Integrating the World: The WorldInfo Assistant

Craig A. Knoblock[†‡], Jose Luis Ambite[†], Steven Minton[‡], Cyrus Shahabi[†], Mohammad Kolahdouzan[†], Maria Muslea[†], Jean Oh[†], and Snehal Thakkar[†]

[†]University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292

[‡]Fetch Technologies
4676 Admiralty Way, 10th floor
Marina del Rey, CA 90292

Abstract *The Internet is an extraordinary resource for information about countries throughout the world. There is detailed information about weather, geography, transportation, politics, news, etc. This data comes in a variety of forms including web pages, databases, maps, satellite imagery, newspapers, online radio and television, and so on. The problem, of course, is how to navigate through this abundance of information without getting bogged down in the details of the location, organization, and format of the information. Towards this end we have developed the* WorldInfo Assistant*, which extracts and integrates geographic-related data about countries throughout the world.*

Keywords: Information integration, wrappers, constraint reasoning, Heracles, WorldInfo Assistant

1 Introduction

The Internet provides an enormous amount of data about countries throughout the world, but the problem is how to locate, organize and present all of this information in a way that is natural and simple to use. To address this challenge, we have developed an application called the *WorldInfo Assistant*, which integrates all of this geographical and multimedia information about countries into a single easy-to-use framework. The user specifies the region of the world and the timeframe of interest, and the system provides the information relative to these parameters organized into a hierarchical representation that can be conveniently navigated.

There are two key underlying technologies that make the *WorldInfo Assistant* possible. First, we have developed a general integration framework, called *Heracles*, for building information assistants. The core of *Heracles* is an interactive constraint reasoning system that makes it easy for the user to navigate through the large amount of available data without being overwhelmed with information. Second, we have developed a set of tools for building wrappers that can turn semistructured data sources into structured sources. This allows the system to access and integrate data from online Web sources since much of the information is available in this form and is not available as databases.

2 The WorldInfo Assistant

The *WorldInfo Assistant* provides integrated access to a variety of sources about countries throughout the world. The initial system provides only a small fraction of the many sources available, but the system is very extensible and demonstrates the ability to integrate and organize a wide variety of sources. In this section we describe the key components of the system.

2.1 The Heracles Constraint Reasoning System

We have developed a hierarchical constraint reasoning system called Heracles [3], which provides the underlying reasoning support for

the *WorldInfo Assistant*. In general, information does not exist in isolation but it is related to or depends on other pieces of information. Heracles captures these relationships and dependencies in a constraint network. Each piece of information that the system accesses or computes is represented as a variable in a constraint network. The relationships among variables are specified as constraints. Conceptually, a constraint defines the valid combinations of values for a set of variables. A constraint is a computable component which may be implemented by a local table look-up, by the computation of a local function, by retrieving a set of tuples from a remote wrapper, or by calling an arbitrary external program.

The constraint reasoning system propagates the information entered by the user as well as the system's suggestions, and decides when to launch information requests, evaluate constraints, and compute preferences. All of these tasks run as asynchronous processes to give the user as much support as possible without interfering with his work.

In order to manage the complexity and capture the task structure of an application, closely related variables and constraints are encapsulated into templates. The templates are organized hierarchically so that a higher-level template representing an abstract task (e.g., Weather) may be decomposed into a set of more specific subtasks, called subtemplates (e.g., Current, Future, Statistical). This hierarchical task network structure further helps to manage the complexity of an application.

In the *WorldInfo Assistant* the top-level template has the major categories of information that the system can provide, which includes weather, news, maps, imagery, etc. At the top-level the user specifies the region, country, city, and the time frame of interest. Figure 1 shows the system after the user has specified Dakar, Senegal as the region of interest and March 2001 as the time frame, and the user has expanded the subtemplate that provides the detailed weather.

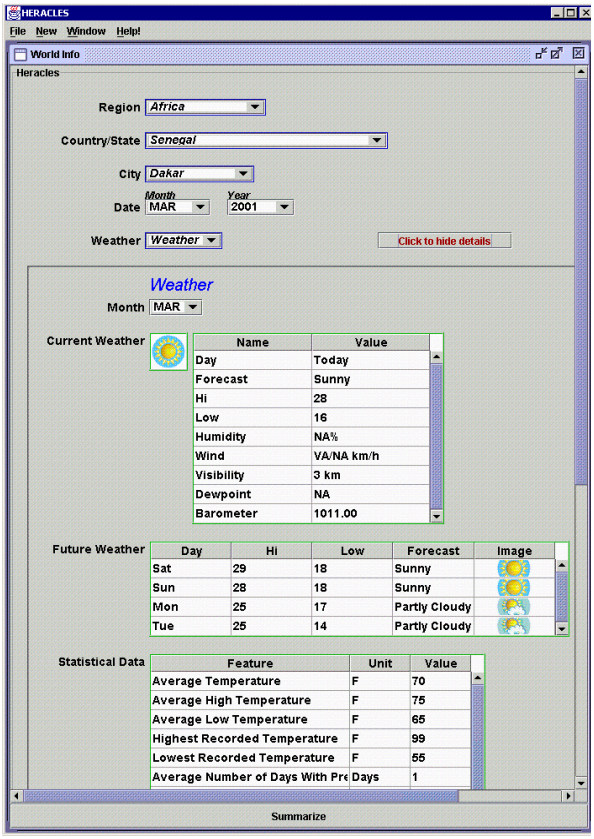

Figure 1: After the user specifies the region of interest and timeframe, the system provides the corresponding data on the weather

2.2 Converting Web Sources into Live Databases

Access to on-line data sources is a critical component of our information assistants. In the *WorldInfo Assistant* most of the information is accessed directly from Web sources, although the system also accesses local and remote databases and programs. To access web data we build wrappers that turn web sources into structured data sources. This allows the system to reason with the data and integrate the information with other data sources.

We have developed a set of tools for semi-automatically creating wrappers for web sources [8]. The tools allow a user to specify by example what the wrapper should extract from a source. The examples are then fed to an inductive learning system that generates a set of

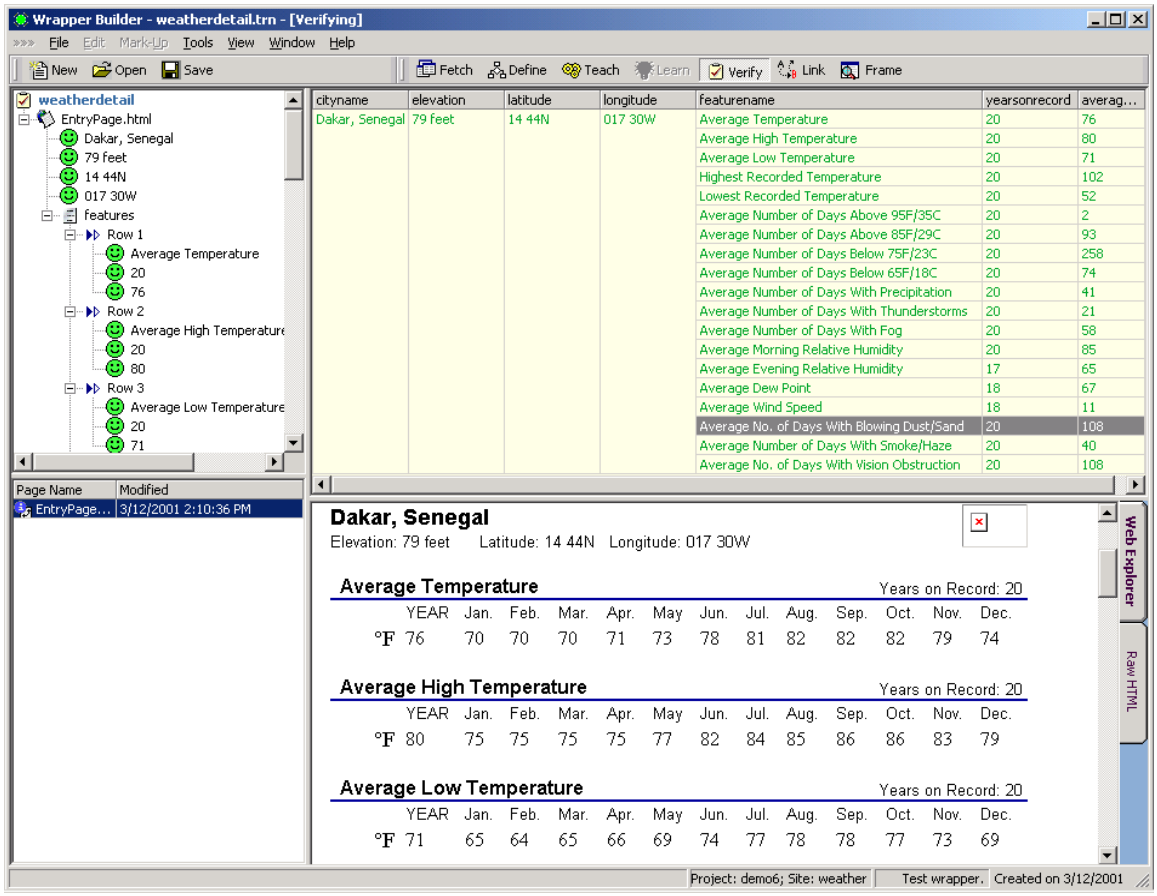

Figure 2: The Graphical User Interface for Building a Wrapper

rules for extracting the required data from a site. Beyond just creating the rules, we have also developed techniques for ensuring that the system is extracting the right data [7], monitoring the source to ensure the it continues to function properly [5], and automatically repairing wrappers in response to format changes in a site [2]. Once a wrapper for a site has been created, the system can use that site programmatically.

For example, we use a wrapper to extract historical weather data from the Weatherbase.com site. The wrapper builder tool is shown in Figure 2. The bottom right pane in Figure 2 shows the the original HTML page for Dakar, Senegal at weatherbase.com. The two upper panes shown the data that has been extracted from that page. With this wrapper the system can send a request to get the historical weather for a given city and time period. The wrapper will return the corresponding XML data. Note that the historical weather data shown in Figure 1 is for the month of March as was specified by the user.

2.3 Accessing Multimedia Information Sources

The *WorldInfo Assistant* integrates a variety of multimedia and geospatial information sources. The different types of sources that are integrated into the system include the following:

- **Databases** – This includes sources such as the database of information about all airports throughout the world published on-line by the National Imagery and Mapping Agency (NIMA).

- **Semistructured data sources** – We use wrappers to extract information from web sources such as weather conditions for any place in the world (from weather.yahoo.com and weatherbase.com), basic country information from the CIA World Factbook, holidays for any country in the world (from holidayfestival.com), etc.
- **Web pages** – Some Web sites are better left as Web sites, so we provide direct links to sources such as an on-line travel guide (www.travel-guide.com) as well as links to newspaper and magazine sites throughout the world (from www.newsdirectory.com).
- **Text sources** – We present the relevant news for a region by presenting the original pages of text from CNN.
- **Images** – We provide direct links to images of maps for cities and countries throughout the world, for example, from the on-line map collection of the Perry-Castañeda Library at the University of Texas, Austin (www.lib.utexas.edu).
- **Audio** – These include the on-line audio sources for radio stations throughout the world (from www.comfm.com/live/radio).
- **Video** – This includes on-line video for television stations throughout the world (from www.comfm.com/live/tv).

In addition to the different types of multimedia data, the *WorldInfo Assistant* also integrate a variety of the more specialized types of geospatial data. The geospatial data includes:

- **Imagery** – We have imagery with resolutions down to 1 meter from NIMA, USGS, and SpaceImaging for locations throughout the world.
- **Maps** – We also have maps at multiple scales from both NIMA and on-line map sources (such as the Perry-Castañeda Library).
- **Gazetteer** – We have a gazetteer with over 500 different feature types, such as airports, rail stations, hospitals, storage tanks, etc., on locations throughout the world.
- **Vector Data** – We have over 60 different layers of vector data including roads, coastlines, railroads, runways, waterways, etc.

Figure 3 shows the template for selecting maps for cities and countries throughout the world. The system shows only the available maps for Senegal which is the country of interest. Once the user selects one of these, the corresponding map is displayed in a Web browser since the remote site stores the maps in multiple formats (such as jpeg, gif, and pdf). In our example the user selects the map for Cap Vert (Figure 3), and the resulting image appears in a web browser spawned by the system (Figure 4).

2.4 Providing an Integrated View of the Data

One of the key features of the underlying constraint reasoning system is that it provides an integrated view of the different types of information. When the user changes a value in one of the slots in the interface, the system automatically reconsiders all the values for the dependent slots in order to insure that all values are consistent with the constraints. For example, when the user is looking at an image for a particular area, it also displays the corresponding map as well as the feature data from the gazetteer and relevant vector data. Figure 5 shows a satellite image and a map of the city of Dakar with a set of gazetteer points superimposed both the image and the map. The gazetteer points also appear as values of several slots in the interface (not shown in Figure 5). If the user wants to examine the airport more closely, then he could either select that point from the list of gazetteer points or click directly on the map. Figure 6 shows the result after the user clicks on the airport and requests the vector data for the runways of the airport.

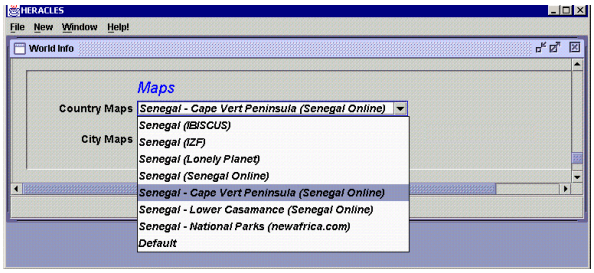

Figure 3: Links to on-line maps extracted by a wrapper. The system shows only the relevant links (maps of Senegal)

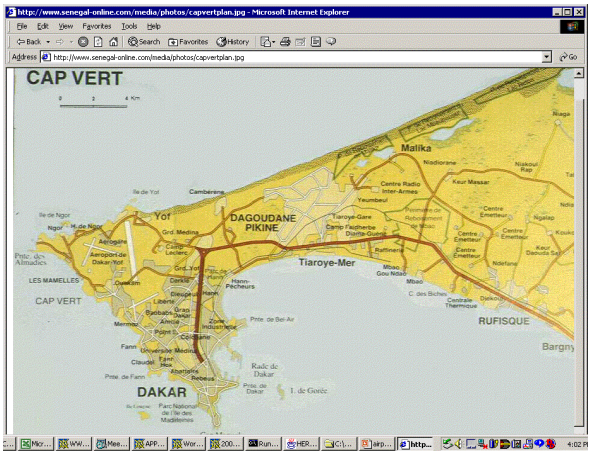

Figure 4: On-line Map of Cap Vert displayed in an external Web browser spawned by the system

Since the information is all linked using the constraint reasoner, once the user selects this airport, the system also retrieves the detailed data on the airport closest to the selected location. Figure 7 shows the airport data for the Leopold Sedar Senghor International Airport, which includes detailed information on location, codes, operating hours, communications, navigational aides, as well as details on the individual runways. If the user selected a different airport on the map, the system would automatically display the corresponding data for that airport.

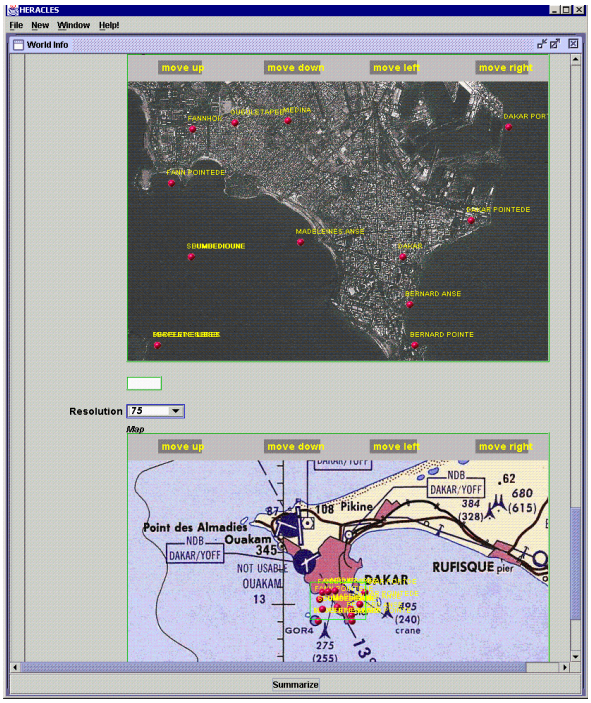

Figure 5: Satellite image of the port of Dakar with the corresponding map and features

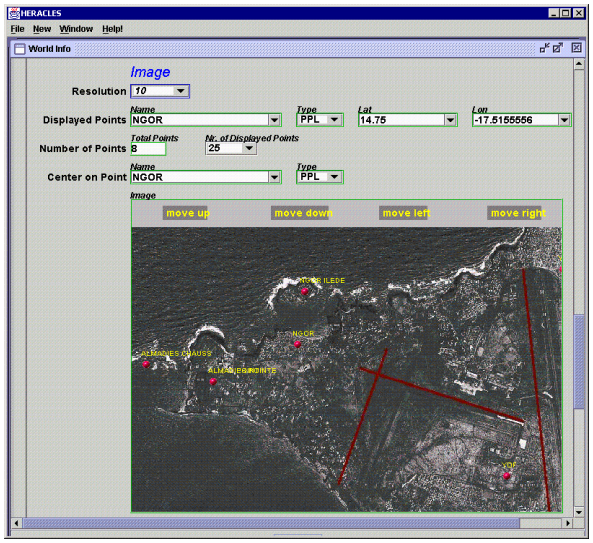

Figure 6: Satellite image of the selected airport with the vector data for the three runways

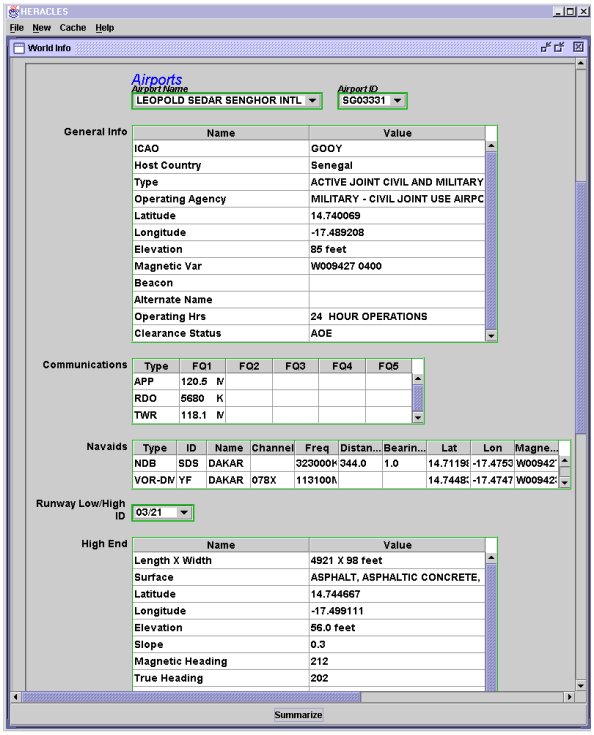

Figure 7: Detailed information on the Leopold Sedar Senghor International Airport

3 Related Work

There has been a lot of research in the area of constraint programming [9], but not much attention has been paid to the interplay between information gathering, constraint propagation, and user interaction, which is the focus of Heracles. Bressan and Goh [1] have applied constraint reasoning technology to information integration. However, the specific problem they are addressing is quite different from ours. They use constraint logic programming to find relevant sources and construct an evaluation plan for a user query. In our system the relevant sources have already been identified. We focus on user interactivity, flexible integration of information gathering and other computational constraints in an uniform framework, and on information propagation in service of the user tasks.

The growth and changes to the constraint network that occur in Heracles as a result of the hierarchical expansion of templates can be seen as a form of dynamic constraint satisfaction [6]. In dynamic constraint satisfaction the variables and constraints present in the network are allowed to change with time. Heracles imposes a structure to these changes as they correspond to meaningful units in the application: the templates.

Lamma et al. [4] propose a framework for interactive constraint satisfaction problems (ICSP) in which the acquisition of values for each variable is interleaved with the enforcement of constraints. The interactive behavior of our constraint reasoner also can be seen as a form of ICSP. However, our approach includes a notion of hierarchical decomposition and task orientation. Their application domain is on visual object recognition, while our focus is on information integration

4 Conclusion

We have presented an initial prototype of the *WorldInfo Assistant*. The system demonstrates an advance in three different areas. First, the system provides an integrated view of the huge amount of information available on countries throughout the world. Given the large amount of information available, the key contribution here is the ability to provide the user with the relevant data without overwhelming him with information. Our hierarchical constraint system facilitates the organization of the information, ensures that displayed data is always consistent, and offers a fast response to user interactions. Second, the system demonstrates the ability to access semi-structured web data by using wrappers. A wrapper provides a structured query interface to a web source and allow the system to extract precisely the information needed by the user. Third, the system provides access to the many different types of multimedia data that is available over the Internet.

In the future, we plan to extend this system in a number of directions. First, we are in the process in identifying and integrating a much larger set of sources to build a system

that goes beyond a simple demonstration and provides a useful operational system. Second, we are building a version of the system that will run within a web browser. The challenge in doing this is that the system updates the slots in real-time as the data becomes available and the standard web protocols used within a browser do not naturally support real-time updates. Third, we are developing the next generation of the constraint reasoning system, which will support a richer language of constraints and allow the user to incorporate their own sources and corresponding constraints. Finally, we are exploring the problems of efficiently integrating spatial data to support spatial queries.

Acknowledgements

The research reported here was supported in part by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-00-1-0504, in part by the Rome Laboratory of the Air Force Systems Command and DARPA under contract number F30602-98-2-0109, in part by the Air Force Office of Scientific Research under Grant Number F49620-01-1-0053, in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152, and in part by a research grant from NCR. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

References

[1] Stéphane Bressan and Cheng Hian Goh. Semantic integration of disparate information sources over the internet using constraint propagation. In *Workshop on Constraint Reasoning on the Internet, at the Third International Conference on Principles and Practice of Constraint Programming (CP97)*, Linz, Austria, 1997.

[2] Craig A. Knoblock, Kristina Lerman, Steven Minton, and Ion Muslea. Accurately and reliably extracting data from the web: A machine learning approach. *Data Engineering Bulletin*, 23(4), December 2000.

[3] Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Maria Muslea, Jean Oh, and Martin Frank. Mixed-initiative, multi-source information assistants. In *Proceedings of the World Wide Web Conference*, Hong Kong, May 2001.

[4] Evelina Lamma, Paola Mello, Michela Milano, Rita Cucchiara, Marco Gavanelli, and Massimo Piccardi. Constraint propagation and value acquisition: why we should do it interactively. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence*, pages 468–474, Stockholm, Sweden, 1999.

[5] Kristina Lerman and Steven Minton. Learning the common structure of data. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 609–614, 2000.

[6] Sanjay Mittal and Brian Falkenhainer. Dynamic constraint satisfaction problems. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 25–32, 1990.

[7] Ion Muslea, Steven Minton, and Craig A. Knoblock. Selective sampling with redundant views. In *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000.

[8] Ion Muslea, Steven Minton, and Craig A. Knoblock. Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, 4(1/2), March 2001.

[9] Vijay Saraswat and Pascal van Hentenryck, editors. *Principles and Practice of Constraint Programming.* MIT Press, Cambridge, MA, 1995.