# contributed articles

DOI:10.1145/1897816.1897839

## Google's WebTables and Deep Web Crawler identify and deliver this otherwise inaccessible resource directly to end users.

BY MICHAEL J. CAFARELLA, ALON HALEVY, AND JAYANT MADHAVAN

# Structured Data on the Web

THOUGH THE WEB is best known as a vast repository of shared documents, it also contains a significant amount of structured data covering a complete range of topics, from product to financial, public-record, scientific, hobby-related, and government. Structured data on the Web shares many similarities with the kind of data traditionally managed by commercial database systems but also reflects some unusual characteristics of its own; for example, it is embedded in textual Web pages and must be extracted prior to use; there is no centralized data design as there is in a traditional database; and, unlike traditional databases that focus on a single domain, it covers everything. Existing data-management systems do not address these challenges and assume their data is modeled within a well-defined domain.

This article discusses the nature of Web-embedded structured data and the challenges of managing it. To begin, we present two relevant research projects developed at Google over the past five years. The first, WebTables, compiles a huge collection of databases by crawling the Web to find small relational databases expressed using the HTML table tag. By performing data mining on the resulting extracted information, WebTables is able to introduce new data-centric applications (such as schema completion and synonym finding). The second, the Google Deep Web Crawler, attempts to surface information from the Deep Web, referring to data on the Web available only by filling out Web forms, so cannot be crawled by traditional crawlers. We describe how this data is crawled by automatically submitting relevant queries to a vast number of Web forms. The two projects are just the first steps toward exposing and managing structured Web data largely ignored by Web search engines.

## Web Data

Structured data on the Web exists in several forms, including HTML tables, HTML lists, and back-end Deep Web databases (such as the books sold on Amazon.com). We estimate in excess of one billion data sets as of February 2011. More than 150 million sources come from a subset of all English-language HTML tables,<sup>4,5</sup> while Elmeleegy et al<sup>11</sup> suggested an equal number from HTML lists, a total that does not account for the non-English Web. Finally, our experience at Google

## » key insights

- Because data on the Web is about everything, any approach that attempts to leverage it cannot rely on building a model of the data ahead of time but on domain-independent methods instead.
- The sheer quantity and heterogeneity of structured data on the Web enables new approaches to problems involving data integration from multiple sources.
- While the content of structured data is typically different from what is found in text on the Web, each content collection can be leveraged to better understand other collections.

suggests the Deep Web alone can generate more than one billion pages of valuable structured data. The result is an astounding number of distinct structured data sets, most still waiting to be exposed more effectively to users.

This structured data differs from data stored in traditional relational databases in several ways:

Data in "page context" must be extracted. Consider a database embedded in an HTML table (such as local coffeehouses in Seattle and the U.S. presidents in Figure 1). To the user the data set appears to be structured, but a computer program must be able to automatically distinguish it from, say, a site's navigational bar that also uses an HTML table. Similarly, a Web form that gives access to an interesting Deep Web database, perhaps containing all Starbucks locations in the world, is not that different from a form offering simple mailing-list signup. The computer program might also have to automatically extract schema information in the form of column labels sometimes appearing in the first row of an HTML table but that sometimes do not exist at all. Moreover, the subject of a table may be described in the surrounding text, making it difficult to extract. There is nothing akin to traditional relational metadata that leaves no doubt as to how many tables there are and the relevant schema information for each table.

No centralized data design or dataquality control. In a traditional database, the relational schema provides a topic-specific design that must be observed by all data elements. The database and the schema may also enforce certain quality controls (such as observing type consistency within a column, disallowing empty cells, and constraining data values to a certain legal range). For example, the set of coffeehouses may have a column called year-founded containing integers constrained to a relatively small range. Neither data design nor quality control exists for Web data; for



example, if a year-founded string is in the first row, there is nothing to prevent the string *macchiatone* from appearing beneath it. Any useful application making use of Web data must also be able to address uncertain data design and quality.

Vast number of topics. A tradi-

tional database typically focuses on a particular domain (such as products or proteins) and therefore can be modeled in a coherent schema. On the Web, data covers everything, and is also one of its appeals. The breadth and cultural variations of data on the Web make it inconceivable that any

🏟 • 🚸 • 🥑 😳 🏠 🗆 https:	//www.enchantedlear	ming.com	n/histor	ry/us/p	ores/fe	st.shtml							1		<b>C</b>  •				
	As a thank-you	bonus, s	ite mer	mbers	have	r access t	a banna	er-ad-free v	ersion o	f the sit	e, with p	rint-fri	iendly p	age	s				
					(Al	ready a	membe	r? Click h	ere.)										
US Flags	US Flags					Histo									USG	eograph	œ		
A B C D E	E 55 1	1	L	K		L	M	N Q	P	Q	R	S	Т	1	UV	3	K S	K I	X I
S frican American	Artists		Expl	lorers	of the	eUS		Invento	CS	1	US Bred	identr.			US Sy	mbols		US	States
							served	Alphabet	ALED CHUC	. sin	HI LADIC C	of Data						brah	m Line
						e at leas	35 year	s of age, t	hey mu	st be no			tens of	the	United	States,			and the second
						e at leas ected to	35 year	s of age, t erm as Pre	hey mu	st be no			tens of	the	United				and the second
een residents of the U.S. for at lea President		s, a persi		nnot b	be ele Part	e at leas ected to	35 year a third to	s of age, t erm as Pre	hey mu sident.)	st be no		m citiz	rens of	the					and the second
een residents of the U.S. for at leas President . <u>George Washington</u> (1732-1799)		s, a perse	on car	nnot b Federal	be ele Part	e at leas ected to	35 year a third to	s of age, t erm as Pre Term a	hey mu sident.)	st be no	ative-bo	rn citiz dams		the					and the second
een residents of the U.S. for at leas President George Washington (1732-1799) John Adams (1735-1826)		s, a perso N F	on car (one, F	nnot b Federal	be ele Part dist	e at leas ected to ty	35 year a third to	s of age, t erm as Pre Term a 789-1797	hey mu sident.)	st be no	tive-bo	en citia dams s Jeffer			Vice-Pr				and the second
en residents of the U.S. for at leas President George Washington (1732-1799) John Adams (1735-1826) Thomas Jefferson (1743-1826)		s, a perse N F D	on car lone, F loderali Democr	Federal ist ratic-R	Part Part dist tepubl	e at leas ected to ty lican lican	35 year a third to 1	s of age, ti erm as Pre Term a 789-1797 797-1801	hey mu sident.)	st be no	John A Thoma Aaron George	rn citia dams s Jeffer Burr, G : Clintor	sion eorge C n, Elbric	linte	Vice-Pr				and the second
een residents of the U.S. for at leas President George Washington (1732-1799) John Adams (1735-1826) Thomas Afferson (1743-1826) James Madison (1751-1836) James Madison (1751-1836)		s, a perso N F D D	on car lone, F ederali Democr Democr	Federal ist ratic-R ratic-R	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1	s of age, t erm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825	hey mu sident.)	st be no	ative-bo John A Thoma Aaron George Daniel	en citia dams s Jeffer Burr, G : Clintor Tompk	sion eorge C n, Elbric	linte	Vice-Pr				and the second
een residents of the U.S. for at leas President George Washington (1725-1799) <u>Johnna Afferson</u> (1745-1826) <u>James Madison</u> (1751-1836) James Monree (1758-1831) <u>Johnson (1757-1848)</u>		s, a perso N F D D	on car lone, F loderali Democr	Federal ist ratic-R ratic-R	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1 1	s of age, t erm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C	rn citia dams s Jeffer Burr, G c Clintor Tompk alboun	son corge C n. Elbris ins	linte Ige G	Vice-Pr on Geny				and the second
een residents of the U.S. for at leas President George Washington (1732-1799) Ashn Adams (1735-1826) Thomas Afferens (1748-1836) James Maduon (1748-1836) James Mourse (1748-1888) Anders Askow (1767-1845)		o, a perso N F D D D D D D	on car lone, F oderadi Democr Democr Democr Democr	Federal ist ratic-R ratic-R ratic-R ratic-R	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1 1 1 1	s of age, t erm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829 829-1837	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C	rn citia dams s.Jeffer Burr, G c Clintor Tompk alboun alboun	son corpe C n, Elbric ins Martin	linte Ige G	Vice-Pr on Geny				and the second
een residents of the U.S. for at leas President George Washington (1723-1799) - Abin Adam (1753-1826) - Immun Jefferson (1743-1826) - James Monieu (1758-1830) - John Quinez, Adams (1767-1848) - Andrew Jackson (1767-1848) - Andrew Jackson (1767-1848)		s, a perso N F D D D D D D D D	on car Sone, F Sederadi Democr Democr Democr Democr Democr	Federal ist ratic-R ratic-R ratic-R ratic-R	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1 1 1 1 1 1	s of age, t rm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829 829-1837 837-1841	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C Richard	dams s Jeffers Burr, G Clintor Tompk alboun alboun, d Johnso	son corpe C n, Elbric ins Martin	linte Ige G	Vice-Pr on Geny				and the second
een residents of the U.S. for at leas President George Washington (1722-1799) <u>John Adams</u> (1735-1826) <u>James Meinne</u> (1735-1826) <u>James Moinne</u> (1735-1830) <u>Janes Moinne</u> (1737-1848) <u>Jandius Aakson</u> (1707-1848) <u>Mairin van Bieren</u> (1723-1841)		s, a perse N D D D D D V W	on car fone, F federali Democr Democr Democr Democr Democr Democr	Federal ist ratic-R ratic-R ratic-R ratic-R	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third te 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	s of age, t rm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829 829-1837 837-1841 841	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C	dams s Jeffers Burr, G Clintor Tompk alboun alboun, d Johnso	son corpe C n, Elbric ins Martin	linte Ige G	Vice-Pr on Geny				and the second
een residents of the U.S. for at leas President George Washington (172-1799) - 6th Adam (173-1850) - Inneus Adams (173-1850) - Inneus Matalum (173-1833) - 8th Obies, Adams (178-1848) - Andrew Jackson (177-1848) - Martin van Bieren (172-1842) - Martin van Bieren (172-1842) - Martin van Bieren (172-1842) - Martin van Bieren (172-1842)		s, a perse N D D D D W W	on car federali Democr Democr Democr Democr Democr Democr Democr Nhig Whig	Federal ist ratic-R ratic-R ratic-R ratic-R rat	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third te 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	s of age, t erm as Pre 789-1797 797-1801 801-1809 809-1817 817-1825 829-1837 837-1841 841 841-1845	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C Richary John T	en citia dams s Jeffer Burr, G : Clinto Tompk alboun alboun, d Johnso yler	son corge C n. Elbric ins Martin on	linte Ige G	Vice-Pr on Geny				and the second
een residents of the U.S. for at leas President George Washington (1723-1799) - John Adams (1735-1826) Thomas Affreem (1745-1826) - James Monies (1758-1831) - John Quinez, Adams (1767-1848) - Andrew Ackson (1767-1848) - Andrew Ackson (1782-1842) - William H. Harrison (1723-1841) 0. John Tyle (1796-1849) 1. James K. Pok (1796-1849)		o, a perso N F D D D D D W W W	on car lone, F ederali Democr Democr Democr Democr Democr Democr Whig Democr	Federal ist ratic-R ratic-R ratic-R ratic-R rat	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	s of age, t erm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829 825-1829 837-1841 841 841-1845 845-1849	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C Richard John T George	en citia dams s Jeffer Burr, G c Clinto Tompk alboun, d Johnso ykr : Dallas	son corpe C n. Elbric ins Martin on	linte Ige G	Vice-Pr on Geny				and the second
een residents of the U.S. for at leas President (172-1799) - demarkanington (173-1799) - demarkanington (173-1786) - James Mainon (173-1786) - James Mainon (173-1786) - James Mainon (173-1845) - Andrew Jakeou (1775-1845) - Martin van Buern (1782-1802) - Wallman H. Harrow (1772-1844) - Onderw Jakeou (1772-1845) - Danley (1790-1802) 1. James K. Polk (1792-1849) 1. James K. Polk (1792-1849)		s, a perse N D D D D D W W W W W	on car Sone, F Sederali Democr Democr Democr Democr Democr Democr Whig Whig Democr Whig	Federal ist ratic-R ratic-R ratic-R ratic-R rat	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	s of age, t rrm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829 829-1837 837-1841 841 841-1845 845-1849 849-1850	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C Richard John T George	en citia dams s Jeffer Burr, G : Clinto Tompk alboun alboun, d Johnso yler	son corpe C n. Elbric ins Martin on	linte Ige G	Vice-Pr on Geny				and the second
he President and Vice-President are een residents of the U.S. for at leas een residents of the U.S. for at leas 1. <u>George Washington</u> (1732-1789) 2. <u>Borns Adlance</u> (1735-1876) 3. <u>Dannes Allecton</u> (1745-1876) 5. <u>Janes Malone</u> (1751-1876) 5. <u>Janes Malone</u> (1775-1878) 5. <u>Andres Jackow</u> (1775-1878) 5. <u>Martin van Bieren (1775-1884)</u> 5. <u>Martin van Bieren (1775-1884)</u> 10. <u>John Tyker (1796-1892)</u> 11. <u>Janes K. Folk (1796-1892)</u> 11. <u>Janes K. Folk (1796-1892)</u> 11. <u>Janes K. Folk (1796-1897)</u> 12. <u>Sachur Tyker (1796-1897)</u> 13. <u>Millard H. Harrow</u> (1896-1874) 15. <u>Millard H. Harrow</u> (1896-1874)		s, a perse N F D D D D D D V W W W W	on car lone, F ederali Democr Democr Democr Democr Democr Democr Whig Democr	nnot b Federal ist ratic-R ratic-R ratic-R ratic-R rat rat rat	Part Part dist tepubl tepubl	e at leas ected to ty lican lican	35 year a third to 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	s of age, t erm as Pre Term a 789-1797 797-1801 801-1809 809-1817 817-1825 825-1829 825-1829 837-1841 841 841-1845 845-1849	hey mu sident.)	st be no	John A John A Thoma Aaron George Daniel John C John C Richard John T George	rn citiz danis « Jeffer Burr, G Clinto Tompk alhoun alhoun, J Johnso J Johnso Yer : Dallas Filluso	son corpe C n. Elbric ins Martin on	linte Ige G	Vice-Pr on Geny				and the second

Figure 1. Typical use of the table tag to describe relational data that has structure never explicitly declared by the author, including metadata consisting of several typed and labeled columns, but that is obvious to human observers. The navigation bars at the top of the page are also implemented through the table tag but do not contain relational-style data.

Status         Status<	Control         Control <t< th=""><th>abort to         show           Rank(1)         City /           V00         City /           City /         City /</th><th>entrine table ////ban area(b) //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight //weight ///weight //weig</th><th>Country Japan USA USA USA Baeti Rosa Japan Philippine USA USA</th><th>Pepulation(1) 33.000.0003.0007 47.000.0003.7007 47.000.0003.7007 47.000.0003.7007 47.000.0003.4007 47.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.2007</th><th>225) Land area (n sqk(n)(1, 2) #.8399830 #.899830 #.899830 #.899830 #.899800 #.899800 #.899800 #.399780 #.399780 #.399780 #.399780 #.399780 #.399780</th><th>cay/u</th><th>Plot Map</th><th>Tables 1 - 10 of 15 Line Graph flar Graph Rank City / Brban area County Population Land area (n sqKm) Density (people per</th><th>20(20.0) Paris France 9,645,000,0645000.0 2,723(2723.0)</th><th>8</th></t<>	abort to         show           Rank(1)         City /           V00         City /           City /         City /	entrine table ////ban area(b) //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight //weight ///weight //weig	Country Japan USA USA USA Baeti Rosa Japan Philippine USA USA	Pepulation(1) 33.000.0003.0007 47.000.0003.7007 47.000.0003.7007 47.000.0003.7007 47.000.0003.4007 47.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.2007	225) Land area (n sqk(n)(1, 2) #.8399830 #.899830 #.899830 #.899830 #.899800 #.899800 #.899800 #.399780 #.399780 #.399780 #.399780 #.399780 #.399780	cay/u	Plot Map	Tables 1 - 10 of 15 Line Graph flar Graph Rank City / Brban area County Population Land area (n sqKm) Density (people per	20(20.0) Paris France 9,645,000,0645000.0 2,723(2723.0)	8
Autorest Laneat cities in the world by nonulation (1 to 123)           Date is         Image: Imag	Normality         Normality <t< th=""><th>show         show           Rask(1)         City /           Villa         City /           Villa            2000</th><th>entrine table ////ban area(b) //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight //weight ///weight //weig</th><th>Country Japan USA USA USA Baeti Rosa Japan Philippine USA USA</th><th>Pepulation(1) 33.000.0003.0007 47.000.0003.7007 47.000.0003.7007 47.000.0003.7007 47.000.0003.4007 47.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.2007</th><th>Land area (in sk(A)(1, 2) Association (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (1) (A)(1) (1)</th><th>cay/u</th><th>and the second se</th><th>Line Graph Bar Graph Rank City / Whoin area County Population Land area (in sp(m) Density (people per</th><th>20(20.0) Paris France 9,645,000,0645000.0 2,723(2723.0)</th><th>8</th></t<>	show         show           Rask(1)         City /           Villa         City /           Villa            2000	entrine table ////ban area(b) //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight //weight ///weight //weig	Country Japan USA USA USA Baeti Rosa Japan Philippine USA USA	Pepulation(1) 33.000.0003.0007 47.000.0003.7007 47.000.0003.7007 47.000.0003.7007 47.000.0003.4007 47.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.2007	Land area (in sk(A)(1, 2) Association (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (1) (A)(1) (1)	cay/u	and the second se	Line Graph Bar Graph Rank City / Whoin area County Population Land area (in sp(m) Density (people per	20(20.0) Paris France 9,645,000,0645000.0 2,723(2723.0)	8
Status         Status<	Normality         Normality <t< th=""><th>show         show           Rask(1)         City /           Villa         City /           Villa            2000</th><th>entrine table ////ban area(b) //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight //weight ///weight //weig</th><th>Country Japan USA USA USA Baeti Rosa Japan Philippine USA USA</th><th>Pepulation(1) 33.000.0003.0007 47.000.0003.7007 47.000.0003.7007 47.000.0003.7007 47.000.0003.4007 47.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.2007</th><th>Land area (in sk(A)(1, 2) Association (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (1) (A)(1) (1)</th><th>cay/u</th><th>and the second se</th><th>Rank City / Urban area County Population Land area (in sqKm) Density (people per</th><th>20(20.0) Paris France 9.645.000.0645000.0 2.723(2723.0)</th><th>8</th></t<>	show         show           Rask(1)         City /           Villa         City /           Villa            2000	entrine table ////ban area(b) //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight //aksiane //weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight ///weight //weight ///weight //weig	Country Japan USA USA USA Baeti Rosa Japan Philippine USA USA	Pepulation(1) 33.000.0003.0007 47.000.0003.7007 47.000.0003.7007 47.000.0003.7007 47.000.0003.4007 47.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.4007 43.000.0003.2007	Land area (in sk(A)(1, 2) Association (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (A)(1) (1) (1) (A)(1) (1)	cay/u	and the second se	Rank City / Urban area County Population Land area (in sqKm) Density (people per	20(20.0) Paris France 9.645.000.0645000.0 2.723(2723.0)	8
Chryl         Chryl         Pspulation(1)         Land area (10)         Court           100         Tage Theorem         Association         Association         Court         Pspulation(1)         Court         Court         Pspulation(1)         Court         Court <td< td=""><td>Chryl         Chryl         Pegulation (1)         Land area (10)         Corr           100         Taget Yankana         Jaget Status (1)         Land area (10)         Corr           200         Taget Yankana         Jaget Status (1)         Land area (10)         Corr           2010         Taget Yankana         Jaget Status (1)         Land area (10)         Corr           2010         Taget Yankana         Galaxies (1)         Land area (10)         Corr           2020         Taget Yankana         Galaxies (1)         Land area (10)         Corr           2030         Taget Yankana         Galaxies (1)         Land area (10)         Corr         <td< td=""><td>Chip         Chip           \$1,00         20,00           20,00         20,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,000         44,00</td><td>f Urban sreat) Téger Valatana Neuro Kalan Saraya Sarahashan Nanis Gy Casalinakan Manan Ma</td><td>Japan USA USA USA USA USA Japan Philophone India USA USA USA</td><td>33,500,000(3,3087) (7,500,000(1,7187) (7,700,000(1,7187) (7,700,000(1,7187) (7,400,000(1,7187) (4,200,000(1,7187) (4,200,000(1,4087) (4,200,000(1,4087) (1,200,000(1,4087) (1,200,000(1,2187) (1,200,000(1,2187)</td><td>1,24(xn)(1,22) 6,240,00030) 1,240,00030) 1,240,00030) 1,240,00030 1,240,000 1,240</td><td>cay/u</td><td>and the second se</td><td>Rank City / Urban area County Population Land area (in sqKm) Density (people per</td><td>20(20.0) Paris France 9.545.000(9645000.0 2.723(2723.0)</td><td>a 14</td></td<></td></td<>	Chryl         Chryl         Pegulation (1)         Land area (10)         Corr           100         Taget Yankana         Jaget Status (1)         Land area (10)         Corr           200         Taget Yankana         Jaget Status (1)         Land area (10)         Corr           2010         Taget Yankana         Jaget Status (1)         Land area (10)         Corr           2010         Taget Yankana         Galaxies (1)         Land area (10)         Corr           2020         Taget Yankana         Galaxies (1)         Land area (10)         Corr           2030         Taget Yankana         Galaxies (1)         Land area (10)         Corr         Corr <td< td=""><td>Chip         Chip           \$1,00         20,00           20,00         20,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,000         44,00</td><td>f Urban sreat) Téger Valatana Neuro Kalan Saraya Sarahashan Nanis Gy Casalinakan Manan Ma</td><td>Japan USA USA USA USA USA Japan Philophone India USA USA USA</td><td>33,500,000(3,3087) (7,500,000(1,7187) (7,700,000(1,7187) (7,700,000(1,7187) (7,400,000(1,7187) (4,200,000(1,7187) (4,200,000(1,4087) (4,200,000(1,4087) (1,200,000(1,4087) (1,200,000(1,2187) (1,200,000(1,2187)</td><td>1,24(xn)(1,22) 6,240,00030) 1,240,00030) 1,240,00030) 1,240,00030 1,240,000 1,240</td><td>cay/u</td><td>and the second se</td><td>Rank City / Urban area County Population Land area (in sqKm) Density (people per</td><td>20(20.0) Paris France 9.545.000(9645000.0 2.723(2723.0)</td><td>a 14</td></td<>	Chip         Chip           \$1,00         20,00           20,00         20,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,00         44,00           20,000         44,00	f Urban sreat) Téger Valatana Neuro Kalan Saraya Sarahashan Nanis Gy Casalinakan Manan Ma	Japan USA USA USA USA USA Japan Philophone India USA USA USA	33,500,000(3,3087) (7,500,000(1,7187) (7,700,000(1,7187) (7,700,000(1,7187) (7,400,000(1,7187) (4,200,000(1,7187) (4,200,000(1,4087) (4,200,000(1,4087) (1,200,000(1,4087) (1,200,000(1,2187) (1,200,000(1,2187)	1,24(xn)(1,22) 6,240,00030) 1,240,00030) 1,240,00030) 1,240,00030 1,240,000 1,240	cay/u	and the second se	Rank City / Urban area County Population Land area (in sqKm) Density (people per	20(20.0) Paris France 9.545.000(9645000.0 2.723(2723.0)	a 14
Cash (U         Cash (U         Page (2014)         Page (2014)         Page (2014)         Page (2014)           V10         Tear Packs         Area         9.202.0002.007)         0.8060000         0.8060000         0.8060000         0.8060000         0.8060000         0.8060000         0.8060000         0.80600000         0.8060000         0.8060000         0.8060000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.80600000         0.8060000000000         0.8060000000000000000000000000000000000	Calcular	11.0) 30.04 44.0 60.0 46.0 46.0 46.0 46.0 46.0	Figure Visitanan Bace Yield Mole Sacetfinaten Unanan Oly Danaktinet Hyte Manita Manita Manita Baliti Janat Lago Kadas Lago Kadas Kadas Kadas Kadas Kadas Kadas Kadas Kadas	Japan USA USA USA USA USA Japan Philophone India USA USA USA	33,500,000(3,3087) (7,500,000(1,7187) (7,700,000(1,7187) (7,700,000(1,7187) (7,400,000(1,7187) (4,200,000(1,7187) (4,200,000(1,4087) (4,200,000(1,4087) (1,200,000(1,4087) (1,200,000(1,2187) (1,200,000(1,2187)	1,24(xn)(1,22) 6,240,00030) 1,240,00030) 1,240,00030) 1,240,00030 1,240,000 1,240		riban prea	Rank City / Urban area County Population Land area (in sqKm) Density (people per	Paris France 9 (545.000(9545000.0 2.723(2723.0)	a 14
44.40         Issuence         Description         Description <thdescription< th=""> <thdesc< th=""><th>44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000</th><th>2000 2000 4400 6000 7000 8000 1000 1000 1000 1000 1000 1</th><th>New York Mater Sear Parke Searchtowen Deutes City Deathford Type Manita</th><th>USA Beatl Secti Kasea Japan Palippines India Ind</th><th>17,800,800(17887) 17,700,800(17887) 17,700,800(17887) 17,700,800(17887) 14,401,800(17887) 14,400,800(14887) 14,200,800(14887) 14,200,800(14887) 14,200,800(14887) 14,200,800(14887) 14,200,800(14887) 12,200,800(12887) 12,200,800(12887)</th><th>8,403,663,03 (, 4,404,1646,03) (, 2,404,264,04) 2,2072,0072,03 (, 2,404,264,04) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03)</th><th></th><th></th><th>City / Urban area Country Population Land area (in sqKm) Density (people per</th><th>Paris France 9 (545.000(9545000.0 2.723(2723.0)</th><th>a 14</th></thdesc<></thdescription<>	44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000	2000 2000 4400 6000 7000 8000 1000 1000 1000 1000 1000 1	New York Mater Sear Parke Searchtowen Deutes City Deathford Type Manita	USA Beatl Secti Kasea Japan Palippines India Ind	17,800,800(17887) 17,700,800(17887) 17,700,800(17887) 17,700,800(17887) 14,401,800(17887) 14,400,800(14887) 14,200,800(14887) 14,200,800(14887) 14,200,800(14887) 14,200,800(14887) 14,200,800(14887) 12,200,800(12887) 12,200,800(12887)	8,403,663,03 (, 4,404,1646,03) (, 2,404,264,04) 2,2072,0072,03 (, 2,404,264,04) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03) (, 2,404,256,03)			City / Urban area Country Population Land area (in sqKm) Density (people per	Paris France 9 (545.000(9545000.0 2.723(2723.0)	a 14
44.40         Issuence         Description         Description <thdescription< th=""> <thdesc< td=""><td>44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000</td><td>3(3.2) 4(4.0) 6(5.2) 4(6.2) 4(6.2) 4(6.2) 4(6.2) 4(6.2) 4(1.2) 4(</td><td>Bat Pado Senditanean Manito Ciry Oseaakani Yayao Manita Manita Baha Jakata Lagro Kakata Cagro Kakata Cagro Kakata Cagro Kakata Cagro Kakata Sensee Alma</td><td>Badi Sech Jana Marina Japan Philiphen India Indi</td><td>17.200.000(1.2787) 17.400.000(1.2787) 17.400.000(1.2787) 14.420.000(1.4687) 14.420.000(1.4687) 14.300.000(1.4687) 14.300.000(1.4687) 14.300.000(1.4687) 14.300.000(1.4687) 14.200.000(1.2887) 12.200.000(1.2887)</td><td>1,046(1466.0) (,045(046.0) 2,075(072.0) 2,044(264.6) (3,344(1566.0) 3,344(1566.0) 1,244(1566.0) 7,36(1566.0) 7,36(1566.0) 1,344(1566.0)</td><td></td><td>, I</td><td>Country Population Land area (in sqKm) Density (people per</td><td>France 9.645.000(9645000.0 2.723(2723.0)</td><td>100</td></thdesc<></thdescription<>	44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000	3(3.2) 4(4.0) 6(5.2) 4(6.2) 4(6.2) 4(6.2) 4(6.2) 4(6.2) 4(1.2) 4(	Bat Pado Senditanean Manito Ciry Oseaakani Yayao Manita Manita Baha Jakata Lagro Kakata Cagro Kakata Cagro Kakata Cagro Kakata Cagro Kakata Sensee Alma	Badi Sech Jana Marina Japan Philiphen India Indi	17.200.000(1.2787) 17.400.000(1.2787) 17.400.000(1.2787) 14.420.000(1.4687) 14.420.000(1.4687) 14.300.000(1.4687) 14.300.000(1.4687) 14.300.000(1.4687) 14.300.000(1.4687) 14.200.000(1.2887) 12.200.000(1.2887)	1,046(1466.0) (,045(046.0) 2,075(072.0) 2,044(264.6) (3,344(1566.0) 3,344(1566.0) 1,244(1566.0) 7,36(1566.0) 7,36(1566.0) 1,344(1566.0)		, I	Country Population Land area (in sqKm) Density (people per	France 9.645.000(9645000.0 2.723(2723.0)	100
44.40         Issuence         Description         Description <thdescription< th=""> <thdesc< td=""><td>44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000</td><td>4(40) 6(60) 6(70) 6(</td><td>Secultoritoria Manica City (Insubinet rispo Manita Manita Manita Jakata Lapro Kokas Carro Lar Argeto Revens Atoo</td><td>Soch Kong Marico Japan Phispheo India Indi</td><td>(7.00.00(1.787) 17.400.000(1.747) 14.420.00(1.64517) 14.470.000(1.4517) 14.300.000(1.4517) 14.200.000(1.4517) 13.400.000(1.3417) 13.400.000(1.2417) 12.200.000(1.2217)</td><td>1,049(1048,0) 2,073(007,2,0) 2,284(2064,4) 1,284(1060,0) 1,284(1060,0) 1,284(1000,0) 1,284(1000,0) 1,284(1000,0) 1,284(1000,0) 1,284(1000,0)</td><td></td><td>4 L 4</td><td>Population Land area (in sqKm) Density (people per</td><td>9 j545.000(9645000.0 2.723(2723.0)</td><td>100</td></thdesc<></thdescription<>	44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000	4(40) 6(60) 6(70) 6(	Secultoritoria Manica City (Insubinet rispo Manita Manita Manita Jakata Lapro Kokas Carro Lar Argeto Revens Atoo	Soch Kong Marico Japan Phispheo India Indi	(7.00.00(1.787) 17.400.000(1.747) 14.420.00(1.64517) 14.470.000(1.4517) 14.300.000(1.4517) 14.200.000(1.4517) 13.400.000(1.3417) 13.400.000(1.2417) 12.200.000(1.2217)	1,049(1048,0) 2,073(007,2,0) 2,284(2064,4) 1,284(1060,0) 1,284(1060,0) 1,284(1000,0) 1,284(1000,0) 1,284(1000,0) 1,284(1000,0) 1,284(1000,0)		4 L 4	Population Land area (in sqKm) Density (people per	9 j545.000(9645000.0 2.723(2723.0)	100
44.40         Issuence         Description         Description <thdescription< th=""> <thdesc< td=""><td>44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000</td><td>05.0) 05.0) 70.0) 06.0) 06.0) 10(10) 10(</td><td>Marine City Statabilities Types Manita Baths Jatats Capy Capy Capy Capy Capy Capy Capy Capy</td><td>Mantos Japan Philippines India India Nigeria India Egypt USA</td><td>17,400,000(15,74E7) 18,405,000(1,6405E7) 18,200,000(1,44987) 18,300,000(1,44987) 18,300,000(1,44987) 18,400,000(1,40887) 12,400,000(1,42987) 12,500,000(1,2287)</td><td>2.0732072.03 2.8%420%40 1.3%1(1500.05 %4%44.05 1.2%1(1500.05) 7.3%1(1500.05) 7.3%13.05 9.3%13.05 1.2%1(1506.05)</td><td></td><td>, La</td><td>Land area (in sqKm) Density (people per</td><td>2.723(2723.0)</td><td>100</td></thdesc<></thdescription<>	44.40         Exertine is and in the interval         0.70000001787         1.246(960)           9.50         Munice Givy         Munice Givy         0.70000001787         2.270007           9.50         Munice Givy         Munice Givy         0.70000001787         2.27007           9.50         Munice Givy         Munice Givy         0.7000001787         2.27007           9.50         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.50         Munice Givy         Munice Givy         Munice Givy         2.27007         Laced state fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.250000         State fields (end)         2.20007           9.60         Munice Givy         Munice Givy         Munice Givy         2.20007         State fields (end)         2.20007           9.01         Addes         Munice Givy         Munice Givy         1.2000000000         State fields (end)         2.2000000000           9.010         Munice Givy         Munice Givy         Munice Givy         2.20000000000         Givy         4.2000000000000000000000000000000000000	05.0) 05.0) 70.0) 06.0) 06.0) 10(10) 10(	Marine City Statabilities Types Manita Baths Jatats Capy Capy Capy Capy Capy Capy Capy Capy	Mantos Japan Philippines India India Nigeria India Egypt USA	17,400,000(15,74E7) 18,405,000(1,6405E7) 18,200,000(1,44987) 18,300,000(1,44987) 18,300,000(1,44987) 18,400,000(1,40887) 12,400,000(1,42987) 12,500,000(1,2287)	2.0732072.03 2.8%420%40 1.3%1(1500.05 %4%44.05 1.2%1(1500.05) 7.3%1(1500.05) 7.3%13.05 9.3%13.05 1.2%1(1506.05)		, La	Land area (in sqKm) Density (people per	2.723(2723.0)	100
Open Constraint         Open Const	Open         Open <th< td=""><td>68.0) 77.0) 68.0) 98.0) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10)</td><td>Subahara Papis Manta Manta Daha Jakata Lapa Kakata Cana Cana Euro Angelos Rus de Janata</td><td>Japan Philippines India India India India India India India India India India India India India India India India India</td><td>14.401.000(1.542187) 14.700.000(1.47987) 14.300.000(1.4887) 14.300.000(1.4887) 14.200.000(1.4887) 13.400.000(1.4877) 13.400.000(1.3487) 12.200.000(1.2877)</td><td>2.84420448) 1.304(1564.6) 4.844448) 1.204(1568.0) 1.304(1560.0) 5.304(1560.0) 614(81.6) 1.204(1560.0)</td><td></td><td>T. Hereit</td><td>Density (people per</td><td></td><td>1000</td></th<>	68.0) 77.0) 68.0) 98.0) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10) 19(10)	Subahara Papis Manta Manta Daha Jakata Lapa Kakata Cana Cana Euro Angelos Rus de Janata	Japan Philippines India India India India India India India India India India India India India India India India India	14.401.000(1.542187) 14.700.000(1.47987) 14.300.000(1.4887) 14.300.000(1.4887) 14.200.000(1.4887) 13.400.000(1.4877) 13.400.000(1.3487) 12.200.000(1.2877)	2.84420448) 1.304(1564.6) 4.844448) 1.204(1568.0) 1.304(1560.0) 5.304(1560.0) 614(81.6) 1.204(1560.0)		T. Hereit	Density (people per		1000
1700         Made         Philgree         14420000-4787         13200000         1           688         Mades         Made         Mades         Mades <t< td=""><td>170         Mede         PAlgene         14420000-4787         13900000         1           08.00         Manda         Mada         Mada</td><td>7(7.0) B(0.0) B(0.0) 15(11.0) 15(12.0) 15</td><td>Manita Manita Balini datata Capin Kokata Canin Lan Angatan Reason Xiao Pite da Janata</td><td>Philippines India India India India India India Egypt USA</td><td>14.700.000(1.47987) 14.300.000(1.4987) 14.300.000(1.4987) 14.200.000(1.4987) 19.400.000(1.347) 12.700.000(1.3787) 12.700.000(1.2787)</td><td>1,306(1300.0) 484(04.0) 1,206(1300.0) 3,306(1300.0) 536(730.0) 634(01.0) 1,206(1206.0)</td><td></td><td></td><td>Density (people per settin)</td><td>3.550(3550.0)</td><td>9 9</td></t<>	170         Mede         PAlgene         14420000-4787         13900000         1           08.00         Manda         Mada	7(7.0) B(0.0) B(0.0) 15(11.0) 15(12.0) 15	Manita Manita Balini datata Capin Kokata Canin Lan Angatan Reason Xiao Pite da Janata	Philippines India India India India India India Egypt USA	14.700.000(1.47987) 14.300.000(1.4987) 14.300.000(1.4987) 14.200.000(1.4987) 19.400.000(1.347) 12.700.000(1.3787) 12.700.000(1.2787)	1,306(1300.0) 484(04.0) 1,206(1300.0) 3,306(1300.0) 536(730.0) 634(01.0) 1,206(1206.0)			Density (people per settin)	3.550(3550.0)	9 9
1770         Markat         Fridgerer         14/20.0001, effer         13.0001001           580         Market         Market         14/20.0001, effer         0.4400.0001, effer         0.4400.0000, effer	17/20         Marka         Philippere         Nature (Marka)         <	88.0) 99.0) 15(10) 16(10) 16(10) 16(10) 16(10) 16(10) 16(10) 16(10) 16(10) 16(10) 16(10) 16(10)	Manshai Dudhi Jakans Lapro Kokais Cario Lare Angeleo Beasen Anno Per de Janeiro	India India India Nigenia India Egypt USA	14,360,000(1,48627) 14,300,000(1,4867) 14,250,000(1,42627) 13,400,000(1,3467) 12,786,000(1,2467) 12,786,000(1,2767) 12,200,000(1,2267)	484,494.0) 1,266(1565.0) 1,368(1300.0) 7,26(738.0) 634,631.0) 1,286(1266.0)		Ę.	September 201	9 9 9	9 9 9
000         Hote         Hote <thh< td=""><td>000         both         166         1.00001.407         1.24(126.0)           0(16)         JJJAN         Milana         Milana         Milana         Milana           0(17)         JJAN         Milana         Milana         Milana         Milana         Milana           0(17)         JJAN         Milana         Milana         Milana         Milana         Milana           0(12)         Jakata         Milana         Milana         Milana         Milana         Milana           0(12)         Jakata         Milana         Milana</td><td>9(9.0) 15(10.0) 16(11.0) 42(12.0) 16(13.0) 16(14.0) 16(14.0) 17(17.0) 17(17.0) 17(17.0) 19(18.0)</td><td>Build Jakata Lagm Kalaka Cam Cam Lan Aquina Bunnet Alma Rin de Janeiro</td><td>India Indonesia Nigeta India Egypt USA</td><td>14.300.000(1.4387) 14.250.000(1.42887) 13.400.000(1.3487) 12.700.000(1.2787) 12.200.000(1.2787)</td><td>1,266(1296.0) 1,266(1360.0) 736(736.0) 834(831.0) 1,266(1266.0)</td><td></td><td>4</td><td>19</td><td>9.9</td><td>9 9</td></thh<>	000         both         166         1.00001.407         1.24(126.0)           0(16)         JJJAN         Milana         Milana         Milana         Milana           0(17)         JJAN         Milana         Milana         Milana         Milana         Milana           0(17)         JJAN         Milana         Milana         Milana         Milana         Milana           0(12)         Jakata         Milana         Milana         Milana         Milana         Milana           0(12)         Jakata         Milana	9(9.0) 15(10.0) 16(11.0) 42(12.0) 16(13.0) 16(14.0) 16(14.0) 17(17.0) 17(17.0) 17(17.0) 19(18.0)	Build Jakata Lagm Kalaka Cam Cam Lan Aquina Bunnet Alma Rin de Janeiro	India Indonesia Nigeta India Egypt USA	14.300.000(1.4387) 14.250.000(1.42887) 13.400.000(1.3487) 12.700.000(1.2787) 12.200.000(1.2787)	1,266(1296.0) 1,266(1360.0) 736(736.0) 834(831.0) 1,266(1266.0)		4	19	9.9	9 9
0,110         J.d.M.         Makes         14.55.000.4.007         1.30(100.0         I           1(11.0)         4.39         Bares         1.346.000.1.307         70700.00         I           1(11.0)         4.39         Bares         1.346.000.0.207         70700.00         I           1(11.0)         6.39         Bares         1.346.000.0.207         1.70700.00         I           1(11.0)         Cares         Earse         1.330.000.0.207         1.340.0000         I         Image: States	GYU10         Jacks         Makes         1450000-0207         Jacky 10000         I           1Y113         Aaps         Bars         13.400000-0207         T/0700.00         I           1Y113         Aaps         Bars         13.400000-0207         T/0700.00         I           1Y113         Aaps         Bars         13.400000-0207         T/0700.00         I         Image: State Sta	02(10.0) 19(11.0) 12(13.0) 12(13.0) 14(14.0) 15(18.0) 15(18.0) 15(18.0) 15(18.0) 16(18.0) 16(18.0)	Jakata Lagos Kalkata Callo Late Angeles Beanus Alles Rio de Janeiro	Informatio Nigeria India Egypt USA	14250.000(1.42587) 13.400.000(1.3487) 12.700.000(1.2787) 12.200.000(1.2287)	1,200(1300,0) 738(738,0) 834(831,0) 1,296(1266,0)		đ		9 9 9	9
NY10         Lage         Page         1.640.0001340°         7.707.00         1.707.000           0x12.0         Kakak         Mark         0.200.001240°         Payline         Payline           1x12.0         Cam         Signed         0.200.001240°         Payline         Payline           1x12.0         Cam         Signed         0.200.001240°         Payline         Payline           1x12.0         Lin Againe         0.200.001240°         Payline         Payline         Payline           1x12.0         Market         Base         Bibbo 200.001240°         Payline         Payline         Payline           1x17.0         Base         Bibbo 200.001240°         Payline         Payline         Payline         Payline           1x17.0         Base         Bibbo 200.001240°         Payline	NY10         Laps         Neges         3.460.000134000         7.707.000         1           NY10         Lank again	11(11.0) 12(12.0) 13(13.0) 14(14.0) 15(18.0) 15(18.0) 17(17.0) 18(18.0) 13(18.0)	Lagen Kabata Carin Late Angeles Busines Alter Rise de Janeiro	Nigeria India Egyst VDA	13,400,000(1,34E7) (2,700,000(1,37E7) (2,200,000(1,22E7)	736(738.0) 834(831.0) 1,266(1266.0)				9.9	
Status         Status<	Str.20         Str.80         Str.90         Str.90<	42(12.0) 19(13.0) 14(14.0) 19(18.0) 19(18.0) 17(17.0) 19(18.0) 19(18.0)	Kabata Carro Los Angeles Buenes Aires Ris de Janeiro	india Egyst USA	12,700,000(1.2787) 12,200,000(1.2287)	833(831.0) 1,295(1295.0)			- The second second	- 9 Y	Q.
Bit No.         Figure         12.00.00011287         1.244(306.0)           16(3.4)         Lan Agains         U.M. (11/96.00011087)         4.3596.0201           16(3.4)         Lan Agains         U.M. (11/96.00011087)         2.3596.0201           10(3.6)         Re 64 assess         Buail         9.0000010871         2.3996.0001           10(3.8)         Re 64 assess         Buail         9.00000010871         2.3996.0000           10(3.8)         Re 64 assess         Buail         9.00000000001001         2.2976.0010           10(3.8)         Re 74 asses         Buail         9.000000000000000000000000000000000000	UX30         Came         Figs         U.20030122EF         1.244(306.0)           14(4.8)         Laf. Ageta         U.0.4         1/1/16001(19807)         4.359(326.0)           14(4.8)         Laf. Ageta         U.0.4         1/1/16001(19807)         4.359(326.0)           10(10)         Refs failed         Base         9.0000(1087)         1.289(106.0)           10(10)         Refs failed         Base         9.0000(1087)         1.289(106.0)           10(10)         Refs failed         Base         9.0000(1087)         2.199(106.0)           10(10)         Refs failed         Base         9.0000(1087)         2.199(106.0)           10(10)         Refs failed         9.0000(1087)         2.199(106.0)         1.299(106.0)           10(10)         Refs failed         9.0000(1087)         2.199(106.0)         1.299(106.0)           10(10)         Refs failed         9.0000(1087)         2.199(107.000)         1.299(107.000)           10(10)         Refs failed         1.199(108.0)         1.299(107.000)         1.299(107.000)           10(10)         Refs failed         1.490(109.00000)         1.299(107.000)         1.299(107.000)           10(20)         Linke         0.100000000000         1.1299(109.00000000)         1	19(13.0) 19(14.0) 19(18.0) 19(18.0) 19(18.0) 19(18.0) 19(18.0)	Cann Los Angeles Buenes Atles Rio de Janeiro	Egyrt USA	12,200,000(1.2287)	1,296(1296.0)			a course	Garante Outermant of	V
N1N0         Lak Agento         0.04         1/13/000/157807         4.2390000           01180         Basena Allano         Agento         1/13/2000/157807         2.2802260.01           01180         Basena Allano         1/13/2000/15077         2.2802260.01         International Allano           010180         Basena Allano         Basena Allano         1/13/2000/15077         2.2802260.01           010180         Basena Allano         Basena Allano         1/13/2000/15077         2.2802260.01           010180         Basena Allano         Basena Allano         1.0300/00010077         2.2802260.01           00180         Basena Allano         Basena Allano         Basena Allano         Edu Statosobornico Allano           00180         Pene         Farere         Basena Allano         Basena Allano         Basena Allano           00180         Pene         Farere         Basena Allano         Basena Allano         Basena Allano         Basena Allano           00180         Pene         Farere         Basena Allano         Basena Allano         Basena Allano         Basena Allano           00180         Pene         Farere         Basena Allano         Basena Allano         Basena Allano         Basena Allano         Basena Allano         Basena Allano	14140         Los Asguto         0.04         1/13/000/170877         4.2390000           10180         Basen Allas         Againa         11/22000/170877         4.2390000         I           10180         Bis d Asenso         Bis d 10/22000/18077         1.08(1060)         I         I           10170         Maxees         Bis d 10/22000/18077         1.08(1060)         I<	14(14.0) 15(18.0) 15(18.0) 17(17.0) 16(18.0) 16(18.0)	Las Angeles Buenes Alies Pier de Janeiro	USA					Same France		
K1000         Hanne (Mes)         Apple fragmentary         112200001 (107)         222002000           K1000         File of Assesso         Bild 000001 (107)         2.30000000         1	KYUR         Baser, Mail         Agenta         112000011007         224952805           KYUR         Ref & Samor         Biolito 200010007         2380000010007         2380000010007         2380000010007         2380000010007         2380000010007         238000000000         2480000000000         228000000000         228000000000         228000000000         2280000000000         228000000000         228000000000         228000000000         228000000000         228000000000         228000000000         2280000000000         2280000000000         2280000000000         2280000000000         2280000000000         2280000000000         2280000000000         2280000000000         2280000000000         2280000000000         228000000000         228000000000         2280000000000         2280000000000         2800000000000         28000000000000         28000000000000000000000000000000000000	15(18.0) 19(18.0) 17(17.0) 19(18.0) 19(18.0)	Buenes Alies Rie de Janeire		ALCONG CONTRACTORS					A Loom	Simourly .
NY10         Pie 44 Jahrs         Pied         H 80 JOB 2001 (MT)         1 JOB 1001 (MT)           V1010         Jahrs         Pied         Status	NYX0         Pie 44 James         Pied         H 80 J0000011807         1.380/1000.1           UYX0         America         Rodo 000011807         1.380/1000.1         1           UYX0         America         Rodo 000011807         1.280/1000.1         1           UYX0         America         Rodo 000011807         1.280/1000.1         1           UYX0         America         Rodo 000011807         1.280/1000.1         1           UYX0         America         Rodo 00000000         0.00000000000         0.000000000000           UYX0         America         Rodo 000000000         0.270/2000         2.270/2000           UYX0         America         America         America         2.276/2000         2.275/2000           UYX0         America         America         2.276/2000         2.277/2000         2.277/2000           D2020         America         America         2.276/2000         2.277/2000         2.277/2000         2.277/2000           D2020         America         America         2.276/2000         2.277/2000         2.277/2000         2.277/2000           D2020         America         America         2.27/200002/2000000         1.267/200002/20000000         2.277/200002/20000000000 <td< td=""><td>19(10.0) 17(17.0) 19(18.0) 18(18.0)</td><td>Ris de Janeiro</td><td>America</td><td>11/10/06(1.170007)</td><td>4,328(4020.0)</td><td></td><td></td><td>Errant, 1978</td><td>min mealth</td><td>anger No.</td></td<>	19(10.0) 17(17.0) 19(18.0) 18(18.0)	Ris de Janeiro	America	11/10/06(1.170007)	4,328(4020.0)			Errant, 1978	min mealth	anger No.
01720         Masse         Pass         0.05000010077         2.1002/00.05           01780         Disarged         Disar         Disarged         Disargedd	VICTO         Hease         P. Bas         0.0500000100077         22.1002/00.05           VICTO         Fanagari         Chana         VICTO         6000000         7000000           VICTO         Fanagari         Chana         VICTO         6000000         6000000           DOUTO         Period         Fanagari         0.000000000000         0.00000000000         0.0000000000000           DOUTO         Period         Fanagari         0.00000000000000000         0.00000000000000000000000000000000000	17(17.0) 19(18.0) 19(18.0)							the first of		Anne C-
0(136)         Description         0(136)         Particle         1700/061         Particle         1700/061         Particle         1700/061         Particle	6(3)(6)         Strauti         Conso         10000001487         7407-061           0(3)(6)         Fasario         F	19(18.0) 19(18.0)		Brazil	10,800,000(1.0887)	1,500(1500.0)			Campies Variation	THE YOUNG	ALC: NOT THE
GUND         Faces         Faces         Statume         Statu	KUM0         Ramel         Pallare         Statuse         Statuse <thstatuse< th=""> <thstatuse< th=""> <thstatus< td=""><td>43(18.0)</td><td>Metadee</td><td>Posta</td><td>10,500,000(1.05E7)</td><td>2.168(2150.0)</td><td></td><td>Ó</td><td>under diament</td><td></td><td>· Correction</td></thstatus<></thstatuse<></thstatuse<>	43(18.0)	Metadee	Posta	10,500,000(1.05E7)	2.168(2150.0)		Ó	under diament		· Correction
2018/0         Perio         Fanto         0.445.000644000.00         2.2720/27.01         Annu Marco         Annu Marco<	2003/0         Perio         Fuero         8.46.000644000.00         2.2720/27.81           2013/0         General         Internet         1.000110000         COLORIDADE         C		Thorghol	Etina	10,000,000(1.067)	745(746.0)		Y	Espatu o	And Address of the	East
JOLTA         Messes         Lawy         #2000000000000000000000000000000000000	JOCY 0         Meanse         Lawy         #2000000000000000000000000000000000000		Karashi	Pakistan	0.00000800000.001.0	S18(518.0)		Calles	- Usupt Strater		and a
20230         Averative         3/2010/0000000000         2.81*020*0.00           20230         Rejorgia         Onesa         6.81*0.000000000000         740*0400           202400         Oblinacija         UAA         6.81*0.0000000000000000000000000000000000	20220         Asyarya         Asyarya         Second         2.817020750           20230         Eleving         Oleva         8.811400001400000         7.6074400           202400         Olevage         UAA         8.811400001400000         4.6074400           202400         Olevage         UAA         8.0300000000000         4.6064400           202400         Olevage         UAA         6.27920002700000         4.6064400	13(20.0)		France			Cove	-	Case Dianan		
DQ230         Energy         Ones         6.11.400001-600.00         740/740.0           DQ340         Ones         VA         8.00.0000000.00         6.495466.0           DQ340         Exercise         VA         8.00.0000000.00         6.495466.0           DQ340         Exercise         VA         8.00.0000000.00         6.495466.0           DQ350         Exercise         VA         6.00.000000.00         1.4201400.00	DB230         Energy         Chara         6.47.450091-600.01         740740.0           DB2400         Charas         UA         5.00.0000000.00         6.464546.0           DB251         Landes         UA         6.30.00000000.00         6.495466.0							Contraction and	William and Annual and and	in the first of the second second	
SEC 40)         Chinespe         VDA         6.306.00000000)         6.446.446.0)           2005.00         Landow         36         6.376.0000270000.00)         1.4221402.00)	Skyladi         Chinespe         V/LA         6.306.000000000)         6.4465.446.0)           SDSE0         Landen         SK         6.376.0000270000.0)         1.4221403.0)										
2525.0) Lunders 04 6.274.000(274000.0) (A221402.0)	2505.0) Landam (W 8.279.000)279000 0) (8221623.0)										
	Hannes mannes maines maines maines	2525.01		11.1	8.278.000(8278000.0)	1,823(1623.0)					
none. more more more	*	-manal.	- 49984 m	(males)		man.					
•							•				
anne, mare, mare, mare, mare,				11.1			·				
		MATING CI	TY POPULATION	5							
MATING CITY POPULATIONS	MATING CITY POPULATIONS	ort to., show	r entire table								
		and the state	RE	GION P	eople per Hectare	Margin of Error	se l				
	ort to., show entire table		Citizes of A	Longing	955	36.249	-				
At to show entre table REDION People ser Hectore Margin of Error	ort to] (show entire table REGION People set Hectare Margin of Error		Cities	of Islam	250	29-244					

Figure 2. Results of a keyword query search for "city population," returning a relevanceranked list of databases. The top result contains a row for each of the most populous 125 cities and columns for "City/Urban Area," "Country," "Population," and "rank" (by population among all the cities in the world). The system automatically generated the image at right, showing the result of clicking on the "Paris" row. The title ("City Mayors...") links to the page where the original HTML table is located.

manual effort would be able to create a clean model of all of it.

Before addressing the challenges associated with accessing structured data on the Web, it is important to ask what users might do with such data. Our work is inspired by the following example benefits:

Improve Web search. Structured Web data can help improve Web search in a number of ways; for example, Deep Web databases are not generally available to search engines, and, by surfacing this data, a Deep Web exploration tool can expand the scope and quality of the Web-search index. Moreover, the layout structure can be used as a relevance signal to the search ranker; for example, an HTML table-embedded database with a column calories and a row latte, should be ranked fairly high in response to the user query latte calories. Traditionally, search engines use the proximity of terms on a page as a signal of relatedness; in this case, the two terms are highly related, even though they may be distant from each other on the page.

Enable question answering. A longstanding goal for Web search is to return answers in the form of facts; for example, in the latte calories query, rather than return a URL a search engine might return an actual numerical value extracted from the HTML table. Web search engines return actual answers for very specific query domains (such as weather and flight conditions), but doing so in a domain-independent way is a much greater challenge.

Enable data integration from multiple Web sources. With all the data sets available on the Web, the idea of combining and integrating them in ad hoc ways is immensely appealing. In a traditional database setting, this task is called data integration; on the Web, combining two disparate data sets is often called a "mashup." While a traditional database administrator might integrate two employee databases with great precision and at great cost, most combinations of Web data should be akin to Web searchrelatively imprecise and inexpensive; for example, a user might combine the set of coffeehouses with a database of WiFi hotspots, where speed

is more important than flawless accuracy. Unlike most existing mashup tools, we do not want users to be limited to data that has been prepared for integration (such as already available in XML).

The Web is home to many kinds of structured data, including embedded in text, socially created objects, HTML tables, and Deep Web databases. We have developed systems that focus on HTML tables and Deep Web databases. WebTables extracts relational data from crawled HTML tables, thereby creating a collection of structured databases several orders of magnitude larger than any other we know of. The other project surfaces data obtained from the Deep Web, almost all hidden behind Web forms and thus inaccessible. We have also constructed a tool (not discussed here) called Octopus that allows users to extract, clean, and integrate Web-embedded data.3 Finally, we built a third system, called Google Fusion Tables,<sup>13</sup> a cloud-based service that facilitates creation and publication of structured data on the Web, therefore complementing the two other projects.

## WebTables

The WebTables system<sup>4,5</sup> is designed to extract relational-style data from the Web expressed using the HTML table tag. Figure 1 is a table listing American presidents (http://www. enchantedlearning.com/history/us/ pres/list.shtml) with four columns, each with topic-specific label and type (such as **President** and **Term as President**) as a date range; also included is a tuple of data for each row. Although most of the structured-data metadata is implicit, this Web page essentially contains a small relational database anyone can crawl.

Not all table tags carry relational data. Many are used for page layout, calendars, and other nonrelational purposes; for example, in Figure 1, the top of the page contains a table tag used to lay out a navigation bar with the letters A–Z. Based on a human-judged sample of raw tables, we estimate up to 200 million true relational databases in English alone on the Web. In general, less than 1% of the content embedded in the HTML table tags represents good tables. In



## Any useful application making use of Web data must also be able to address uncertain data design and quality.

deed, the relational databases in the WebTables corpus form the largest database corpus we know of, by five orders of decimal magnitude.<sup>a</sup>

WebTables focuses on two main problems surrounding these databases: One, perhaps more obvious, is how to extract them from the Web in the first place, given that 98.9% of tables carry no relational data. Once we address this problem, we can move to the second—what to do with the resulting huge collection of databases.

Table extraction. The WebTables table-extraction process involves two steps: First is an attempt to filter out all the nonrelational tables. Unfortunately, automatically distinguishing a relational table from a nonrelational table can be difficult. To do so, the system uses a combination of handwritten and statistically trained classifiers that use topic-independent features of each table; for example, high-quality data tables often have relatively few empty cells. Another useful feature is whether each column contains a uniform data type (such as all dates or all integers). Google Research has found that finding a column toward the left side of the table with values drawn from the same semantic type (such as country, species, and institution) is a valuable signal for identifying highquality relational tables.

The second step is to recover metadata for each table passing through the first filter. Metadata is information that describes the data in the database (such as number of columns, types, and names). In the case of the presidents, the metadata contains the column labels President, Party, and so on. For coffeehouses, it might contain Name, Speciality, and Roaster. Although metadata for a traditional relational database can be complex, the goal for WebTables metadata is modest—determine whether or not the first row of the ta-

a The second-largest collection we know is due to Wang and Hu,<sup>22</sup> who also tried to gather data from Web pages but with a relatively small and focused set of input pages. Other research on table extraction has not focused on large collections.<sup>10,12,23</sup> Our discussion here refers to the number of distinct databases, not the number of tuples. Limaye et al<sup>16</sup> described techniques for mapping entities and columns in tables to an ontology.

ble includes labels for each column. When inspecting tables by hand, we found 70% of good relational-style tables contain such a metadata row. As with relational filtering, we used a set of trained classifiers to automatically determine whether or not the schema row is present.

The two techniques together allowed WebTables to recover 125 million high-quality databases from a large general Web crawl (several billion Web pages). The tables in this corpus contained more than 2.6 million unique "schemas," or unique sets of attribute strings. This enormous data set is a unique resource we explore in the following paragraphs.

Leveraging extracted data. Aggregating data over the extracted WebTables data, we can create new applications previously difficult or impossible through other techniques. One such application is structured data search. Traditional search engines are tuned to return relevant documents, not data sets, so users searching for data are generally ill-served. Using the extracted WebTables data, we implemented a search engine that takes a keyword query and returns a ranked list of databases instead of URLs; Figure 2 is a screenshot of the prototype system. Because WebTables extracted structural information for each object in the search engine's index, the results page can be more interesting than in a standard search engine. Here, the page of search results contains an automatically drawn map reflecting the cities listed in the data set; imagine the system being used by knowledge workers who want to find data to add to a spreadsheet.

In addition to the data in the tables, we found significant value in the collection of the tabular schemata we collected. We created the Attribute Correlation Statistics Database (ACSDb) consisting of simple frequency counts for each unique piece of metadata WebTables extracts; for example, the database of presidents mentioned earlier adds a single count to the four-element set president, party, term-as-president, vicepresident. By summing individual attribute counts over all entries in the ACSDb, WebTables is able to compute various attribute probabilities, given a An important lesson we learned is there is significant value in analyzing collections of metadata on the Web, in addition to the data itself. randomly chosen database; for example, the probability of seeing the name attribute is far higher than seeing the roaster attribute.

WebTables also computes conditional probabilities, so, for example, we learn that p(roaster |house-blend) is much higher than p(roaster | album-title). It makes sense that two coffee-related attributes occur together much more often than a combination of a coffeerelated attribute and, say, a musicrelated attribute. Using these probabilities in different ways, we can build interesting new applications, including these two:

*Schema autocomplete*. The database schema auto-complete application is designed to assist novice database designers. Like the tab-complete feature in word processors, schema autocomplete takes a few sample attributes from the user and suggests additional attributes to complete the table; for example, if a user types roaster and house-blend, the auto-complete feature might suggest speciality, opening-time and other attributes to complete the coffeehouse schema. Table 1 lists example outputs from our auto-complete tool, which is also useful in scenarios where users should be encouraged to reuse existing terminologies in their schemas.

The auto-complete algorithm is easily implemented with probabilities from the ACSDb. The algorithm repeatedly emits the attribute from the ACSDb to yield the highest probability, when conditioned on the attributes the user (or algorithm) has already suggested. The algorithm terminates when the attribute yielding the highest probability is below a tunable threshold.

Synonym finding. The WebTables synonym-finding application uses AC-SDb probabilities to automatically detect likely attribute synonyms; for example, phone-number and phone-# are two attribute labels that are semantically equivalent. Synonyms play a key role in data integration. When we merge two databases on the same topic created by different people, we first need to reconcile the different attribute names used in the two databases. Finding these synonyms is generally done by the application designer or drawn automatically from a pre-compiled linguistic resource (such as a thesaurus). However, the task of synonym finding is complicated by the fact that attribute names are often acronyms or word combinations, and their meanings are highly contextual. Unfortunately, manually computing a set of synonyms is burdensome and error-prone.

WebTables uses probabilities from the ACSDb to encode three observations about good synonyms:

► Two synonyms should not appear together in any known schema, as it would be repetitive on the part of the database designer;

► Two synonyms should share common co-attributes; for example, phone-number and phone-# should both appear along with name and address; and

► The most accurate synonyms are popular in real-world use cases.

WebTables can encode each of these observations in terms of attribute probabilities using ACSDb data. Combining them, we obtain a formula for a synonym-quality score WebTables uses to sort and rank every possible attribute pair; Table 2 lists a series of input domains and the output pairs of the synonym-finding system.

## **Deep Web Databases**

Not all structured data on the Web is published in easily accessible HTML tables. Large volumes of data stored in back-end databases are often made available to Web users only through HTML form interfaces; for example, a large chain of coffeehouses might have a database of store locations that are retrieved by zip code using the HTML form on the company's Web site, and users retrieve data by performing valid form submissions. On the back-end, HTML forms are processed by either posing structured queries over relational databases or sending keyword queries over text databases. The retrieved content is published on Web pages in structured templates, often including HTML tables.

While WebTables-harvested tables are potentially reachable by users posing keyword queries on search engines, the content behind HTML forms was for a long time believed to be beyond the reach of search engines; few hyperlinks point to Web pages resulting from form submissions, and Web crawlers did not have the ability to automatically fill out forms. Hence, the names "Deep," "Hidden," and "Invisible Web" have all been used to refer to the content accessible only through forms. Bergman<sup>2</sup> and He et al<sup>14</sup> have speculated that the data in the Deep Web far exceeds the data indexed by contemporary search engines. We estimate at least 10 million potentially useful distinct forms<sup>18</sup>; our previous work<sup>17</sup> has a more thorough discussion of the Deep Web literature and its relation to the projects described here.

The goal of Google's Deep Web Crawl Project is to make Deep Web content accessible to search-engine users. There are two complementary approaches to offering access to it: create vertical search engines for specific topics (such as coffee, presidents, cars, books, and real estate) and surface Deep Web content. In the first, for each vertical, a designer must create a mediated schema visible to users and create semantic mappings from the Web sources to the mediated schema. However, at Web scale, this approach suffers from several drawbacks:

► A human must spend time and effort building and maintaining each mapping;

► When dealing with thousands of domains, identifying the topic relevant to an arbitrary keyword query is extremely difficult; and

► Data on the Web reflects every topic in existence, and topic boundaries are not always clear.

The Deep Web Crawl project followed the second approach to surface DeepWeb content, pre-computing the most relevant form submissions for all interesting HTML forms. The URLs resulting from these submissions can then be added to the crawl of a search engine and indexed like any other HTML page. This approach leverages the existing search-engine infrastructure, allowing the seamless inclusion of Deep Web pages into Web-search results. The system currently surfaces content for several million Deep Web databases spanning more than 50 languages and several hundred domains, and the surfaced pages contribute results to more than 1,000 Web-search queries per second on Google.com. For example, as of the writing of this article, a search query for citibank atm 94043 will return in the first position a parameterized URL surfacing

Table 1. Sample output from the schema autocomplete tool. To the left is a user's input attribute; to the right are sample schemas.

Input attribute	Auto-completer output
name	name, size, last-modified, type
instructor	instructor, time, title, days, room, course
elected	elected, party, district, incumbent, status, opponent, description
ab	ab, h, r, bb, so, rbi, avg, lob, hr, pos, batters
sqft	sqft, price, baths, beds, year, type, lot-sqft, days-on-market, stories

Table 2. Sample output from the synonym-finding tool. To the left are the input context attributes; to the right are synonymous pairs generated by the system.

Input context	Synonym-finder outputs
name	e-maillemail, phoneItelephone, e-mail addressIemail address, dateIlast-modified
instructor	course-title title, day days, course course-#, course-name course-title
elected	candidate name, presiding-officer speaker
ab	k so, h hits, avg ba, name player
sqft	bath baths, list list-price, bed beds, price rent

results from a database of ATM locations—a very useful search result that would not have appeared otherwise.

Pre-computing the set of relevant form submissions for any given form is the primary difficulty with surfacing; for example, a field with label roaster should not be filled in with value toyota. Given the scale of a Deep Web crawl, it is crucial there be no human involvement in the process of pre-computing form submissions. Hence, previous work that either addressed the problem by constructing mediator systems one domain at a time<sup>8,9,21</sup> or needed site-specific wrappers or extractors to extract documents from text databases<sup>1,19</sup> could not be applied.

Surfacing Deep Web content involves two main technical challenges:

► Values must be selected for each input in the form; value selection is trivial for select menus but very challenging for text boxes; and

► Forms have multiple inputs, and using a simple strategy of enumerating all possible form submissions can be wasteful; for example, the search form on cars.com has five inputs, and a cross product will yield more than 200 million URLs, even though cars. com lists only 650,000 cars for sale.<sup>7</sup>

The full details on how we addressed these challenges are in Madhavan et al.<sup>18</sup> Here, we outline how we approach the two problems:

Selecting input values. A large number of forms have text-box inputs and require valid input values for the retrieval of any data. The system must therefore choose a good set of values to submit in order to surface useful result pages. Interestingly, we found it is not necessary to have a complete understanding of the semantics of the form to determine good candidate text inputs. To understand why, first note that text inputs fall into one of two categories: generic search inputs that accept most keywords and typed text inputs that accept only values in a particular topic area.

For search boxes, the system predicts an initial set of candidate keywords by analyzing text from the form site, using the text to bootstrap an iterative probing process. The system submits the form with candidate keywords; when valid form submissions result, the system extracts more keywords from the resulting pages. This iterative process continues until either there are no new candidate keywords or the system reaches a prespecified target number of results. The set of all candidate keywords can then be pruned, choosing a small number that ensures diversity of the exposed database content. Similar iterative probing approaches have been used to extract text documents from specific databases.<sup>1,6,15,19,20</sup>

For typed text boxes, the system attempts to match the type of the text box against a library of types common across topics (such as U.S. zip codes). Note that probing with values of the wrong type results in invalid submissions or pages with no results. We found even a library of just a few types can cover a significant number of text boxes.

Selecting input combinations. For HTML forms with more than one input, a simple strategy of enumerating the entire cross-product of all possible values for each input will result in a huge number of output URLs. Crawling too many URLs drains the resources of a search engine Web crawler while posing an unreasonable load on Web servers hosting the HTML forms. Choosing a subset of the cross-product that yields results that are nonempty, useful, and distinct is an algorithmic challenge.18 The system incrementally traverses the search space of all possible subsets of inputs. For a given subset, it tests whether it is informative, or capable of generating URLs with sufficient diversity in their content. As we showed in Madhavan et al,<sup>18</sup> only a small fraction of possible input sets must be tested, and, for each subset, the content of only a sample of generated URLs must be examined. Our algorithm is able to extract large fractions of underlying Deep Web databases without human supervision, using only a small number of form submissions. Furthermore, the number of form submissions the system generates is proportional to the size of the database underlying the form site, rather than the number of inputs and input combinations in the form.

Limitations of surfacing. By creating Web pages, surfacing does not preserve the structure or semantics of the data gathered from the underlying DeepWeb databases. But the loss in semantics is also a lost opportunity for query answering; for example, suppose a user searched for "used ford focus 1993" and a surfaced used-car listing page included Honda Civics, with a 1993 Honda Civic for sale, but also said "has better mileage than the Ford Focus." A traditional search engine would consider such a surfaced Web page a good result, despite not being helpful to the user. We could avoid this situation if the surfaced page had a search-engine-specific annotation that the page was for used-car listings of Honda Civics. One challenge for an automated system is to create a set of structure-aware annotations textual search engines can use effectively.

## **Next Steps**

These two projects represent first steps in retrieving structured data on the Web and making it directly accessible to users. Searching it is not a solved problem; in particular, search over large collections of data is still an area in need of significant research, as well as integration with other Web search. An important lesson we learned is there is significant value in analyzing collections of metadata on the Web, in addition to the data itself.

Specifically, from the collections we have worked with—forms and HTML tables—we have extracted several artifacts:

► A collection of forms (input names that appear together and values for select menus associated with input names);

• A collection of several million schemata for tables, or sets of column names appearing together; and

► A collection of columns, each with values in the same domain (such as city names, zip codes, and car makes).

**Semantic services.** Generalizing from our synonym finder and schema auto-complete, we build from the schema artifacts a set of semantic services that form a useful infrastructure for many other tasks. An example of such a service is that, given a name of an attribute, return a set of values for its column; such a service can automatically fill out forms in order to surface Deep Web content. A second

example is, given an entity, return a set of possible properties—attributes and relationships—that may be associated with it. Such a service would be useful for both information-extraction tasks and query expansion.

**Structured data from other sources.** Some of the principles of our previous projects are useful for extracting structured data from other growing sources on the Web:

Socially created data sets. These data sets (such as encyclopedia articles, videos, and photographs) are large and interesting and exist mainly in site-specific silos, so integrating them with information extracted from the wider Web would be useful;

*Hypertext-based data models.* These models, in which page authors use combinations of HTML elements (such as a list of hyperlinks), perform certain data-model tasks (such as indicate that all entities pointed to by the hyperlinks belong to the same set); this category can be considered a generalization of the observation that HTML tables are used to communicate relations; and

*Office-style documents.* These documents (such as spreadsheets and slide presentations) contain their own structured data, but because they are complicated, extracting information from them can be difficult, though it also means they are a tantalizing target.

Creating and publishing structured data. The projects we've described are reactive in the sense that they try to leverage data already on the Web. In a complementary line of work, we created Google Fusion Tables,<sup>13</sup> a service that aims to facilitate the creation, management, and publication of structured data, enabling users to upload tabular data files, including spreadsheets and CSV, of up to 100MB. The system provides ways to visualize the data-maps, charts, timelines-along with the ability to query by filtering and aggregating the data. Fusion Tables enables users to integrate data from multiple sources by performing joins across tables that may belong to different users. Users can keep the data private, share it with a select set of collaborators, or make it public. When made public, search engines are able to crawl the tables, thereby providing additional incentive to publish data. Fusion Tables also includes a set of social features (such as collaborators conducting detailed discussions of the data at the level of individual rows, columns, and cells). For notable uses of Fusion Tables go to https://sites.google.com/ site/fusiontablestalks/stories.

## Conclusion

Structured data on the Web involves several technical challenges: difficult to extract, typically disorganized, and often messy. The centralized control enforced by a traditional database system avoids all of them, but centralized control also misses out on the main virtues of Web data—that it can be created by anyone and covers every topic imaginable. We are only starting to see the benefits that might accrue from these virtues. In particular, as illustrated by WebTables synonym finding and schema auto-suggest, we see the results of large-scale data mining of an extracted (and otherwise unobtainable) data set.

It is often argued that only select Web-search companies are able to carry out research of the flavor we've described here. This argument holds mostly for research projects involving access to logs of search queries, but the research described here was made easier by having access to a large Web index and computational infrastructure, and much of it can be conducted at academic institutions as well, in particular when it involves such challenges as extracting the meaning of tables on the Web and finding interesting combinations of such tables. ACSDb is freely available to researchers outside of Google (https://www. eecs.umich.edu/ michjc/acsdb.html); we also expect to make additional data sets available to foster related research.

#### References

- Barbosa, L. and Freire, J. Siphoning Hidden-Web data through keyword-based interfaces. In *Proceedings* of the Brazilian Symposium on Databases, 2004, 309–321.
- 2. Bergman. M.K. The Deep Web: Surfacing hidden value. Journal of Electronic Publishing 7, 1 (2001).
- Cafarella, M.J., Halevy, A.Y., and Khoussainova, N. Data integration for the relational Web. *Proceedings of the* VLDB Endowment 2, 1 (2009), 1090–1101.
- Cafarella, M.J., Halevy, A.Y., Wang, D.Z., Wu, E., and Zhang, Y. WebTables: Exploring the power of tables on the Web. *Proceedings of the VLDB Endowment* 1, 1

(Aug. 2008), 538-549.

- Cafarella, M.J., Halevy, A.Y., Zhang, Y., Wang, D.Z., and Wu, E. Uncovering the relational Web. In *Proceedings* of the 11th International Workshop on the Web and Databases (Vancouver, B.C., June 13, 2008).
- Callan, J.P. and Connell, M.E. Query-based sampling of text databases. ACM Transactions on Information Systems 19, 2 (2001), 97–130.
- 7. Cars.com (faq); http://siy.cars.com/siy/qsg/
- faqgeneralinfo.jsp#howmanyads 8. Cazoodle apartment search; http://apartments. cazoodle.com/
- Chang, K.C.-C., He, B., and Zhang, Z. Toward large-scale integration: Building a metaquerier over databases on the Web. In Proceedings of the Conference on Innovative Data Systems Research (Asilomar, CA, Jan. 2005).
- Chen, H., Tsai, S., and Tsai, J. Mining tables from large-scale html texts. In Proceedings of the 18th International Conference on Computational Linguistics (Saarbrucken, Germany, July 31–Aug. 4, 2000), 166–172.
- 11. Elmeleegy, H., Madhavan, J., and Halevy, A. Harvesting relational tables from lists on the Web. *Proceedings of the VLDB Endowment 2*, 1 (2009), 1078–1089.
- Gatterbauer, W., Bohunsky, P., Herzog, M., Krüupl, B., and Pollak, B. Towards domain-independent information extraction from Web tables. In Proceedings of the 16th International World Wide Web Conference (Banff, Canada, May 8–12, 2007), 71–80.
- Gonzalez, H., Halevy, A., Jensen, C., Langen, A., Madhavan, J., Shapley, R., Shen, W., and Goldberg-Kidon, J. Google Fusion Tables: Web-centered data management and collaboration. In Proceedings of the SIGMOD ACM Special Interest Group on Management of Data (Indianapolis, 2010). ACM Press, New York, 2010, 1061–1066.
- He, B., Patel, M., Zhang, Z., and Chang, K.C.-C. Accessing the Deep Web. *Commun. ACM 50*, 5 (May 2007), 94–101.
- Ipeirotis, P.G. and Gravano, L. Distributed search over the Hidden Web: Hierarchical database sampling and selection. In Proceedings of the 28th International Conference on Very Large Databases (Hong Kong, Aug. 20–23, 2002), 394–405.
- Limaye, G., Sarawagi, S., and Chakrabarti, S. Annotating and searching Web tables using entities, types, and relationships. *Proceedings of the VLDB Endowment* 3, 1 (2010), 1338–1347.
- Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., and Halevy, A.Y. Google's Deep Web Crawl. Proceedings of the VLDB Endowment 1, 1 (2008), 1241–1252.
- Madhavan, J., Cohen, S., Dong, X.L., Halevy, A.Y., Jeffery, S.R., Ko, D., and Yu, C. Web-scale data integration: You can afford to pay as you go. In *Proceedings of the Second Conference on Innovative Data Systems Research* (Asilomar, CA, Jan. 7–10, 2007). 342–350.
- Ntoulas, A., Zerfos, P., and Cho, J. Downloading textual Hidden Web content through keyword queries. In Proceedings of the Joint Conference on Digital Libraries (Denver, June 7–11, 2005), 100–109.
- Raghavan, S. and Garcia-Molina, H. Crawling the Hidden Web. In Proceedings of the 27th International Conference on Very Large Databases (Rome, Italy, Sept. 11–14, 2001), 129–138.
- 21. Trulia; http://www.trulia.com/
- 22. Wang, Y. and Hu, J. A machine-learning-based approach for table detection on the Web. In *Proceedings of the 11th International World Wide Web Conference* (Honolulu, 2002), 242–250.
- Zanibbi, R., Blostein, D., and Cordy, J. A survey of table recognition: Models, observations, transformations, and inferences. *International Journal on Document Analysis and Recognition* 7, 1 (2004), 1–16.

Michael J. Cafarella (michjc@umich.edu) is an assistant professor of computer science and engineering at the University of Michigan, Ann Arbor, MI.

Alon Halevy (halevy@google.com) is Head of the Structured Data Management Research Group, Google Research, Mountain View, CA.

Jayant Madhavan (jayant@google.com) a senior software engineer at Google Research, Mountain View, CA.

© 2011 ACM 0001-0782/11/0200 \$10.00

Copyright of Communications of the ACM is the property of Association for Computing Machinery and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.