This separation also means that you can make changes to the name disambiguation module without affecting rest of the system. This could potentially open up the system to incorporate some sort of a computer-assisted or unsupervised name disambiguation.

Another benefit of this system is that we do not need to authenticate all the researchers listed in the system. Because all privacy issues regarding consent are met by each data source, MathPeople only needs to authenticate relatively small number of data controllers. Because all data comes from trusted data sources, the quality and trustworthiness of records are assured. With this system, the likelihood of malicious or spamming webpages being added to the system is minimized.

Finally, because the record matching is done by the editors, who are known to be trustworthy, MathPeople can give certain amount of assurance as to the accuracy of the disambiguation.

Although there are many benefits, MathPeople's design also has some drawbacks. Because it must accept data from trusted data sources that are known to comply with privacy laws, it inherently limits the number of possible data sources. Furthermore, by delegating all user-interaction to the data sources and requiring the records to be in certain format, it may discourage potential data source due to overhead that is required.

Another drawback of delegating control of data over to the third party data source is that we need to actively maintain data. If the data source does not update information in a timely manner, the records can easily contain information that is out of date.

There are drawbacks in the name disambiguation module as well. Because the actual disambiguation is done by a small number of people in a closed system, there is limit as to how fast records are associated with a person. The limit on the number of people responsible for disambiguation could hamper how well the system can handle large amount of contributions from unauthenticated users. Also, as MathPeople allows the merge of people objects, this may result in the person ID maintained by the data controller become out of synch if there is an error in name disambiguation.

Finally, because a researcher can't actually change the information within the system, the only way for a researcher to opt-out is contacting the maintainers for all of the relevant data sources. The fact that there is not a direct way to opt out of the system may reduce the attractiveness of the system to users and data providers.

## Chapter 6

### Populi Scientiae - The User-filtered Name-Authority System

In this chapter, we provide an in-depth discussion on the implementation of Populi Scientiae, a prototype user-filtered name-authority system focussed on providing biographical and research information on researchers in mathematical science fields. Populi Scientiae was hosted on a server maintained by D-Drive lab at Faculty of Computer Science, Dalhousie University. This deployment of Populi Scientiae was used to profile pages for most members of D-Drive where information on D-Drive members' past research as well as short biographical information was provided to general public.

This chapter is organized into the following sections. In Section 6.1, we provide an overview on the design of Populi Scientiae. In Section 6.2, we examine the different types of users defined by Populi Scientiae and how they are managed. In Section 6.3, we examine how Populi Scientiae manages contributions from different users based on the privilege granted to them. In Section 6.4, we examine the user interfaces for Populi Scientiae's search page and profile pages. Finally, in Section 6.5, we discuss the strengths and weaknesses of the implementation.

#### 6.1 System Overview

As discussed Section 4.4.2, the design of the user-filtered name-authority system is based on filtering the user contribution based on the identity of the user contributing instead of the content of the contribution, as is the case in the data-filtered name-authority system. In Populi Scientiae, this is realized by allowing each researcher to define what types of users he/she wants to trust for a given type of data on his/her

profile page. The overview of this design is shown in Figure 6.1. The database for Populi Scientiae is based on following five types of entities:

1. Person.
2. Resource.
3. Resource\_Blacklst.
4. User.
5. Right.

In Populi Scientiae, the access control is managed through the *user* and the *right* entities. As in MathPeople, an instance of the user entity represents an authenticated user, containing the associated username and the encrypted password. The difference between a user in Populi Scientiae and MathPeople is that, instead of being assigned a role as in MathPeople, an user in Populi Scientiae represents a researcher and the types of task the user can do is defined by the relationships to other researchers. Due to this difference, the *right* entity in Populi Scientiae is significantly different from the *right* entity in MathPeople. The *right* entity in Populi Scientiae contains the trust level and data type it is associated with, along with the IDs of the person and the user it is associated with. We will discuss the trust level and data type in Section 6.2.2.

The management of the information on the researchers, denoted as the *profile management* in Figures 6.1, is accomplished through *person*, *resource*, and *resource\_blacklist* entities. As it is the case with MathPeople, the information regarding a researcher is encapsulated through the *person* entity. In Populi Scientiae, each instance of Person entity is associated with a user. As well, it contains the “canonical” names. The resource entity captures the data currently associated to researchers. Because the data is maintained directly by the researchers, Populi Scientiae allows larger number of data types to be entered, compare to MathPeople. Populi Scientiae also

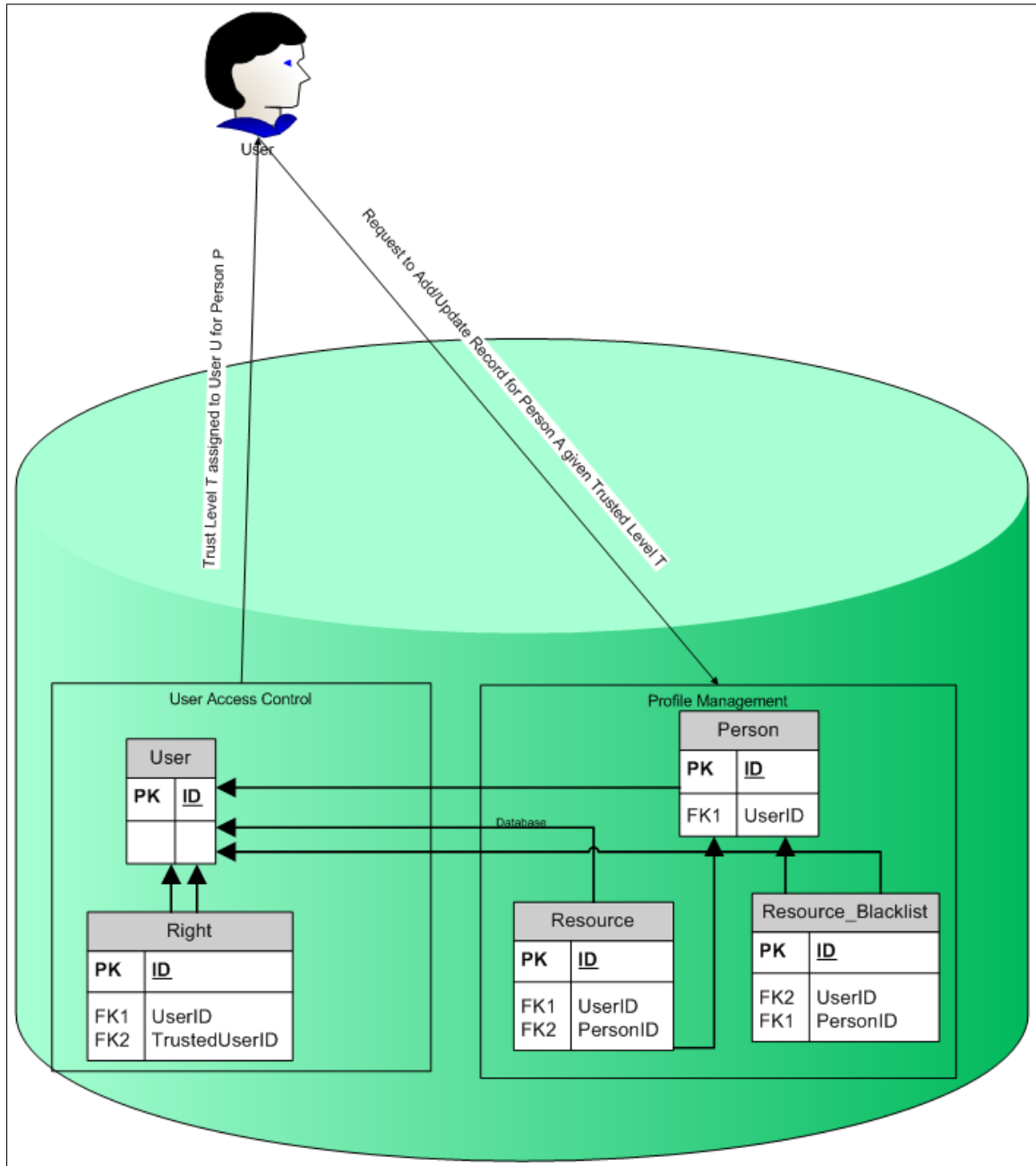


Figure 6.1: The overview diagram of how the user's contribution interacts with the database. The ER-diagram for the database only shows the redacted view of the database, with Resource and Resource.Blacklist entity constituting amalgamation of multiple tables.



keeps track of resources that have been disassociated from researchers. This is done through the `resource_blacklist` entity. Data management will be further discussed in Section 6.3.

## 6.2 User Management

In Populi Scientiae, we introduce a trust-based access control for the user management. Each researcher in Populi Scientiae defines how much he/she trusts a user. Based on this trust level, Populi Scientiae determines whether a user can contribute data to a profile page.

Two key aspects of the trust-based access control are *the authentication and verification* and *per-profile based trusted user and confidence level maintenance*. Populi Scientiae needs to ensure that each profile is actually assigned to the said researcher. As well, Populi Scientiae needs to ensure that each researcher can configure the access control for his/her profile page.

### 6.2.1 Account Creation

As discussed in Section 2.5, many of the professional societies and post-secondary institution provides directories of member researchers in the mathematical science fields. Populi Scientiae's leverages off of the email address lists made available through these directories to authenticate user during the account creation process.

The creation of the account for a researcher in Populi Scientiae involves four stages.

1. submission of the user information,
2. verification of the user information,
3. extraction of the identification information and profile creation, and
4. user account creation confirmation.

In the submission of the user information stage, Populi Scientiae collects the information from a researcher through the user account creation form. The information collected at this phase includes the personal name, the email address, and the name of the institution or the society whose membership directory contains the given email address, along with the new username and the password.

Populi Scientiae searches for the user from the membership directory using the information collected from the user. The email address is used to confirm that the researcher is a member of the given organization. Once the identity of the researcher has been confirmed, Populi Scientiae extracts the name of the researcher from the membership directory and uses this name as canonical name for the new profile (i.e. an instance of person) that it generated.

Once the researcher has been verified and the profile has been generated, the confirmation email is sent to the user. This email contains the activation link, a link to the page that needs to be visited by the researcher so he/she can be authenticated and the profile activated.

In the Populi Scientiae prototype, the authentication is done using the Dalhousie Computer Science email search engine. To create a profile page in Populi Scientiae, the researcher supplies the Dal CS email address, along with the new username and password he/she wants to use to login. The given email address is checked against the Dal CS email search engine, and if the email address is found, Populi Scientiae authentication system extracts name and position from the search results. After a successful search, Populi Scientiae sends a confirmation to the researcher, which contains a link to a verification webpage. When the researcher visits the verification webpage, a new profile page containing name and position extracted from email search is generated and user is redirected to the newly-generate profile page.

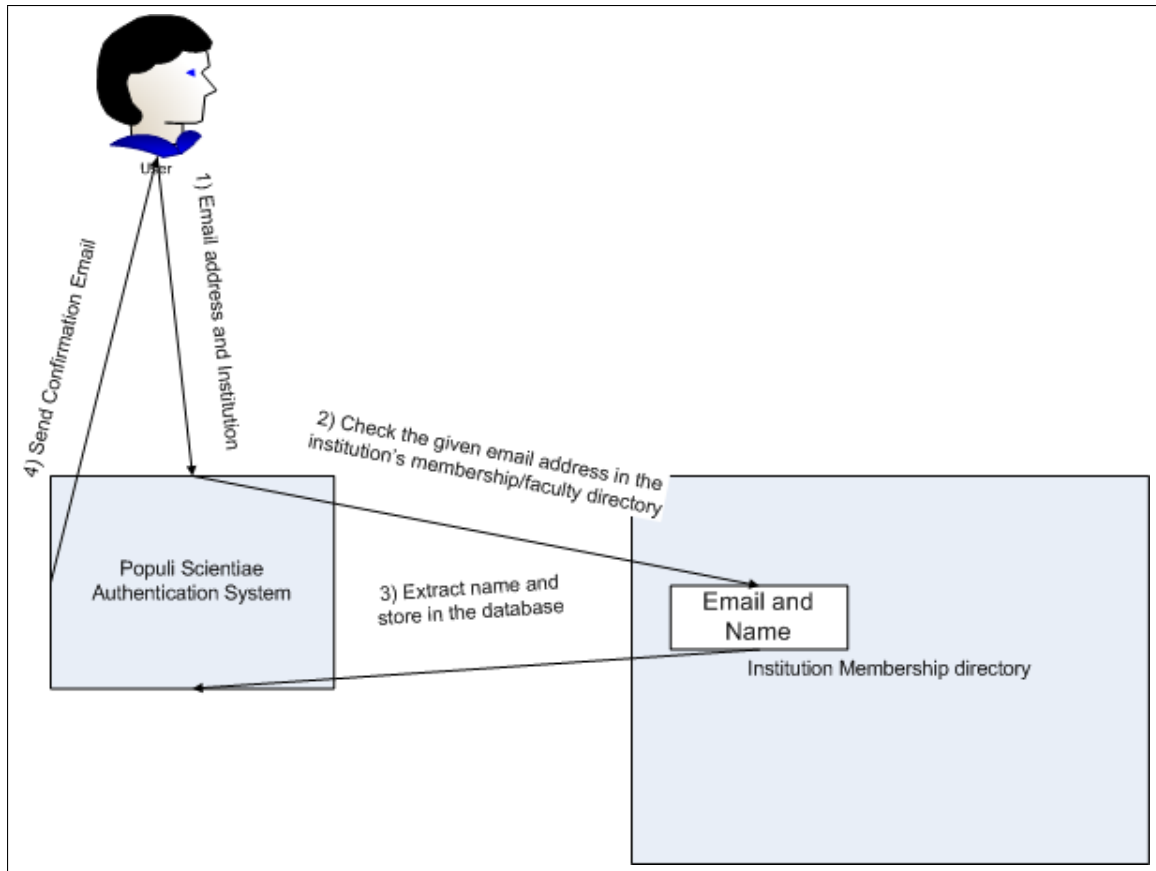


Figure 6.2: The overview of the four-stage authentication process in Populi Scientiae account creation

The screenshot shows the sign-up page for PopuliScientiae. The page includes a login section at the top right and a sign-up section at the bottom left.

**PopuliScientiae**

**Login**

Login

Password

[Forgot Password? or Signup](#)

---

**Sign up**

Login

Dalhousie CS Email

Password

Confirm Password

[Change Required Privilege Level](#)  
[Edit Trusted Users](#)  
[Change Password](#)

Figure 6.3: The sign-up page for Populi Scientiae: New users can register by providing an email address, new log-in name, and new password

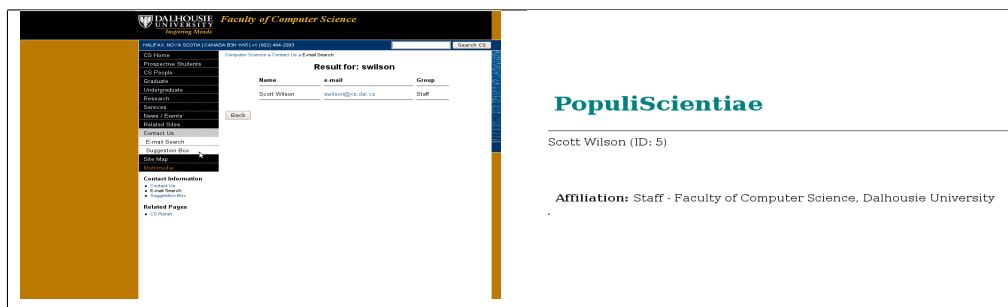


Figure 6.4: The Successful Registration to Populi Scientiae: When a new user provides a valid email, the values from Name and Group fields are extracted from the Dalhousie CS Email Search’s result page (see on the left). The extracted value are then appended to the new users profile page (see on the right) once the user confirms the registration.

## 6.2.2 Trusted Users and Confidence Level

Populi Scientiae introduces the concept of confidence level to enable each researcher to control user contribution to his or her profile page. The confidence-level based access control is based on three principles. The first principle is that every user visiting a profile page is assigned confidence level. The second principle is that any section of a profile page whose content comes from user contribution must have a minimum confidence level associate with it, and that only users with confidence levels that are equal or greater than minimum confidence level can contribute to that section. The final principle is that a user can change a contribution made by another user if and only if the user’s confidence level is equal or higher than the confidence level of the user who made the contribution.

A user visiting a profile page is assigned with one of the four confidence level in the current prototype.

- If the user has not logged in or does not have an account, he/she is assigned the confidence level of -1.
- If the user does not own the profile page and is not one of the trusted users, he/she is assigned the confidence level of 0.

- If the user is one of the trusted users, he/she is assigned the confidence level of 1.
- If the user owns the profile page, he/she is assigned the confidence level of 2.

Further discussion on how the confidence level is set is provided in Section 6.3.

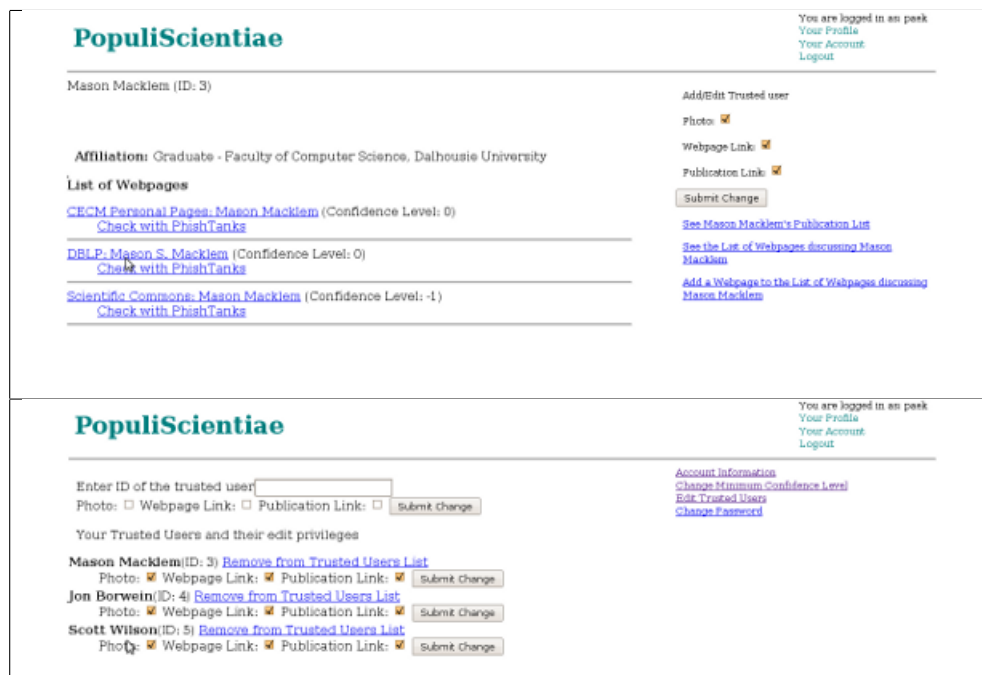


Figure 6.5: A researcher can add or remove a user from trusted user list by visiting the profile page of the user in question (seen on the left) or through the researcher's Account Setting page (seen on the right).

### 6.3 Data Management

The basic concept behind the data management in Populi Scientiae is as follows: a user can add or modify a resource if the confidence level assigned to the resource is not greater than the confidence level of the user. For a resource that has not been associated to the given profile, it is assigned with the minimum confidence level that has been either set by the researcher who owns the profile or the system. The algorithm for managing user contributions is detailed in Algorithm 6.3.

---

**Algorithm 1** Algorithm for inserting or updating data in Populi Scientiae

---

**Input:** a user  $U$ , a profile  $P$ , and a resource  $R$  that  $U$  attempts to load into  $P$ . Also used are auxiliary variables  $PR$  (resource in  $P$ 's resource set that matches data in  $R$ )

**Output:** Updated  $P.resourceSet$

{ If  $R$  is already in the  $P$ 's resource set, compare the confidence level of  $P$  against the confidence level assigned to  $R$  }

- 1: **if**  $R$  in  $P.resourceSet$  **then**
- 2:   **if**  $PR.confidenceLevel \leq P.ConfidenceLevel(U)$  **then**
- 3:     update  $PR$  with data in  $R$
- 4:      $PR.confidenceLevel = P.ConfidenceLevel(U)$
- 5:   **end if**
- { If  $R$  is not in  $P$ 's resource set, compare the confidence level of  $P$  against the minimum confidence level of  $R$  }
- 6: **else if**  $P.MinimumConfidenceLevel(.TypeOf(R)) \leq P.ConfidenceLevel(U)$  **then**
- 7:   add  $R$  to  $P.resourceSet$
- 8:    $R.confidenceLevel = P.ConfidenceLevel(U)$
- 9: **end if**

---

In current prototype of Populi Scientiae, we define seven different types of data. These are:

1. personal names, including the canonical name,
2. current position or current affiliation of the researcher,
3. photographs, including the profile picture,
4. short description of the researcher,
5. current location of the researcher,
6. links to webpages that contain information about the researcher, and
7. links to Google Scholar entries for publications authored or co-authored by the researcher.

For some types of data, Populi Scientiae imposes minimum confidence level for all the profile pages in order to ensure that the control over information that can be construed as sensitive remains with the data subject.

The canonical name and current position/affiliation information are assigned with confidence level of “3,” which means it can only be set by Populi Scientiae system. Both the canonical name and the current position/affiliation are set by the authentication system discussed in Section 6.2.1. This measure has been implemented to safe-guard against the researcher assuming an identity of another researcher. In the current prototype, this can only be set at user account creation, as it only incorporates the Computer Science faculty at Dalhousie University. However, the same authentication system can easily extend to update this information as other membership directories are incorporated.

The minimum confidence levels for the publication of photographs, including the profile picture, the personal description, and the current location of the researcher are set to “2,” restricting the contribution to the owner of the profile page. Our determination is that this information has sufficient sensitivity to warrant limiting the privilege to modify these data to the owner of the profile.

For all other types of information, including web links and Google Scholar links, the owner of the profile page sets the minimum confidence level. This includes the picture suggestion. Although, picture can only be published by the owner, the user can upload pictures onto profile as a suggestion for publication. The suggested pictures can only be accessed by the owner of the profile page.

A researcher who has registered with Populi Scientiae can set the minimum confidence levels through the Account Setting page (see Figure ?? for an example). The subsections for which a user can set the minimum confidence levels are the photograph suggestion, the web page list, and the publication list.

We demonstrate how the user can contribute data to a profile page in Section 6.4.

The screenshot shows the 'PopuliScientiae' account settings page. At the top right, it indicates the user is logged in as 'paek' and provides links for 'Your Profile', 'Your Account', and 'Logout'. Below this, there are links for 'Account Information', 'Change Minimum Confidence Level', 'Edit Trusted Users', and 'Change Password'. The main section is titled 'Minimum Confidence Level' and contains three subsections, each with a 'Change' button and radio button options:

- Adding Publication:** Unregistered User  Registered users  Trusted users and account holder  [Change]
- Adding Web Page Link:** Unregistered User  Registered users  Trusted users and account holder  [Change]
- Suggesting Photo:** Unregistered User  Registered users  Trusted users  [Change]

Figure 6.6: A user of Populi Scientiaie can adjust minimum confidence level for each of three profile subsections through Account Setting page

#### 6.4 User Interface - Search and Profile Pages in Populi Scientiaie

Like MathPeople, Populi Scientiaie supports search engine and profile page. Since the data is richer and how the user can contribute differs, the UI looks quite different.

The same search engine employed for MathPeople has been incorporated into Populi Scientiaie. However, because it is a user-driven system, it displays more information on the result page, which helps the users to easily identify the researchers that they are looking for.

In Populi Scientiaie, each entry contains at minimum the canonical name, and the current position at Dalhousie University's Faculty of Computer Science. Additionally, if the researcher has specified a profile photograph, this picture will be displayed along with the search result. An example of Populi Scientiaie's result page is shown in Figure ??.

Also, each researcher in Populi Scientiaie is associated with an ID number. This ID number can be used to directly link to profile page, photograph gallery, publication list and web page list of a researcher.

The detailed information on each researcher registered with Populi Scientiaie is



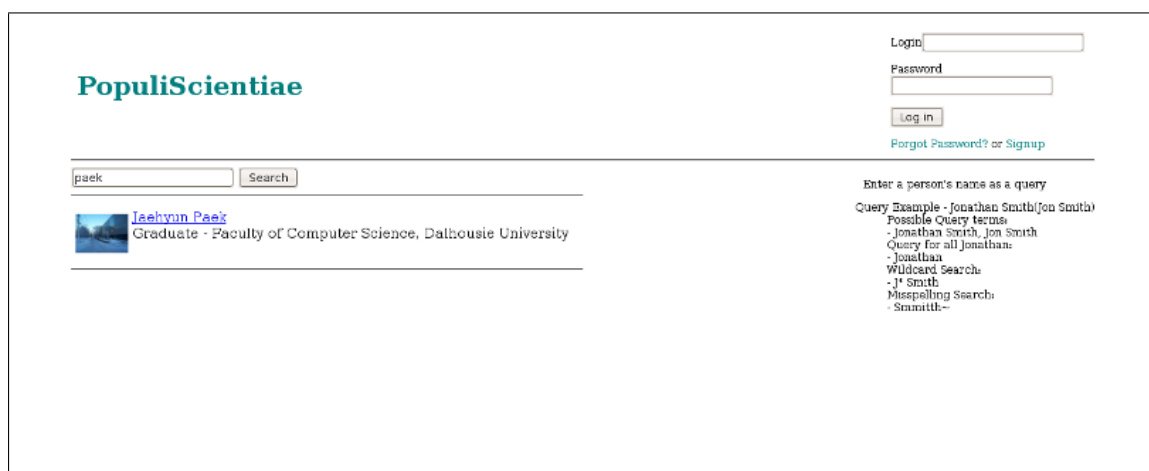


Figure 6.7: A Populi Scientiae page displaying search result.

provided through the Profile page (see Figure ??). The main profile page, the page linked by the result page, consists of two sections. The main window provides a complete description on the researcher, as well as abridged lists of web pages related to researchers and of publications authored by the researcher. The sidebar on the right may contain following links: the publication list, the interface to Google Scholar, the list of web pages, the form to add a new web page link, and the photo gallery.

The publication list contains links to publications that were authored by the researcher. To add a publication, click on the “Add a Publication to ...” link in the sidebar. This will open a Google Scholar interface generated by Populi Scientiae. Enter a search term in the search box, and click on either “Add” to add a link to the publication list or “Remove” to remove from the list.

The web page list contains links that provide some information on the researcher’s research activity. As a precaution against a possible phishing site entered by a malicious user, each link on the list has a mechanism to check against PhishTank (an open database of suspected phishing sites) in real-time. To check, simply click on the link called “Check with PhishTank,” and to add a link, click on the “Add a Webpage...” link on the sidebar and follow the instruction.

If the owner of the profile has added any pictures, a link to photo gallery will be

The screenshot shows a web browser window displaying a profile page on PopuliScientiae. At the top right, there is a login form with fields for 'Login' and 'Password', a 'Log in' button, and a link for 'Forgot Password? or Signup'. The main header features the site name 'PopuliScientiae' in a teal font. Below this, the profile information for Jaehyun Paek (ID: 1) is shown, including his name in English and Korean, and his affiliation with Dalhousie University. A 'Personal Description' section identifies him as the creator of PopuliScientiae and a Master's student. To the left, there is a photo gallery titled 'Picture of Jaehyun Paek' with a single image of a snowy street and a link to 'See more photos of Jaehyun Paek'. To the right, a map titled 'Current Location of Jaehyun Paek' shows a street view of Dalhousie University. At the bottom, there are links for 'Jaehyun Paek', 'Check with PhishTanks', 'Jaehyun Paek's Dal CS profile page', and 'Jaehyun Paek - Homansoo'.

Figure 6.8: A Populi Scientiae Profile Page

listed on the sidebar. If the user has permission to suggest a picture to the researcher, there will be a form to suggest a picture at the bottom of the photo gallery.

When researchers are visiting their own profile pages, they can make further edits to the profile page. A researcher can add a pseudonym or an alias to the list of names by clicking the “Add a name ...” link near the top of the profile page and entering a name into the text box that appears (see Figure ??). To edit the personal description, a researcher simply has to click on the section in that he or she wants to edit and that area becomes editable text box. A researcher can also add his or her current location by editing the current location provided by the Google Maplet interface. A researcher can also approve or delete a picture, as well as designate a photo as a profile picture, in the photo gallery page (see Figure ??).

PopuliScientiae

You are logged in as: paek  
[Your Profile](#)  
[Your Account](#)  
[Logout](#)

---


Jaehyun Paek (ID: 1)  
[Jay Paek](#) [Remove Name](#) [Jae Hyun Paek](#) [Remove Name](#)  
[백재현](#) [Remove Name](#) [白才鉉](#) [Remove Name](#)  
[Add a Name to the List of Jaehyun Paek's Pseudonyms](#)

**Affiliation:** Graduate - Faculty of Computer Science, Dalhousie University

**Personal Description (Click on Description to Edit):**

Creator of PopuliScientiae  
Master's Student

Picture of Jaehyun Paek



[See more photos of Jaehyun Paek](#)

**List of Webpages**

[Jaehyun Paek - Homepage](#) (Confidence Level: 2)  
[Check with PhishTanks](#)

---

[Jaehyun Paek - Dalhousie Homepage](#) (Confidence Level: 2)  
[Check with PhishTanks](#)

---

[LinkedIn: Jaehyun Paek](#) (Confidence Level: 2)  
[Check with PhishTanks](#)

---

[Scientific Commons: Jaehyun Paek](#) (Confidence Level: -1)  
[Check with PhishTanks](#)


---

[Jaehyun Paek](#) (Confidence Level: 2)  
[Check with PhishTanks](#)

---

Current Location of Jaehyun Paek

Map    Satellite    Hybrid



6050 University Ave, Halifax    Change Location  
[Submit Mapplet Location](#)

[See Jaehyun Paek's Publication List](#)

[Add a Publication to Jaehyun Paek's Publication List](#)

[See the List of Webpages discussing Jaehyun Paek](#)

[Add a Webpage to the List of Webpages discussing Jaehyun Paek](#)

[See Photographs of Jaehyun Paek](#)

Figure 6.9: A researcher visiting his Populi Scientiae profile page: The researcher can edit the personal description by clicking the area highlighted with yellow.


PopuliScientiae

You are logged in as: paek  
[Your Profile](#)  
[Your Account](#)  
[Logout](#)

---

[Go back to Profile](#)


## Pictures of Jaehyun Paek



Winter 1

Set as Non-profile

Delete



Winter 2

Set as Profile

Delete

Title of the picture:

Upload A Picture:

[Back to Jaehyun Paek's Profile](#)

[See Jaehyun Paek's Publication List](#)

[Add a Publication to Jaehyun Paek's Publication List](#)

[See the List of Webpages discussing Jaehyun Paek](#)

[Add a Webpage to the List of Webpages discussing Jaehyun Paek](#)

[See Photographs of Jaehyun Paek](#)

Figure 6.10: A researcher visiting his own photo gallery in Populi Scientiae

## 6.5 Discussion

By implementing the user-filtered name-authority system, Populi Scientiae eliminates many administrative and privacy constraints imposed on MathPeople. One advantage of Populi Scientiae is that it provides greater privacy control to the users. To correct or remove information in MathPeople, a user has to contact each data source that contributed the incorrect or undesired data. MathPeople may require each data source to maintain list of users who have opted out of MathPeople. However, because Populi Scientiae gives users full control over their profile page, users can directly correct their profile, and even delete the profile page altogether.

Another advantage of Populi Scientiae is that it is not restricted by data source when getting information. MathPeople were able to only obtain information from a source that has provided proper level privacy protection and is willing to share data with MathPeople. In contrast, Populi Scientiae can potentially include personal information with low sensitivity from any data source, no matter what the privacy policy is at the said data source. This is because Populi Scientiae solicits consent to release low-sensitivity personal information from each user who creates a profile page. This also means that Populi Scientiae does not need to have an agreement from each data source as long as information it accepts does not violate any copyright laws.

Also, because the information is added directly to a profile page by users, Populi Scientiae eliminates need for disambiguation module. In MathPeople, each record sent from a data source needs to be matched to a person. Because of this, MathPeople has to depend on availability of large number trusted users in order to present records in timely manner. By letting users add information directly to the profile, Populi Scientiae eliminates both of these requirements.

Finally, Populi Scientiae presents more variety of information than MathPeople.

By depending solely on participating online data sources, MathPeople is largely restricted to presenting web links to users. However, because it operates with direct consent from users, Populi Scientiae can incorporate additional information, such as biographical details, photographs, and current contact information.

However, Populi Scientiae has some drawbacks compared to MathPeople. One of the drawbacks of Populi Scientiae is that the amount of information it can provide depends on the user base. As it can obtain the bulk of the records from various data sources, MathPeople can present substantial amount of information on researchers from early stage of its development, even when it does not have a large user base. As Populi Scientiae can obtain information on researchers from users, it can only host substantial amount of contents when it has a large and active user base. This is a problem for Populi Scientiae in terms of number of researchers it can provide information on as it can only provide information who has created a profile page. As well, because each user needs to authenticate them, the user who can join the system is limited.

Another drawback for Populi Scientiae is that it cannot provide the same confidence as the data-filtered solution such as MathPeople. Because MathPeople accepts records from reputable data sources, such as MathSciNet, the information it provides has the same credibility as the data source from which the data came from. Furthermore, the confidence level of the system is dependent on how secure the authenticated system is. However, Populi Scientiae cannot give the same credibility as MathPeople due to the fact that its contents comes from users. Because the authentication system is based on the email directory, it would be difficult for a user to create a profile based on someone else's identity without hacking into the email account. However, it could be easy for the owner of a profile page or for a party trusted by the owner to insert inaccurate information to profile page. Also, Populi Scientiae can contain some links to harmful or purely commercial websites, a possibility that is remote in MathPeople.

Lastly, Populi Scientiae suffers from lack of interoperability of with other data sources. In MathPeople, each profile page can be linked using the ID number assigned by each data source that has contributed a record to the said profile page. This can be a useful way for a third-party to disambiguate their results, as well as a way to provide additional information to the users, which could encourage other data sources to contribute data. However, because users can add information from online and offline source to Populi Scientiae, it would be much more difficult to provide such a service in Populi Scientiae. In Populi Scientiae, data elements must be added one-by-one. In contrast, data sources can bulk upload data elements in MathPeople.

## Chapter 7

### Conclusion and Future Works

Today, a vast amount of resources for academic researchers is available online. Some of these resources are generated by researchers or by research institutions. However, many are generated by independent parties without the consent from researchers. When looking for information on a researcher, it is often difficult to determine whether content found on a webpage about the researcher one was searching for is current. Even if one can determine the identity of the researcher, it may be difficult to discern the accuracy of the information presented. Given this climate, it is important to develop systems that will help to authoritatively identify high quality resources and group them based on the researchers' identities. In this thesis, we investigated the problem of making the correct association between a researcher and some online information on a researcher, and presented a novel solution to this problem.

We have proposed a public name authority system as a way to authoritatively match web resources to researchers. Our proposed solution is to implement a centralized repository of data, called a public name authority system, where data that are approved by the researchers are collected and disambiguated by trusted users. By collecting only approved data and allowing only trusted users to disambiguate data, we can ensure the accuracy and the quality of data maintained.

As a result of our research, we have developed two different designs for the public name authority systems. The first - the data-filtered name-authority system, realized through the prototype called MathPeople, incorporates the suggestions from both

the authenticated and unauthenticated users where the contributions from the unauthenticated users are filtered by the authenticated users based on the content of the data. Math People provides high-confidence cross-matching of these resources, which can be used by each participating data source as a name disambiguation tool. This system presupposes existence and participation of numerous data sources that have permission from researchers to collect and share personal information.

The second name-authority system design, the user-filtered name-authority system filters data based on the trust that a researcher has on a user who wants to contribute data against the researcher's profile page. This design was realized through the Populi Scientiae prototype. In Populi Scientiae, each authenticated researcher gives trusted users permission to add data to his or her profile page. As each authenticated researcher administers his or her profile page, the users of the system can have confidence in the accuracy of data. Furthermore, Populi Scientiae has potential to contain more data than MathPeople as it is bound neither to participating data sources nor to the data sources complying with the privacy laws in their jurisdiction. However, as Populi Scientiae is user-driven system, the effectiveness of the system depends on the size of the user base.

## 7.1 Future Work

The two prototype name-authority systems proposed in this thesis, MathPeople and Populi Scientiae, have been focussed on addressing name disambiguation and data control problems pertaining to information on researchers in Mathematical Science fields. This was due to availability of such resources as

- MathSciNet - an author identification tool that identifies previous publication of a Mathematician with high reliability,
- arXiv.org - an open access server providing a significant number of high quality



research materials,

- membership directories from various Mathematical and Statistical societies, such as the Canadian Mathematical Society, and
- numerous other well-maintained web sites providing biographical and research information on Mathematicians.

However, the design alternatives for name-authority system proposed in this thesis can be replicated and extended for serving researchers in other fields of study. For the implementation of a data-filtered name-authority system, only requirements are the availability of suitable name authority data sources and the interest of suitable agents in validating and curating that data. For the implementation of a user-filtered name-authority system, all that is required is the availability of a list of contact information, preferably email address, that has been verified highly reliable source, which include faculty listings provided by accredited post-secondary institutions.

The next step in this research is to implement a public name-authority system that combines the benefits of the data-source independent data acquisition scheme of Populi Scientiae with the user-independent data acquisition scheme of Math People. In order to ensure the privacy control for the data acquired, it may not be possible to implement a system that is both data-source independent and user independent. However, it may be possible to achieve a system that is fully independent of data source, and does not require large user base to acquire initial set of data. For instance, a public name-authority system can suggest a set of links from highly trusted data sources to a researcher when the researcher registers to create a profile page.

In addition, there are a number of interesting technical issues surrounding the implementation of a public name-authority system. Some of these issues include effective management of the security, the incorporation of the automated name disambiguation suggestion system, and the distribution of the system over a number of

web servers.

## Bibliography

- [1] D. L. Baumer, J. B. Earp, and J. C. Pointdexter. Internet privacy law: A comparison between the united stated and the european union. *Computer and Society*, 23:400–412, 2003.
- [2] Ron Bekkerman and Andrew McCallum. Disambiguating web appearances of people in a social network. In *WWW'05, World Wide Web Conference 2005*, pages 463–470, May 2005.
- [3] LEAF Project Consortium. Public progress report 2. <http://www.crxnet.com/leaf/>. accessed on 12/2010.
- [4] Michael Deturbide. personal communication.
- [5] Technology Directorate for Science, Industry Organization for Economic Co-operation, and Development. Oecd guidelines on the protection of privacy and transborder flows of personal data. [http://www.oecd.org/document/18/0/2C3343/2Cen\\_2649\\_34255\\_1815186\\_1\\_1\\_1\\_1/2C00.html](http://www.oecd.org/document/18/0/2C3343/2Cen_2649_34255_1815186_1_1_1_1/2C00.html). accessed on 12/2010.
- [6] Technology Directorate for Science, Industry Organization for Economic Co-operation, and Development. Protection of privacy and personal data. [http://www.oecd.org/document/26/0/2C3343/2Cen\\_2649\\_34255\\_1814170\\_1\\_1\\_1\\_1/2C00.html](http://www.oecd.org/document/26/0/2C3343/2Cen_2649_34255_1814170_1_1_1_1/2C00.html). accessed on 12/2010.
- [7] European Parliament and of the Council. Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML>, October 1995. accessed on 04/2009.
- [8] Association for Computing Machinery. Author profile pages in the acm digital library - overview. [http://www.acm.org/membership/author\\_pages](http://www.acm.org/membership/author_pages), 2009. accessed on 4/2009.
- [9] F. S. Grodzinsky and H. T. Tavani. Some ethical reflections on cyberstalking. *Computer and Society*, pages 193–200, March 2002.
- [10] Hui Han, Hongyuan Zha, Wei Xu, and C. Lee Giles. A hierarchical naive bayes mixture model for name disambiguation in author citations. In *ACM Symposium on Applied Computing*, pages 1065–1069, March 2005.
- [11] Masanori Harada, Shin ya Sato, and Kazuhiro Kazama. Finding authoritative people from the web. In *JCDL'04, Joint ACM/IEEE Conference on Digital Libraries*, pages 306–313, June 2004.

- [12] Dmitri V. Kalashnikov, Sharad Mehrotra, Zhaoqi Chen, Rabia Nuray-Turan, and Naveen Ashish. Disambiguation algorithm for people search on the web. In *ICDE'07, IEEE 23rd International Conference on Data Engineering 2007*, pages 1258–1260, April 2007.
- [13] A. Kobsa. Personalized hypermedia and international privacy. *Communications of the ACM*, 45(5), May 2002.
- [14] Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Conference on Natural Language Learning at HLT-NAACL 2003*, volume 4, pages 33–40, May 2003.
- [15] Duncan M. McRae-Spencer and Nigel R. Shadbolt. Also by the same author: Aktiveauthor, a citation graph approach to name disambiguation. In *JCDL'06, Joint ACM/IEEE Conference on Digital Libraries*, pages 53–54, June 2006.
- [16] Privacy Commissioner of Canada. The personal information protection and electronic documents act. [http://www.privcom.gc.ca/legislation/02\\\_06\\\_01\\\_e.asp](http://www.privcom.gc.ca/legislation/02\_06\_01\_e.asp). accessed on 4/2009.
- [17] V. Patel and R. Juric. Internet users and online privacy. In *23rd International Conference on Information Technology Interfaces*, pages 193–200, 2001.
- [18] P. Perri. Privay law - italy: Lessons learned from the italian law on privacy - part i. *Computer Law & Security Report*, 20(4), 2004.
- [19] American Mathematical Society. Uniquely identifying mathematical authors in the mathematical reviews database. <http://www.ams.org/publications/math-reviews/mr-authors>. accessed on 12/2010.
- [20] H. T. Tavani. Privacy online. *Computers and Society*, December 1999.
- [21] Xiaojun Wan, Jianfeng Ga, Mu Li, and Inggong Ding. Person resolution in person search results: Webhawk. In *CIKM'05, Conference on Information and Knowledge Management*, pages 163–170, November 2005.
- [22] Wikipedia. Wikipedia:consensus. <http://en.wikipedia.org/wiki/Wikipedia:Consensus>. accessed on 12/2010.
- [23] Kai-Hsiang Yang, Kun-Yan Chiou, Hahn-Ming Lee, and Jan-Ming Ho. Web appearance disambiguation of personal names based on network motif. In *WI'06, 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 386–389, December 2006.
- [24] G. Yee and L. Korba. Privacy policy compliance for web services. In *Proceedings of the IEEE International Conference on Web Services*, 2004.