A Computational Alembic for a Web of Knowledge[⋆]

David W. Embley[†], Stephen W. Liddle[‡],
Deryle Lonsdale[**], George Nagy[††], Yuri Tijerino[‡‡]
Yihong Ding[†], Stephen Lynn[†], Jeff Peters[†], and Cui Tao[†]

[†]Department of Computer Science
[‡]Department of Information Systems
[**]Department of Linguistics and English Language
Brigham Young University, Provo, Utah, 84602

[††]Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, New York, 12180

[‡‡]Department of Applied Informatics
Kwansei Gakuin University, Kobe-Sanda, Japan

The current web is a web of linked pages. Frustrated users search for facts by guessing which keywords or keyword phrases might lead them to pages where they can find facts. Can we make it possible for users to search directly for facts? Can we turn the web into a web of facts (in addition to a web of pages containing facts)? Ultimately, can the web also be a knowledgebase—a *WoK* (*Web of Knowledge*)—that can provide direct answers to factual questions?

The answer to these questions calls for distilling knowledge from the web's wealth of heterogeneous digital data. But how? Our computational alembic[1] must turn raw symbols contained in web pages into knowledge and make this knowledge easily accessible via the web. We face three challenges: (1) automatic (or near automatic) creation of ontologies, (2) automatic (or near automatic) annotation of web pages with respect to these ontologies, and (3) simple, but accurate, query specification, usable without specialized training.

To show that a computational alembic with these characteristics is feasible, we are constructing a demo. We begin with an ontology editor [ISTA'05][2] with which we can manually construct extraction ontologies—conceptual models with instance recognizers that can automatically recognize and extract instances embedded in web pages for the concepts in the ontologies. From a populated extraction ontology it is straightforward to generate a database instance, which immediately allows it to be queried with a formal query language [DKE'99].[3] Added to this foundation, we have begun to build tools (1) to automate the generation of ontologies from semi-structured web pages, (2) to automate the

[⋆] Supported in part by the National Science Foundation (#0414644 and #0414854).

[1] A vessel with a beaked cap or head, formerly used in distilling liquids.

[2] All citations refer to www.deg.byu.edu/papers/ where our papers are posted.

[3] Our *High-Level Demo* lets a user apply a pre-built extraction ontology to a web page and then query the extracted information with SQL. (All demos to which we refer reside on www.deg.byu.edu/demos/.)

extraction of data instances with respect to these generated ontologies, and (3) to provide tools for free-form query evaluation. We have made significant progress on all three tool sets:

1. *Ontology generation.* Our TANGO[4] project [WWWJ'05] is well underway. TANGO generates ontologies: (1) it interprets tables by finding table labels and associating them with the table's data, (2) it conceptualizes interpreted tables turning them into conceptual models, and (3) it merges conceptualized tables into a growing ontology that represents a domain described by a set of domain-related tables. To interpret tables, we are building a tool to process arbitrary tables [Jha'07]. For a particular situation—namely, where so-called sibling tables are available, such as those displayed as a result of querying the hidden web—we have completed a table interpretation tool—TISP[5] [ER'07]. We have also completed a conceptualization tool—MOGO[6] [Lynn'08]. For merging ontologies, we have completed a matching and merging framework and an API interface [Lian'08], but we have not yet added into this framework our tools for automated ontology and data integration [InfoSys'06].
2. *Information extraction.* We are continuing to work on automatic semantic annotation [ASWC'06].[7] Currently, we are exploring ways to synergistically run extraction-ontology annotators and pattern-based annotators to reduce our reliance on hand-crafted extraction ontologies [SIGSEMIS'05]. We are also exploring ways to allow ontologies to be expressed as ordinary forms, which users already understand and can readily create and into which information can be harvested [CMLSA'07].
3. *Query specification.* We have completed two projects on free-form query specification: AskOntos [Vickers'06][8] and SerFR [ICDE'07].[9] We have not, however, integrated these projects into our overall WoK project and thus we currently can write queries only in SPARQL.

A current challenge for us is to integrate all these projects together into a single WoK demo. Although not yet finished,[10] we have completed much. Currently, when we begin our demo, we can start from scratch, find a set of sibling pages we wish to process, record their URLs in a simple text file, select

[4] Table ANalysis for Generating Ontologies

[5] Table Interpretation with Sibling-Pages

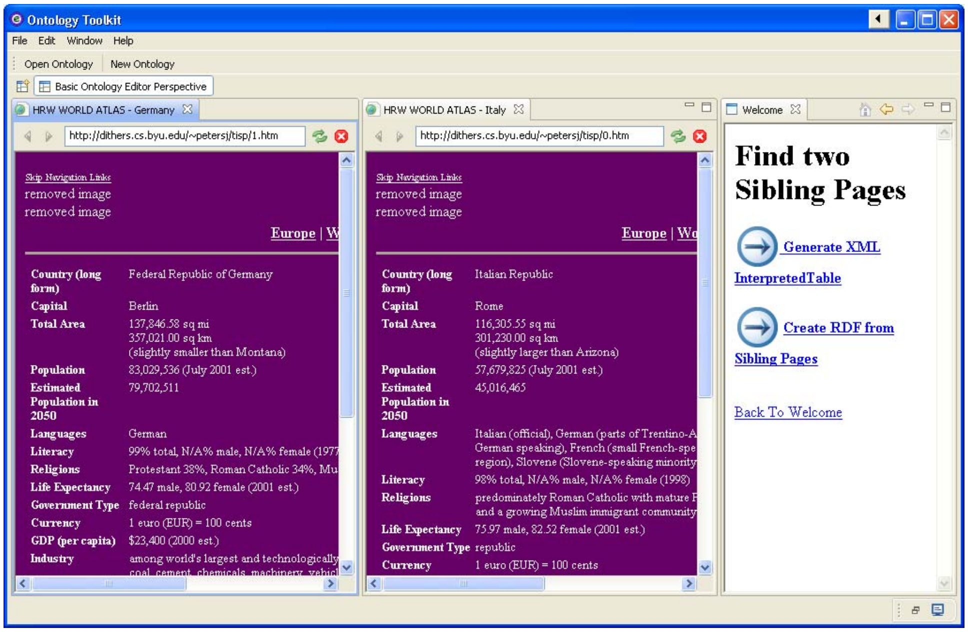
[6] Mini-Ontology GeneratOr—"mini" because it generates a small ontology from a single table to be integrated into the growing ontology

[7] Our *semantic annotation* demo yields annotated web pages—when a user's mouse hovers over an annotated item on a page, the meta-information from the extraction ontology by which the item was annotated appears.

[8] Our *AskOntos* demo allows a user to specify free-form queries for a car-sales domain. The demo shows answers to queries in a table. Each row of the table has a clickable "source" entry, which when clicked yields the original HTML page from which the information was extracted with the extracted information highlighted.

[9] Our *SerFR* demo runs, but is in need of an upgrade.

[10] The demo probably never will be finished because we are continuously testing new ideas for each of the three components of our computational alembic.



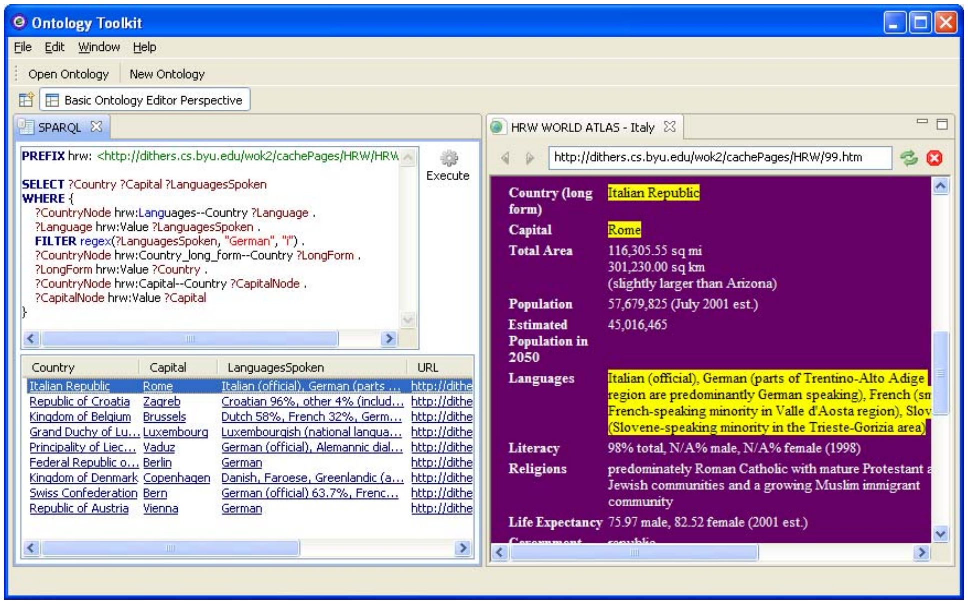
Country (long form)	Federal Republic of Germany
Capital	Berlin
Total Area	137,846.58 sq mi 357,021.00 sq km (slightly smaller than Montana)
Population	83,029,536 (July 2001 est.)
Estimated Population in 2050	79,702,511
Languages	German
Literacy	99% total, N/A% male, N/A% female (1977
Religions	Protestant 38%, Roman Catholic 34%, Mu
Life Expectancy	74.47 male, 80.92 female (2001 est.)
Government Type	federal republic
Currency	1 euro (EUR) = 100 cents
GDP (per capita)	$23,400 (2000 est.)
Industry	among world's largest and technologically coal, cement, chemicals, machinery, vehicl

Country (long form)	Italian Republic
Capital	Rome
Total Area	116,305.55 sq mi 301,230.00 sq km (slightly larger than Arizona)
Population	57,679,825 (July 2001 est.)
Estimated Population in 2050	45,016,465
Languages	Italian (official), German (parts of Trentino-A German speaking), French (small French-spe region), Slovene (Slovene-speaking minority
Literacy	98% total, N/A% male, N/A% female (1998)
Religions	predominately Roman Catholic with mature F and a growing Muslim immigrant community
Life Expectancy	75.97 male, 82.52 female (2001 est.)
Government Type	republic
Currency	1 euro (EUR) = 100 cents

Fig. 1. Sibling Pages Ready for TISP Processing.

any two as a starting place, and then click on a "Go" button to process the pages and prepare them to be queried.[11] The screen-shot of our demo in Figure 1 shows two sibling pages selected as the starting point. TISP takes these two pages and uses them to interpret the entire set of sibling pages (i.e., uses them to identify labels and values and to properly associate labels with values for all sibling pages in the file). Then, via some internal processing, the WoK system turns the interpreted table into an RDF data instance that not only contains all the data but also all the annotation information picked up by TISP. Figure 2 shows the interface from which a user can query the RDF data instance. When a user enters a SPARQL query, the WoK demo displays the result as clickable data instances. If a user clicks on a displayed data value, the WoK demo uses the additional stored information in the RDF data instance to find the annotation information. As Figure 2 shows, the WoK demo displays the web page from which the data instances were taken. If the user clicks on a data item, then prior to display, the WoK demo highlights the instance and scrolls the web page near to the spot in the page from which the instance was taken. In the special case for URL entries, which we always provide, the WoK demo highlights all the

[11] Typically, we look for a set of pages from a single site on the hidden web, but any set of pages with sibling tables works. We can, of course, also save the text file of URLs so that we can run the demo without having to look for a set of sibling pages.



```
PREFIX hrw: <http://dithers.cs.byu.edu/wok2/cachePages/HRW/HRW

SELECT ?Country ?Capital ?LanguagesSpoken
WHERE {
   ?CountryNode hrw:Languages--Country ?Language .
   ?Language hrw:Value ?LanguagesSpoken .
   FILTER regex(?LanguagesSpoken, "German", "i") .
   ?CountryNode hrw:Country_long_form--Country ?LongForm .
   ?LongForm hrw:Value ?Country .
   ?CountryNode hrw:Capital--Country ?CapitalNode .
   ?CapitalNode hrw:Value ?Capital
}
```

Country	Capital	LanguagesSpoken	URL
Italian Republic	Rome	Italian (official), German (parts ...	http://dithe
Republic of Croatia	Zagreb	Croatian 96%, other 4% (includ...	http://dithe
Kingdom of Belgium	Brussels	Dutch 58%, French 32%, Germ...	http://dithe
Grand Duchy of Lu...	Luxembourg	Luxembourgish (national langua...	http://dithe
Principality of Liec...	Vaduz	German (official), Alemannic dial...	http://dithe
Federal Republic o...	Berlin	German	http://dithe
Kingdom of Denmark	Copenhagen	Danish, Faroese, Greenlandic (a...	http://dithe
Swiss Confederation	Bern	German (official) 63.7%, Frenc...	http://dithe
Republic of Austria	Vienna	German	http://dithe

http://dithers.cs.byu.edu/wok2/cachePages/HRW/99.htm

Country (long form)	Italian Republic
Capital	Rome
Total Area	116,305.55 sq mi 301,230.00 sq km (slightly larger than Arizona)
Population	57,679,825 (July 2001 est.)
Estimated Population in 2050	45,016,465
Languages	Italian (official), German (parts of Trentino-Alto Adige region are predominantly German speaking), French (sm French-speaking minority in Valle d'Aosta region), Slov (Slovene-speaking minority in the Trieste-Gorizia area)
Literacy	98% total, N/A% male, N/A% female (1998)
Religions	predominately Roman Catholic with mature Protestant a Jewish communities and a growing Muslim immigrant community
Life Expectancy	75.97 male, 82.52 female (2001 est.)
Government	republic

Fig. 2. SPARQL Query with Results and Highlighted Web Page.

instances in the record for display. Figure 2 shows this case: here we have clicked on the URL entry for the Italian Republic.[12]

In summary, we point out the following: (1) The demo is in harmony with the aims of ER'08. Conceptual modeling plays a foundational role in actualizing the idea of a WoK (Web of Knowledge). (2) The demo will show (a) how to automatically build ontologies from certain types of semi-structured web pages, (b) how to automatically annotate data from web pages with respect to the generated ontology, and (c) how to query the extracted facts with a standard query language. (3) The demo is highly interactive. Users can run the demo by choosing web pages previously cached specifically for the demo or by browsing the web for pages they would like to try. Users can view intermediate results such as generated ontologies, generated annotation information, and generated RDF data files. And users can write their own SPARQL queries (and hopefully, their own free-form queries) and receive the answers to their queries as well as links back to the original pages from which the information was extracted.

[12] By summer's end we hope to replace the SPARQL query processor by a free-form query processor. We also hope to join together the various projects to provide additional ways to automatically generate ontologies and annotate web pages.