Domain-Independent Data Extraction: Person Names

Carl Christensen and Deryle Lonsdale
Brigham Young University
cvchristensen@gmail
lonz@byu.edu

Challenge

- Extraction software and techniques yield good results with domain specific data extraction
- Person names and information rarely domain specific
- Identification and extraction difficult because of noisy data, lack of formatting

WePS task

- Web People Search
- 18 attribute values on person names

  - Date of birth, Birth place, Other name, Occupation, Affiliation, Work, Award, School, Major, Degree, Mentor, Location, Nationality, Relatives, Phone, FAX, Email, Web site

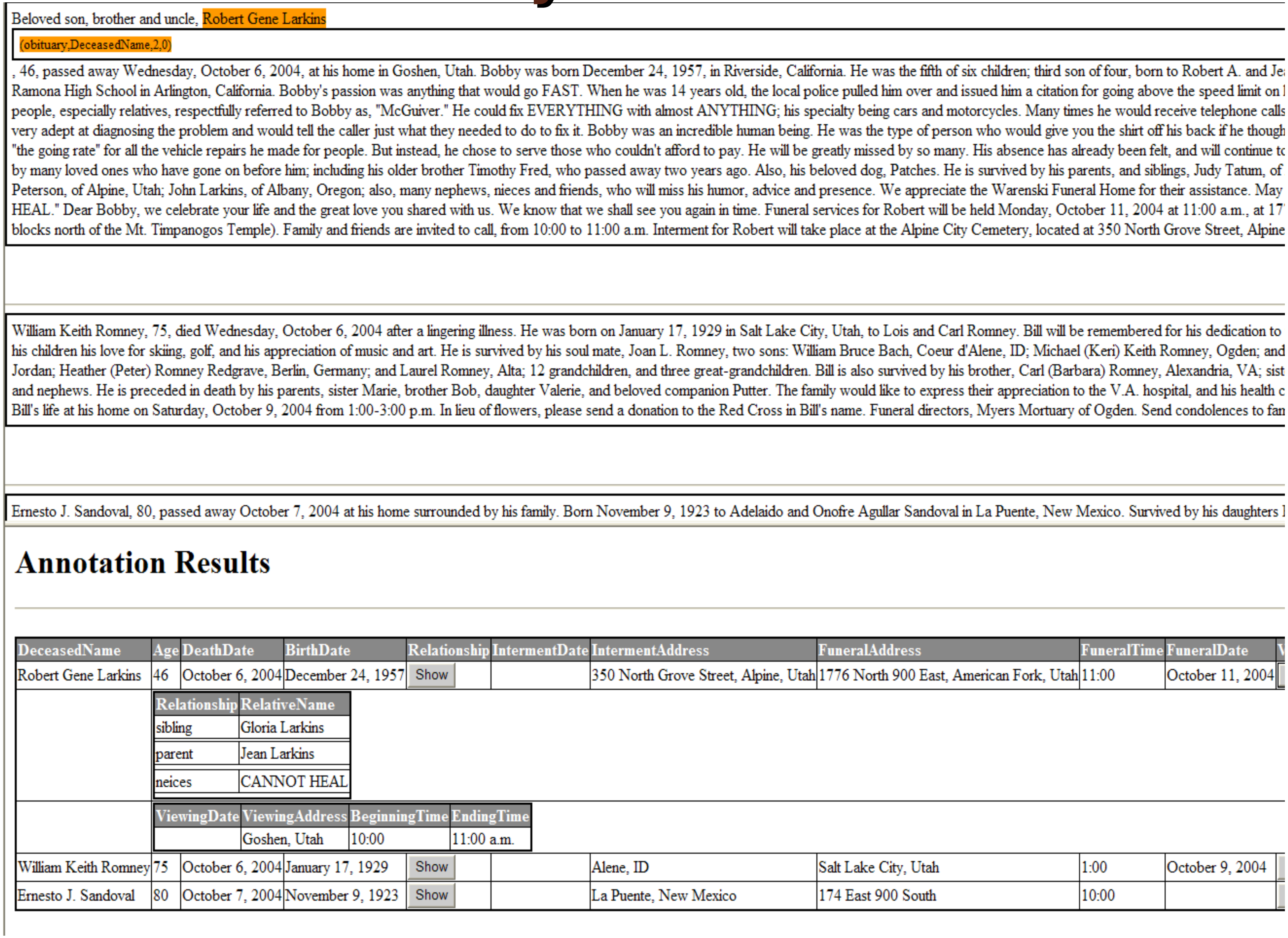
- Training corpus – 17 names, approx. 100 web pages per
- Script given to evaluate performance
- Test corpus of comparable size
- New ground in information extraction

Ontos

- Software developed by BYU data extraction group
- Ontology based method leveraged to organize data
- Off the shelf performance
- Similar uses for obituaries and car ads information extraction

  - Good performance on these tasks

Ontos obituary results



Beloved son, brother and uncle, Robert Gene Larkins

(obituary,DeceasedName,2,0)

, 46, passed away Wednesday, October 6, 2004, at his home in Goshen, Utah. Bobby was born December 24, 1957, in Riverside, California. He was the fifth of six children; third son of four, born to Robert A. and Jea
Ramona High School in Arlington, California. Bobby's passion was anything that would go FAST. When he was 14 years old, the local police pulled him over and issued him a citation for going above the speed limit on
people, especially relatives, respectfully referred to Bobby as, "McGuiver." He could fix EVERYTHING with almost ANYTHING; his specialty being cars and motorcycles. Many times he would receive telephone calls
very adept at diagnosing the problem and would tell the caller just what they needed to do to fix it. Bobby was an incredible human being. He was the type of person who would give you the shirt off his back if he though
"the going rate" for all the vehicle repairs he made for people. But instead, he chose to serve those who couldn't afford to pay. He will be greatly missed by so many. His absence has already been felt, and will continue to
by many loved ones who have gone on before him; including his older brother Timothy Fred, who passed away two years ago. Also, his beloved dog, Patches. He is survived by his parents, and siblings, Judy Tatum, of
Peterson, of Alpine, Utah; John Larkins, of Albany, Oregon; also, many nephews, nieces and friends, who will miss his humor, advice and presence. We appreciate the Warenski Funeral Home for their assistance. May
HEAL." Dear Bobby, we celebrate your life and the great love you shared with us. We know that we shall see you again in time. Funeral services for Robert will be held Monday, October 11, 2004 at 11:00 a.m., at 17
blocks north of the Mt. Timpanogos Temple). Family and friends are invited to call, from 10:00 to 11:00 a.m. Interment for Robert will take place at the Alpine City Cemetery, located at 350 North Grove Street, Alpine

William Keith Romney, 75, died Wednesday, October 6, 2004 after a lingering illness. He was born on January 17, 1929 in Salt Lake City, Utah, to Lois and Carl Romney. Bill will be remembered for his dedication to
his children his love for skiing, golf, and his appreciation of music and art. He is survived by his soul mate, Joan L. Romney, two sons: William Bruce Bach, Coeur d'Alene, ID; Michael (Keri) Keith Romney, Ogden; and
Jordan; Heather (Peter) Romney Redgrave, Berlin, Germany; and Laurel Romney, Alta; 12 grandchildren, and three great-grandchildren. Bill is also survived by his brother, Carl (Barbara) Romney, Alexandria, VA; sist
and nephews. He is preceded in death by his parents, sister Marie, brother Bob, daughter Valerie, and beloved companion Putter. The family would like to express their appreciation to the V.A. hospital, and his health c
Bill's life at his home on Saturday, October 9, 2004 from 1:00-3:00 p.m. In lieu of flowers, please send a donation to the Red Cross in Bill's name. Funeral directors, Myers Mortuary of Ogden. Send condolences to fan

Ernesto J. Sandoval, 80, passed away October 7, 2004 at his home surrounded by his family. Born November 9, 1923 to Adelaido and Onofre Agullar Sandoval in La Puente, New Mexico. Survived by his daughters I

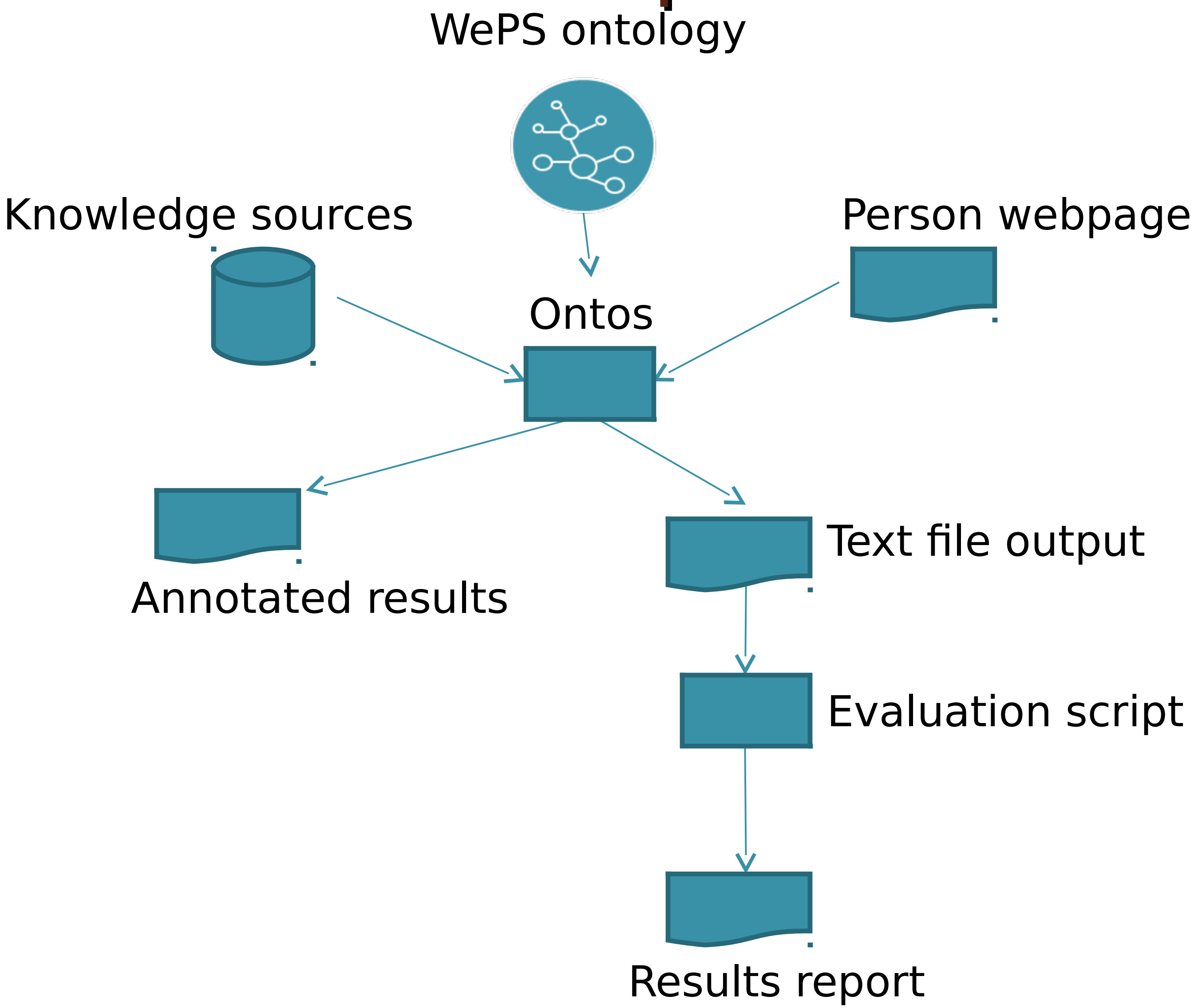
Annotation Results

DeceasedName	Age	DeathDate	BirthDate	Relationship	IntermentDate	IntermentAddress	FuneralAddress	FuneralTime	FuneralDate	V
Robert Gene Larkins	46	October 6, 2004	December 24, 1957	Show		350 North Grove Street, Alpine, Utah	1776 North 900 East, American Fork, Utah	11:00	October 11, 2004	
William Keith Romney	75	October 6, 2004	January 17, 1929	Show		Alene, ID	Salt Lake City, Utah	1:00	October 9, 2004	
Ernesto J. Sandoval	80	October 7, 2004	November 9, 1923	Show		La Puente, New Mexico	174 East 900 South	10:00		

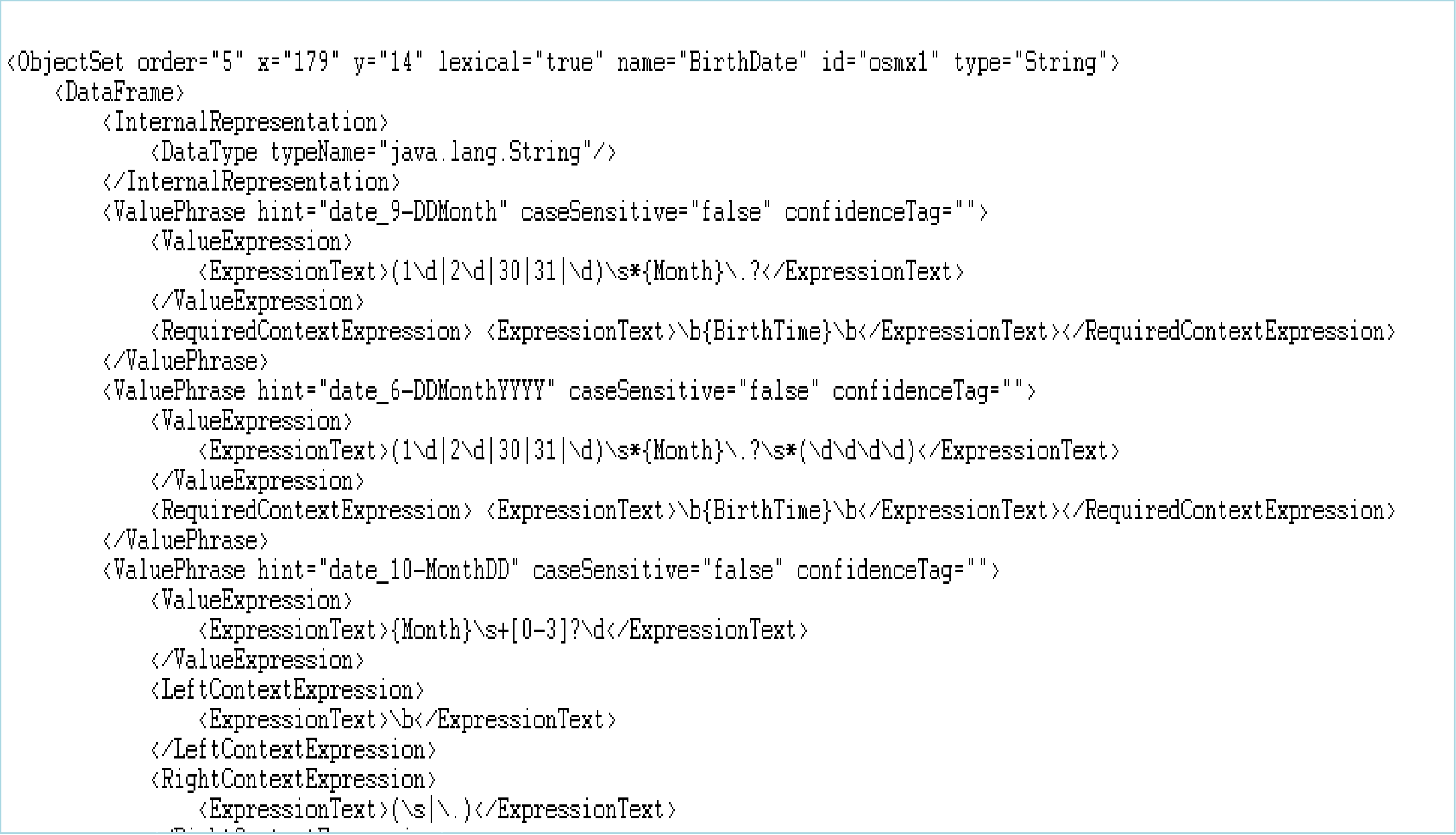
Relationship	RelativeName
sibling	Gloria Larkins
parent	Jean Larkins
neices	CANNOT HEAL

ViewingDate	ViewingAddress	BeginningTime	EndingTime
	Goshen, Utah	10:00	11:00 a.m.

WePS extraction process



Data frames



```
<ObjectSet order="5" x="179" y="14" lexical="true" name="BirthDate" id="osmx1" type="String">
    <DataFrame>
        <InternalRepresentation>
            <DataType typeName="java.lang.String"/>
        </InternalRepresentation>
        <ValuePhrase hint="date_9-DDMonth" caseSensitive="false" confidenceTag="">
            <ValueExpression>
                <ExpressionText>(1\d|2\d|30|31|\d)\s*{Month}\.?</ExpressionText>
            </ValueExpression>
            <RequiredContextExpression> <ExpressionText>\b{BirthTime}\b</ExpressionText></RequiredContextExpression>
        </ValuePhrase>
        <ValuePhrase hint="date_6-DDMonthYYYY" caseSensitive="false" confidenceTag="">
            <ValueExpression>
                <ExpressionText>(1\d|2\d|30|31|\d)\s*{Month}\.?\s*(\d\d\d\d)</ExpressionText>
            </ValueExpression>
            <RequiredContextExpression> <ExpressionText>\b{BirthTime}\b</ExpressionText></RequiredContextExpression>
        </ValuePhrase>
        <ValuePhrase hint="date_10-MonthDD" caseSensitive="false" confidenceTag="">
            <ValueExpression>
                <ExpressionText>{Month}\s+[0-3]?\d</ExpressionText>
            </ValueExpression>
            <LeftContextExpression>
                <ExpressionText>\b</ExpressionText>
            </LeftContextExpression>
            <RightContextExpression>
                <ExpressionText>(\s|\.)</ExpressionText>
            </RightContextExpression>
```

- XML description of extraction ontology components

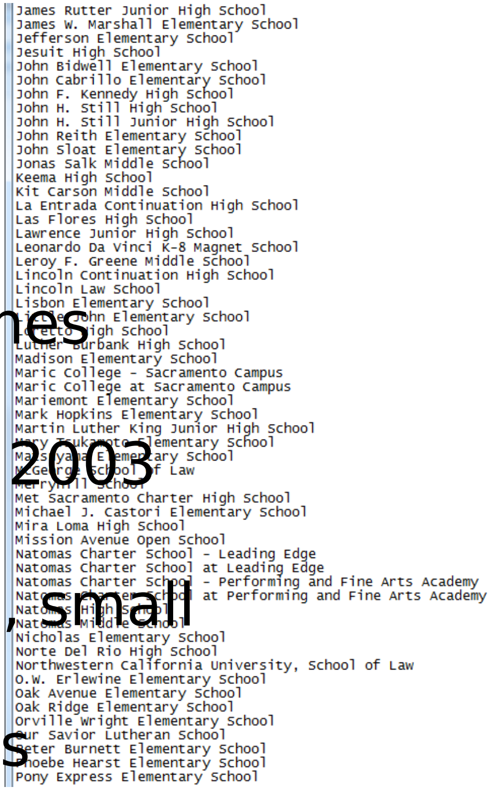
Knowledge files

- Names, cities, countries, hypocoristics, occupations,

- Knowledge gathered from extracting and formatting databases
  - Live Journal
  - Wikipedia
  - Bureau of Labor Statistics
  - etc.

- Approximately 80,000 school names and 30,000 occupations
  - 66% of total schools in U.S. in 2003

- All possible options for some files, small subset for others
  - e.g. Occupations, hypocoristics



```
James Rutter Junior High School
James W. Marshall Elementary School
Jefferson Elementary School
Jesuit High School
John Bidwell Elementary School
John Cabrillo Elementary School
John F. Kennedy High School
John H. Still High School
John H. Still Junior High School
John Reith Elementary School
John Sloat Elementary School
Jonas Salk Middle School
Keema High School
Kit Carson Middle School
La Entrada Continuation High School
Las Flores High School
Lawrence Junior High School
Leonardo Da Vinci K-8 Magnet School
Leroy F. Greene Middle School
Lincoln Continuation High School
Lincoln Law School
Lisbon Elementary School
Little John Elementary School
Loretto High School
Luther Burbank High School
Madison Elementary School
Maric College - Sacramento Campus
Maric College at Sacramento Campus
Mariemont Elementary School
Mark Hopkins Elementary School
Martin Luther King Junior High School
Mary Tsukamoto Elementary School
Matsuyama Elementary School
McGeorge School of Law
Merryhill School
Met Sacramento Charter High School
Michael J. Castori Elementary School
Mira Loma High School
Mission Avenue Open School
Natomas Charter School - Leading Edge
Natomas Charter School at Leading Edge
Natomas Charter School - Performing and Fine Arts Academy
Natomas Charter School at Performing and Fine Arts Academy
Natomas High School
Natomas Middle School
Nicholas Elementary School
Norte Del Rio High School
Northwestern California University, School of Law
O.W. Erlewine Elementary School
Oak Avenue Elementary School
Oak Ridge Elementary School
Orville Wright Elementary School
Our Savior Lutheran School
Peter Burnett Elementary School
Phoebe Hearst Elementary School
Pony Express Elementary School
```

Constraints

- Required context expressions

  <RequiredContextExpression>
  <ExpressionText>\b{BirthTime}\b</Expre
  ssionText></RequiredContextExpression>

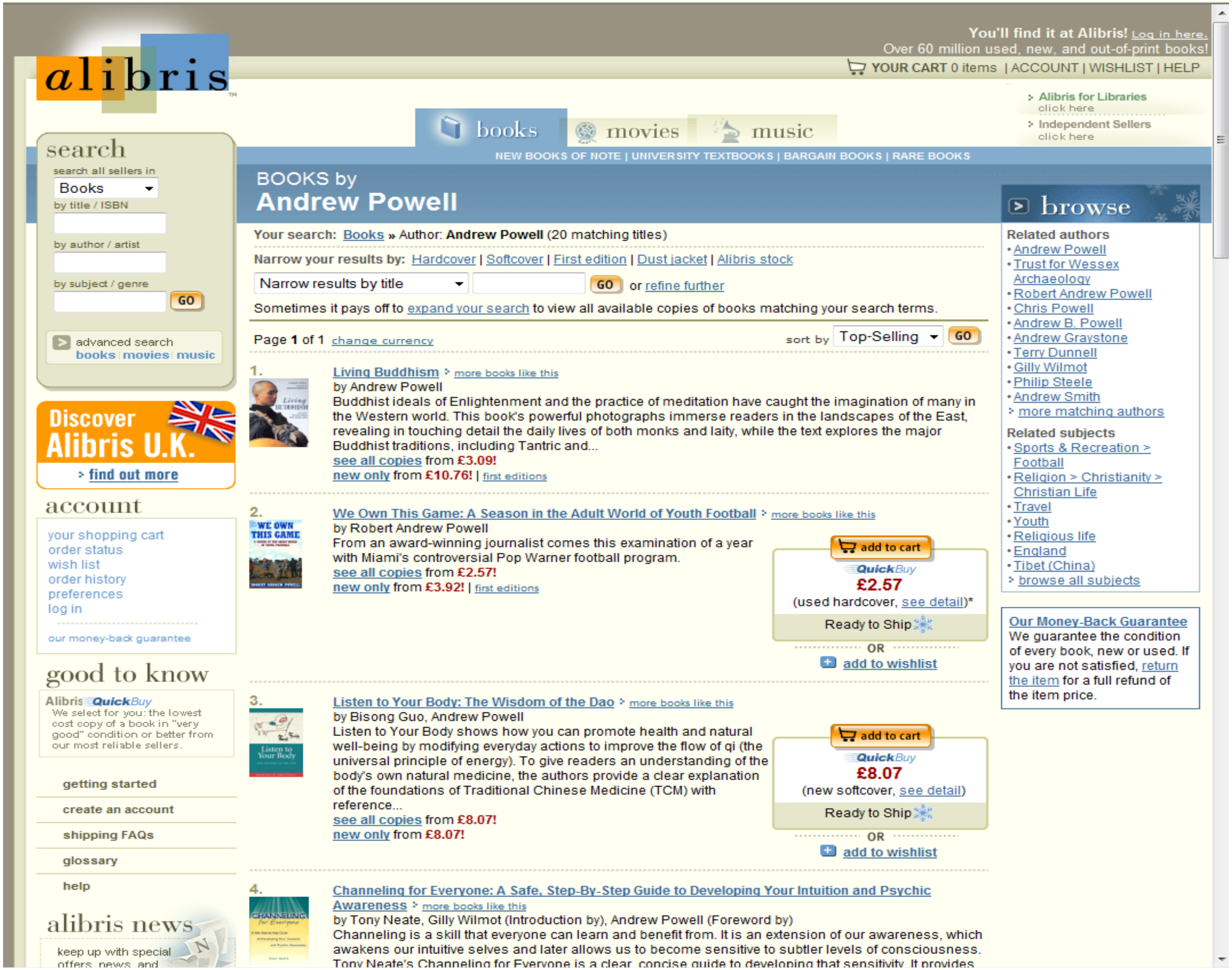
- Cardinality

  Search Person [0:*] has Occupation
  [1:*];
  Search Person [0:*] has Affiliation
  [1:*];
  Search Person [0:*] has Email [1];

- Regular expressions

  Email : [\w]+[\d]*@[\w]+[.]{1}[\w]+

Sample webpage



You'll find it at Alibris! Log in here.
Over 60 million used, new, and out-of-print books!
YOUR CART 0 items | ACCOUNT | WISHLIST | HELP

alibris™

Alibris for Libraries
click here
Independent Sellers
click here

books movies music

NEW BOOKS OF NOTE | UNIVERSITY TEXTBOOKS | BARGAIN BOOKS | RARE BOOKS

search
search all sellers in
Books
by title / ISBN
by author / artist
by subject / genre
GO
advanced search
books movies music

Discover Alibris U.K.
find out more

account
your shopping cart
order status
wish list
order history
preferences
log in
our money-back guarantee

good to know
Alibris QuickBuy
We select for you: the lowest cost copy of a book in "very good" condition or better from our most reliable sellers.

getting started
create an account
shipping FAQs
glossary
help

alibris news
keep up with special offers, news, and

BOOKS by
Andrew Powell

Your search: Books » Author: **Andrew Powell** (20 matching titles)

Narrow your results by: Hardcover | Softcover | First edition | Dust jacket | Alibris stock

Narrow results by title GO or refine further

Sometimes it pays off to expand your search to view all available copies of books matching your search terms.

Page **1** of 1 change currency sort by Top-Selling GO

1. Living Buddhism › more books like this
by Andrew Powell
Buddhist ideals of Enlightenment and the practice of meditation have caught the imagination of many in the Western world. This book's powerful photographs immerse readers in the landscapes of the East, revealing in touching detail the daily lives of both monks and laity, while the text explores the major Buddhist traditions, including Tantric and...
see all copies from £3.09!
new only from £10.76! | first editions

2. We Own This Game: A Season in the Adult World of Youth Football › more books like this
by Robert Andrew Powell
From an award-winning journalist comes this examination of a year with Miami's controversial Pop Warner football program.
see all copies from £2.57!
new only from £3.92! | first editions

add to cart
QuickBuy
£2.57
(used hardcover, see detail)*
Ready to Ship
OR
add to wishlist

3. Listen to Your Body: The Wisdom of the Dao › more books like this
by Bisong Guo, Andrew Powell
Listen to Your Body shows how you can promote health and natural well-being by modifying everyday actions to improve the flow of qi (the universal principle of energy). To give readers an understanding of the body's own natural medicine, the authors provide a clear explanation of the foundations of Traditional Chinese Medicine (TCM) with reference...
see all copies from £8.07!
new only from £8.07!

add to cart
QuickBuy
£8.07
(new softcover, see detail)
Ready to Ship
OR
add to wishlist

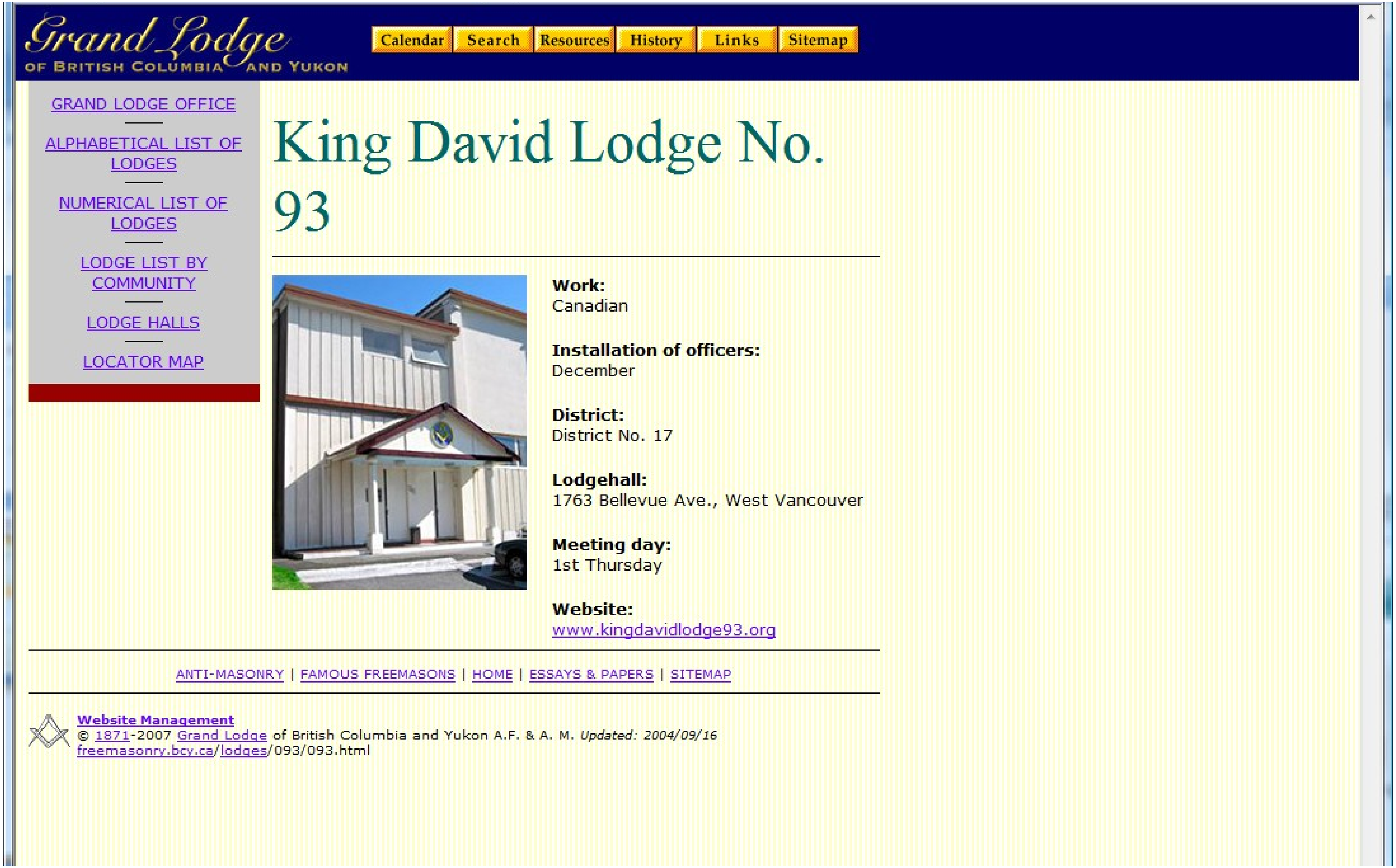
4. Channeling for Everyone: A Safe, Step-By-Step Guide to Developing Your Intuition and Psychic Awareness › more books like this
by Tony Neate, Gilly Wilmot (Introduction by), Andrew Powell (Foreword by)
Channeling is a skill that everyone can learn and benefit from. It is an extension of our awareness, which awakens our intuitive selves and later allows us to become sensitive to subtler levels of consciousness.
Tony Neate's Channeling for Everyone is a clear, concise guide to developing that sensitivity. It provides

browse

Related authors
- Andrew Powell
- Trust for Wessex Archaeology
- Robert Andrew Powell
- Chris Powell
- Andrew B. Powell
- Andrew Graystone
- Terry Dunnell
- Gilly Wilmot
- Philip Steele
- Andrew Smith
- more matching authors

Related subjects
- Sports & Recreation > Football
- Religion > Christianity > Christian Life
- Travel
- Youth
- Religious life
- England
- Tibet (China)
- browse all subjects

Our Money-Back Guarantee
We guarantee the condition of every book, new or used. If you are not satisfied, return the item for a full refund of the item price.



King David Lodge No. 93

Work:
Canadian

Installation of officers:
December

District:
District No. 17

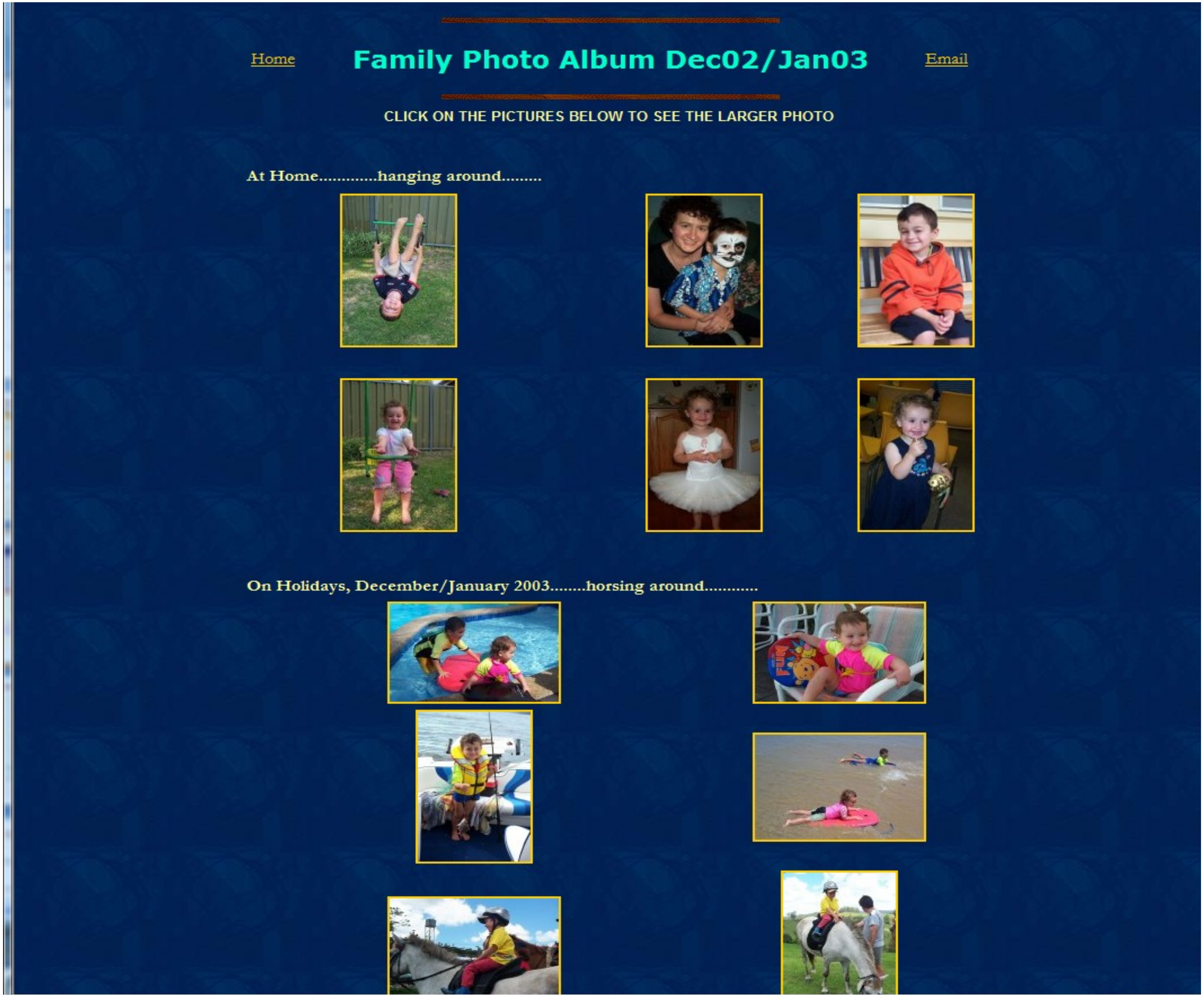
Lodgehall:
1763 Bellevue Ave., West Vancouver

Meeting day:
1st Thursday

Website:
www.kingdavidlodge93.org

ANTI-MASONRY | FAMOUS FREEMASONS | HOME | ESSAYS & PAPERS | SITEMAP

Website Management
© 1871-2007 Grand Lodge of British Columbia and Yukon A.F. & A. M. *Updated: 2004/09/16*
freemasonry.bcy.ca/lodges/093/093.html



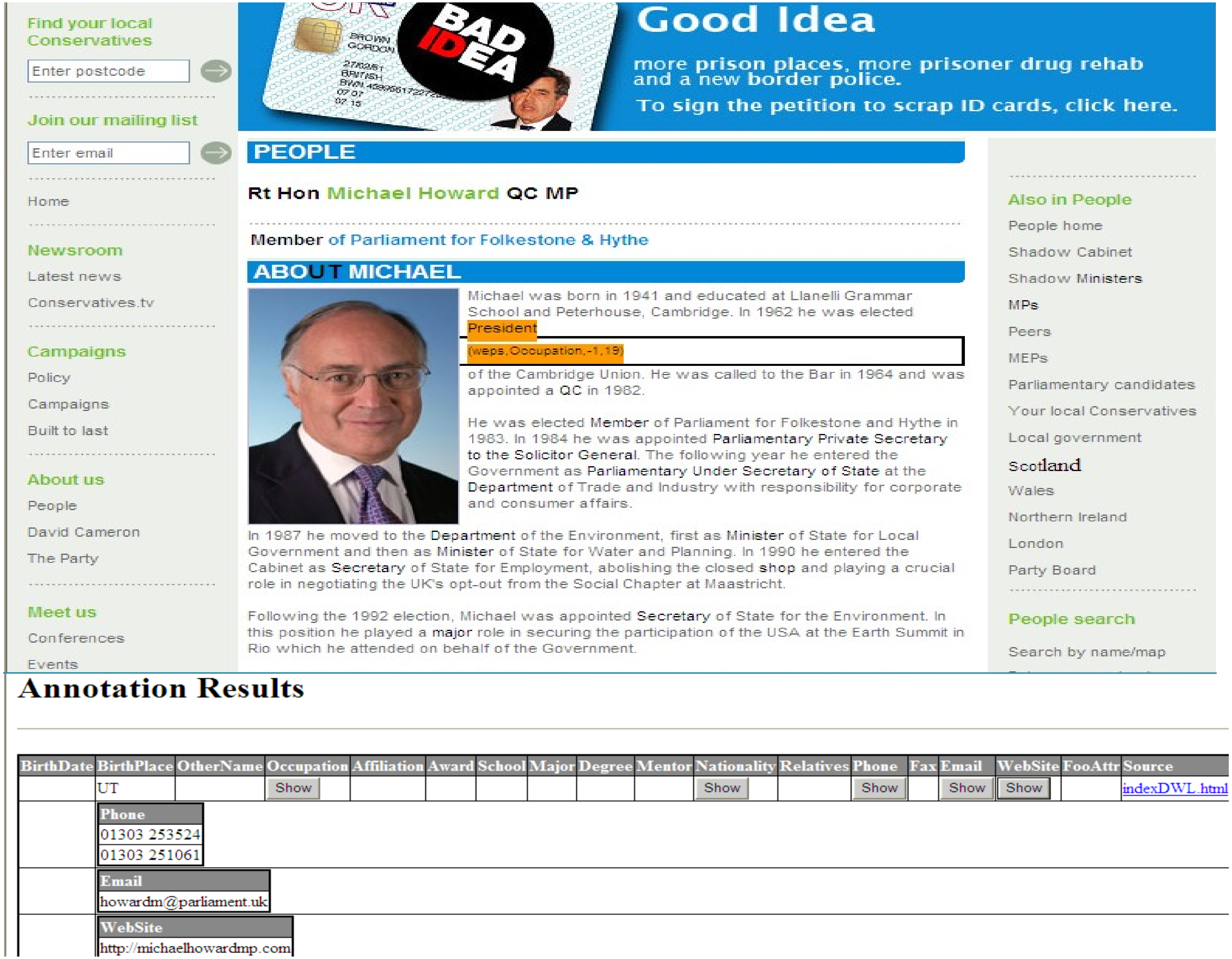
Family Photo Album Dec02/Jan03

CLICK ON THE PICTURES BELOW TO SEE THE LARGER PHOTO

At Home............hanging around.........

On Holidays, December/January 2003........horsing around...........

Sample annotated webpage



Find your local Conservatives

Enter postcode

Join our mailing list

Enter email

Home

Newsroom

Latest news

Conservatives.tv

Campaigns

Policy

Campaigns

Built to last

About us

People

David Cameron

The Party

Meet us

Conferences

Events

Good Idea

more prison places, more prisoner drug rehab and a new border police.

To sign the petition to scrap ID cards, click here.

PEOPLE

Rt Hon Michael Howard QC MP

Member of Parliament for Folkestone & Hythe

ABOUT MICHAEL

Michael was born in 1941 and educated at Llanelli Grammar School and Peterhouse, Cambridge. In 1962 he was elected President

(weps,Occupation,-1,19)

of the Cambridge Union. He was called to the Bar in 1964 and was appointed a QC in 1982.

He was elected Member of Parliament for Folkestone and Hythe in 1983. In 1984 he was appointed Parliamentary Private Secretary to the Solicitor General. The following year he entered the Government as Parliamentary Under Secretary of State at the Department of Trade and Industry with responsibility for corporate and consumer affairs.

In 1987 he moved to the Department of the Environment, first as Minister of State for Local Government and then as Minister of State for Water and Planning. In 1990 he entered the Cabinet as Secretary of State for Employment, abolishing the closed shop and playing a crucial role in negotiating the UK's opt-out from the Social Chapter at Maastricht.

Following the 1992 election, Michael was appointed Secretary of State for the Environment. In this position he played a major role in securing the participation of the USA at the Earth Summit in Rio which he attended on behalf of the Government.

Also in People

People home

Shadow Cabinet

Shadow Ministers

MPs

Peers

MEPs

Parliamentary candidates

Your local Conservatives

Local government

Scotland

Wales

Northern Ireland

London

Party Board

People search

Search by name/map

Annotation Results

BirthDate	BirthPlace	OtherName	Occupation	Affiliation	Award	School	Major	Degree	Mentor	Nationality	Relatives	Phone	Fax	Email	WebSite	FooAttr	Source
	UT		Show							Show		Show		Show	Show		indexDWL.html
	Phone 01303 253524 01303 251061																
	Email howardm@parliament.uk																
	WebSite http://michaelhowardmp.com																

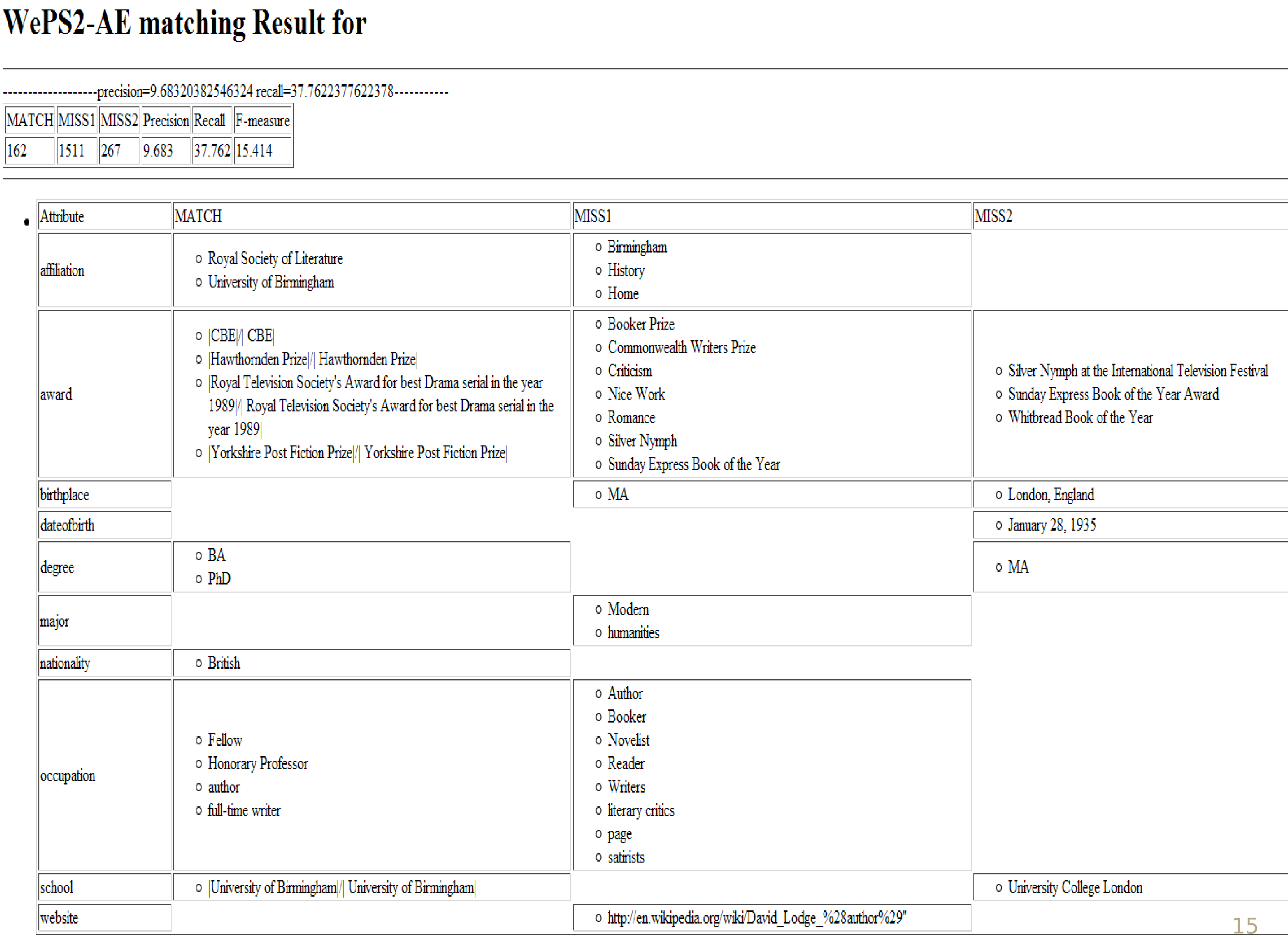
Precision/Recall

$$\text{Recall} = \frac{|\{\text{relevant information}\} \cap \{\text{retrieved information}\}|}{|\{\text{relevant information}\}|}$$

$$\text{Precision} = \frac{|\{\text{relevant information}\} \cap \{\text{retrieved information}\}|}{|\{\text{retrieved information}\}|}$$

$$F_{\text{alpha}} = \frac{(1 + \text{alpha}) * \text{precision} * \text{recall}}{((\text{alpha} * \text{precision}) + \text{recall})}$$

alpha = 0.5 for
WePS



WePS2-AE matching Result for

-------------------precision=9.68320382546324 recall=37.7622377622378-----------

MATCH	MISS1	MISS2	Precision	Recall	F-measure
162	1511	267	9.683	37.762	15.414

-

Attribute	MATCH	MISS1	MISS2
affiliation	○ Royal Society of Literature ○ University of Birmingham	○ Birmingham ○ History ○ Home	
award	○ |CBE|/| CBE| ○ |Hawthornden Prize|/| Hawthornden Prize| ○ |Royal Television Society's Award for best Drama serial in the year 1989|/| Royal Television Society's Award for best Drama serial in the year 1989| ○ |Yorkshire Post Fiction Prize|/| Yorkshire Post Fiction Prize|	○ Booker Prize ○ Commonwealth Writers Prize ○ Criticism ○ Nice Work ○ Romance ○ Silver Nymph ○ Sunday Express Book of the Year	○ Silver Nymph at the International Television Festival ○ Sunday Express Book of the Year Award ○ Whitbread Book of the Year
birthplace		○ MA	○ London, England
dateofbirth			○ January 28, 1935
degree	○ BA ○ PhD		○ MA
major		○ Modern ○ humanities	
nationality	○ British		
occupation	○ Fellow ○ Honorary Professor ○ author ○ full-time writer	○ Author ○ Booker ○ Novelist ○ Reader ○ Writers ○ literary critics ○ page ○ satirists	
school	○ |University of Birmingham|/| University of Birmingham|		○ University College London
website		○ http://en.wikipedia.org/wiki/David_Lodge_%28author%29"	

Challenges

- Smaller/larger match preference

  - Preference for “Arizona” as place over “University of Arizona” as school
- DOM parser

  - Unofficial HTML tags cause system to fail
- Text formatting

  - Record detection for individuals intractable
- System functionality

  - Cardinality bounds, system output file

Performance

- Very low initial precision/recall: < 1%
- Increased drastically with knowledge engineering and system constraints

  - 27 % recall with some person results approaching 40% recall

  - Approaching 10% precision
- Nothing to measure against

  - Official WePS results will be released in April

Future work

- Ontos robustness
- Machine learning for constraints/knowledge files
- Person name disambiguation
- Keyword probability values