

# Automating the Construction of Authority Files in Digital Libraries: A Case Study



J. C. French, A. L. Powell, E. Schulman\*, J. L. Pfaltz

Department of Computer Science, University of Virginia

\*National Radio Astronomy Observatory

# What are we trying to do?



- **Merge bibliographic records**
- **Search distributed collections**
- **Data mining**
- **Bibliometric studies**

# What is the problem?



## **Messy data!**

- variants
- misspellings
- acronyms
- abbreviations
- multiple languages
  - translation and transliteration problems

# But, what is THE problem?



**It's very difficult to tell when two entities are, in fact, the same entity.**

**For example:**

<b>French, James C.</b>	<b>(19)</b>
<b>French, James C.</b>	<b>(2)</b>
<b>French, James</b>	<b>(2)</b>
<b>French, J. C.</b>	<b>(2)</b>
<b>French, J.C.</b>	<b>(7)</b>
<b>French, J.</b>	<b>(2)</b>

# Our solution?



## **We use a combination of**

- data transforms**
- approximate string matching techniques**
- approximate word matching techniques**

**to cluster the data and generate authority files.**

# Authority Files



- **Authority work: detecting variant names for unique entities in the database.**
- **Authority files: files that maintain the correspondence between all allowable forms for strings in a particular bibliographic field.**

# A running example



## **Astrophysics Data System (ADS)**

- collection of bibliographic data, abstracts, and full text from astronomy and astrophysics**
- approximately 240,000 entries from over 1,000 journals and conference proceedings**
- our experiments are based on a subset of 146,000 journal articles**

# Raw affiliation strings for the University of Virginia

Affiliation string	Count
Univ. of Virgina, Charlottesville, VA, US	1
Univ. of Virginia, Charlottesvill, VA, US	1
Univ. of Virginia, Charlottesville, VA, US	44
Univ. of Virginia, Charlottsville, VA, US	1
Univ. of Virginia, VA, US	1
University of VA., Charlottesville	1
University of Virginia, Charlottesville, VA, US	23
University of Virginia, Virginia, US	1
Virgina Univ., Charlottesville, VA, US	1
Virgina, University, Charlottesville, VA	1
Virginia Univ.	2
Virginia Univ., Charlottesville	58
Virginia Univ., Charlottesville, VA	1
Virginia Univ., Charlottesville, VA, US	4
Virginia University, Charlottesville	1
Virginia University, Charlottesville, VA	1
Virginia, University	57
Virginia, University, Charlottesville	204
Virginia, University, Charlottesville, VA	77
Virginia, University, Charlottesville, Va.	83

Cleanup Method	Number of Distinct Affils.	Δ
None	20168	
Remove US, U.S.A., etc. if occurring at the end of an affiliation	19868	300
Remove US ZIP codes from end of affiliations	19850	18
Remove US state abbrevs. occurring at the end of an affiliation	18850	1000
Expand most obvious abbreviations (University, Institute, etc.)	17773	1077
Expand other selected abbreviations and acronyms	17598	175
Remove country names occurring at the end of an affiliation	16427	1171

## Affiliation string

University of VA., Charlottesville  
University of Virginia, Charlottesville  
University of Virginia  
University of Virginia, Charlottesville  
University of Virginia, Charlottesville  
    Univ. of Virginia, Charlottesville, VA, US  
    University of Virginia, Charlottesville, VA, US  
University of Virginia, Charlottesville  
University of Virginia, Virginia  
Virginia University, Charlottesville  
Virginia, University, Charlottesville  
Virginia University  
Virginia University, Charlottesville  
    Virginia University, Charlottesville  
    Virginia Univ., Charlottesville  
    Virginia Univ., Charlottesville, VA  
    Virginia Univ., Charlottesville, VA, US  
    Virginia University, Charlottesville, VA  
Virginia, University  
Virginia, University, Charlottesville  
    Virginia, University, Charlottesville  
    Virginia, University, Charlottesville, VA  
    Virginia, University, Charlottesville, Va.

# Edit Distance



**The edit distance,  $e(u, v)$ , from a string  $u$  to a string  $v$  is the minimum number of simple edit operations (insert, delete, replace, transpose) required to transform one string to the other.**

*Example,*

$$e(\text{"Virginia"}, \text{"Vermont"}) = 5$$

Virginia  
Verginia  
Verminia  
Vermonia  
Vermonta  
Vermont

# Clustering Alternatives



- **Absolute edit distance**

$$e(u, v) \leq \delta$$

	Lexically Cleaned-Up		Raw	
	Number of Clusters	$\Delta$	Number of Clusters	$\Delta$
<b>Edit Dist. Threshold</b>				
	16427		20168	
1	13527	2900	17226	2942
2	12786	741	16357	869
3	12160	608	15665	692
5	10924	1236	13554	2111

# Edit Distance 1

Virginia, University, Charlottesville  
Virginia, University, Charlottesville  
Virginia, University, Charlottesville  
Virginia University, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottesvill  
Virginia, University  
Virginia University  
Virginia, University  
University of Virginia  
Virginia University, Charlottesville  
University of Virginia, Virginia  
University of VA., Charlottesville

## Edit Distance 5

Virginia, University, Charlottesville  
Virginia, University, Charlottesville  
Virgina, University, Charlottesville  
Virginia University, Charlottesville  
Virgina University, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottesville  
University of Virginia, Charlottsville  
University of Virginia, Charlottesvill  
Victoria, University  
Victoria, University  
Victoria University  
Virginia University  
Virginia, University  
University of Arizona  
University of Arizona  
University of Virginia  
University of Virginia, Virginia  
University of VA., Charlottesville

# Clustering Alternatives



- **Absolute edit distance**

$$e(u, v) \leq \delta$$

- **Relative edit distance**

$$e(u, v) \leq \alpha \min(u, v)$$

	EDT1 representatives		Lexically Cleaned-Up	
Edit Dist. Threshold	Number of Clusters	$\Delta$	Number of Clusters	$\Delta$
	13527		16427	
1/10	11825	1702	11872	4555
1/9	11595	230	11641	141
1/7	10888	707	10926	725
1/5	9542	1346	9583	1343

## Edit Distance 1/10

Virginia, University, Charlottesville

Virginia, University, Charlottesville

Virgina University, Charlottesville

University of Virginia, Charlottesville

Virginia, University

University of Virginia

University of Virginia, Virginia

University of VA., Charlottesville

## Edit Distance 1/5

Virginia, University, Charlottesville  
Virginia, University, Charlottesville  
Virginia University, Charlottesville  
University of Virginia, Charlottesville  
Victoria, University  
Victoria, University  
Virginia, University  
Pretoria, University  
University of Virginia  
University of Virginia, Virginia  
University of VA., Charlottesville

# Clustering Alternatives



- **Absolute edit distance**

$$e(u, v) \leq \delta$$

- **Relative edit distance**

$$e(u, v) \leq \alpha \min(u, v)$$

- **Approximate word matching**

## Approximate Word Matching vs. String Matching

$s_1$ : Moskovskii Gosudarstvennyi Pedagogicheskii Institut, Moscow  
 $s_2$ : Moskovskij Pedagogicheskij Gosudarstvennyj University, Moscow  
 $s_3$ : Virginia, University  
 $s_4$ : University of Virginia  
 $s_5$ : University of Vermont

$e(u, v)$

	$s_1$ (59)	$s_2$ (61)	$s_3$ (20)	$s_4$ (22)	$s_5$ (21)
$s_1$	0	36	50	50	50
$s_2$		0	45	52	51
$s_3$			0	17	16
$s_4$				0	5
$s_5$					0

$w(u, v)$

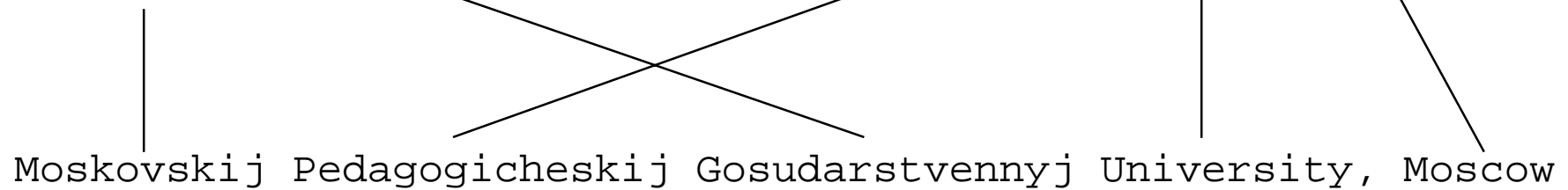
	$s_1$ (55)	$s_2$ (57)	$s_3$ (18)	$s_4$ (20)	$s_5$ (19)
$s_1$	0	11	56	52	52
$s_2$		0	48	44	44
$s_3$			0	2	7
$s_4$				0	5
$s_5$					0

# Alternative distance measures

Approximate word matching

Original distance 36

Moskovskii Gosudarstvenni Pedagogicheskii Institut, Moscow



Distance 11

Sorted surrogates

Gosudarstvenni Institut Moscow Moskovskii Pedagogicheskii

Gosudarstvennyj Moscow Moskovskij Pedagogicheskij University

Distance 22

	Simple Approach (a)		Combination Approach (b)	
Edit Dist. Threshold	Number of Clusters	$\Delta$	Number of Clusters	$\Delta$
	12212		10866	
1/10	10719	1493	10352	514
1/9	10484	235	10165	187
1/7	9844	640	9580	585
1/5	8677	1167	8514	1066

## Affiliation string

charlottesville university virginia  
Virginia, University, Charlottesville  
University of Virginia, Charlottesville  
Virgina University, Charlottesville

university virginia  
Virginia, University

of university virginia  
University of Virginia, Virginia  
University of Virginia

charlottesville of university va  
University of VA., Charlottesville

Affiliation cluster/string	Number of occurrences
Virginia, University, Charlottesville	502
Virginia, University, Charlottesville	431
University of Virginia, Charlottesville	70
Virginia, University	59
University of Virginia	2
University of Virginia	1
University of Virginia, Virginia	1
University of VA., Charlottesville	1

# Conclusions



- **Authority control of some form will be a necessary component of digital libraries.**
- **In general, authority files cannot be generated fully automatically.**
- **Effective tools to aid in the construction of authority files are possible.**

# Acknowledgements



**This work supported in part by:**

- **NSF grant CDA-9529253**
- **DARPA contract N66001-97-C-8542**
- **DOE grant DE-FG05-95ER25254**
- **NASA GSRP fellowship NGT5-50062**