Annales UMCS Informatica AI 1 (2003) 259-270

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

The world of travel: a comparative analysis of classification methods

A. Gilbert[a], M. Gordon[a], Marcin Paprzycki[a*], J. Wright[b]

[a]*Computer Science Department, Oklahoma State Uniwersity, Tulsa, OK 74106, USA,*
[b]*EDS, Tulsa, OK, USA*

Abstract

In this paper we look at the ways leading Internet content providers such as Yahoo! and Google categorize travel-related web resources, and contrast their methods to an alternate scheme underlying the business-communication specification of the Open Travel Alliance. Our analysis reveals some inherent weaknesses in each approach to resource categorization and a need for a more comprehensive solution. In formulating a new classification scheme for travel resources we attempt to take advantage of the strengths and improve on the weaknesses of existing systems.

1. Introduction

Hierarchical directory systems such as those employed by Yahoo! and Google are a relatively convenient and scalable means of organizing, navigating and searching tremendous volumes of information [1]. Google and Yahoo! are the successors to a long line of Internet search and directory sites, pre- and post-World Wide Web. Although these systems can be considered close to optimal for human consumers (an assumption that should be re-evaluated in light of observations made in this paper), they are distinctly lacking when it comes to machine accessibility. More precisely, the qualities that make directory systems such as Yahoo! popular with human users are the same which hinder their employment by machines: their human "usability" often equates to an absence of several prerequisites to machine consumption, namely: formally-defined semantics for navigating hierarchical relationships; business logic for operating on a directory structure; and programmatic interfaces for interacting and manipulating information stored in the system. Machines are reduced to operating on directories and search engines by the same means as human users, which places the machines at a distinct disadvantage. In order to allow

[*]Corresponding author: *e-mail address*: marcin@cs.okstate.edu

computers to work with directories of web resources, the directory systems must: (1) organize their resources to allow discovery via semantic-level lookups and navigation, and (2) describe the interactions required to "consume" these resources in a form and language machines can understand. Programmatic information of this type is necessary to facilitate automatic interactions between web-based suppliers and consumers [2].

A machine interface to directories becomes of particular importance when we consider recent advances in Internet information processing [3], most notably in the area of software agents [4]. It is often claimed that agent-based systems are the future processors of distributed heterogeneous information. Recent advances in multi-agent negotiations [5], auctions [6], commerce [7], and bargaining have brought the vision of a software agent-driven information economy one step closer. The ability of software agents (i.e. computers) to interface with web content will be crucial to the further development of online commerce. While the study of agent-based systems is currently spreading into many domains, our recent focus has been on the development of agent-based systems for supporting travelers [8-16]. It is within this context that we examine the role of directories and search engines such as Yahoo! and Google! as guides for agents discovering travel-related content on the Internet.

While there have been attempts to standardize and type electronic business communications [17-18], none have achieved the widespread support required for industry success. Electronic Data Interchange (EDI) [18] was one of the foremost of these efforts. Although many industries investigated the use of EDI techniques, and defined schemes to exchange common information between businesses within a single industry, the approach only caught hold in a few fields, notably the travel and health care industries. The lack of a united front for the EDI approach obviously contributed to its limited adoption. Evidence of the limitations of EDI within the travel industry lead to a number of recent industry-wide collaborations on data interchange specifications, among them the Open Travel Alliance (OTA) [19]. Over the past few years the OTA has developed XML-based specifications to support business-to-business communication within the travel industry. In the context of this paper the importance of this effort is twofold. First, the OTA specifications are clearly focused on facilitating machine-to-machine communication about the travel world. This being the case, the types of and relationships between information implicit in these specifications are substantially different than the above-described cataloging and searching services. Second, the specifications have been developed by a group of travel professionals, in contrast to the Yahoo! and Google directories, which were created and evolved through a process that was largely based on a “popular vote.”

In summary, we observe two distinct ways of dealing with travel-related content:

- The catalog approach exemplified by Yahoo! and Google, which is geared toward human consumption and is consequently rather poorly suited to automatic information processing
- The OTA specification of a machine-oriented communication protocol for facilitating business-to-business communication about travel-related activities. This communication, while clearly not intended for human consumption, also implicitly defines a categorization of the world of travel that can nevertheless be explicitly compared to the Yahoo! and Google systems.

The aim of this paper is to analyze the categorization of the world of travel as defined by both approaches. The ultimate end of our design analysis will be to utilize this information to define a more formal ontology of the world of travel [20-21]. In Section 2 we will present and analyze the way the travel world is represented by Yahoo!, Google and OTA, respectively. Our analysis will point out strengths and weaknesses of each categorization. Based on this analysis, we will propose a general framework that capitalizes on the strengths of each system, while avoiding some of their weaknesses (Section 3). While the aforementioned formal ontology of the world of travel is beyond the scope of this paper, we hope to provide some direction as to how such a categorization should be developed. Finally, we will summarize the technological means by which we are implementing a general framework for serving travel-related information to both human and machine consumers (Section 4).

2. Existing approaches

We begin our consideration by taking a closer look at two current and popular approaches to hierarchical categorization of the world of travel, Yahoo! and Google, and then proceed with a description of the central categories at the heart of the OTA specifications.

2.1. Yahoo! and Google

The Yahoo! and Google web directories are organized according to a human-defined classification scheme. The categories in the directories are simply parents in a tree. The sites at the leaves of the tree are minimally described: aside from the basic URI, they contain only a very short (few words) description; site descriptions do not include a formal specification of how to interact with the given site, beyond the protocol included in the URI (usually http). When travel is considered, the two web directories use what can be dubbed "resource type," and, to a lesser extent, the geographical location of a resource, to organize the world of travel into a hierarchical structure. Let us now look at each directory in detail.

2.1.1. The Yahoo! directory of travel

A partial depiction of the "world of travel" according to Yahoo! is presented in Figure 1. To obtain this picture we traversed portions of the directory starting from the top-level category *Sports and Recreation*, capturing the subsequent sub-categories in the branches of interest. Categories representing links to other branches (containing items outside of the context of our interest) have been removed, leaving only natural descendants in a branch. While this method is certainly open to criticism and obviously implies that the overall structure represented by the Yahoo! taxonomy is more complicated than it appears, we believe the depiction is sufficient to illustrate our main points.

Even a brief look at Figure 1 indicates that the directory scheme is somewhat aberrant: for instance, why would one consider *Canals*, *Commuting* and *Education* to be subcategories under transportation? Or, why are commercial airlines not a subcategory of *Travel*, when they obviously play such a significant role in travel? While humans are certainly able to make sense of and navigate this alphabetically ordered directory structure, its organization is clearly not intelligible to machines, which require a much less ambiguous semantics of traversal. As a side note, a more general question can be asked: is this really the best way of displaying information for human consumption? One could suggest that when subcategories are listed, they should be listed in a way that groups similar items together (for instance *Canoeing* and *Kayaking* are closer to *Whitewater Rafting* than to other events), however these considerations are outside of the scope of the current paper.

Notice also that the travel categories do not contain any geographical pointers. This is because in Yahoo! the geographical categorization is completely separated from the branch depicted above and starts with the top level *Regional* category.

2.1.2. The Google directory of travel

The second web directory service we consider is Google. The world of travel according to Google is presented in Figure 2. This figure was obtained in the same manner as Figure 1.

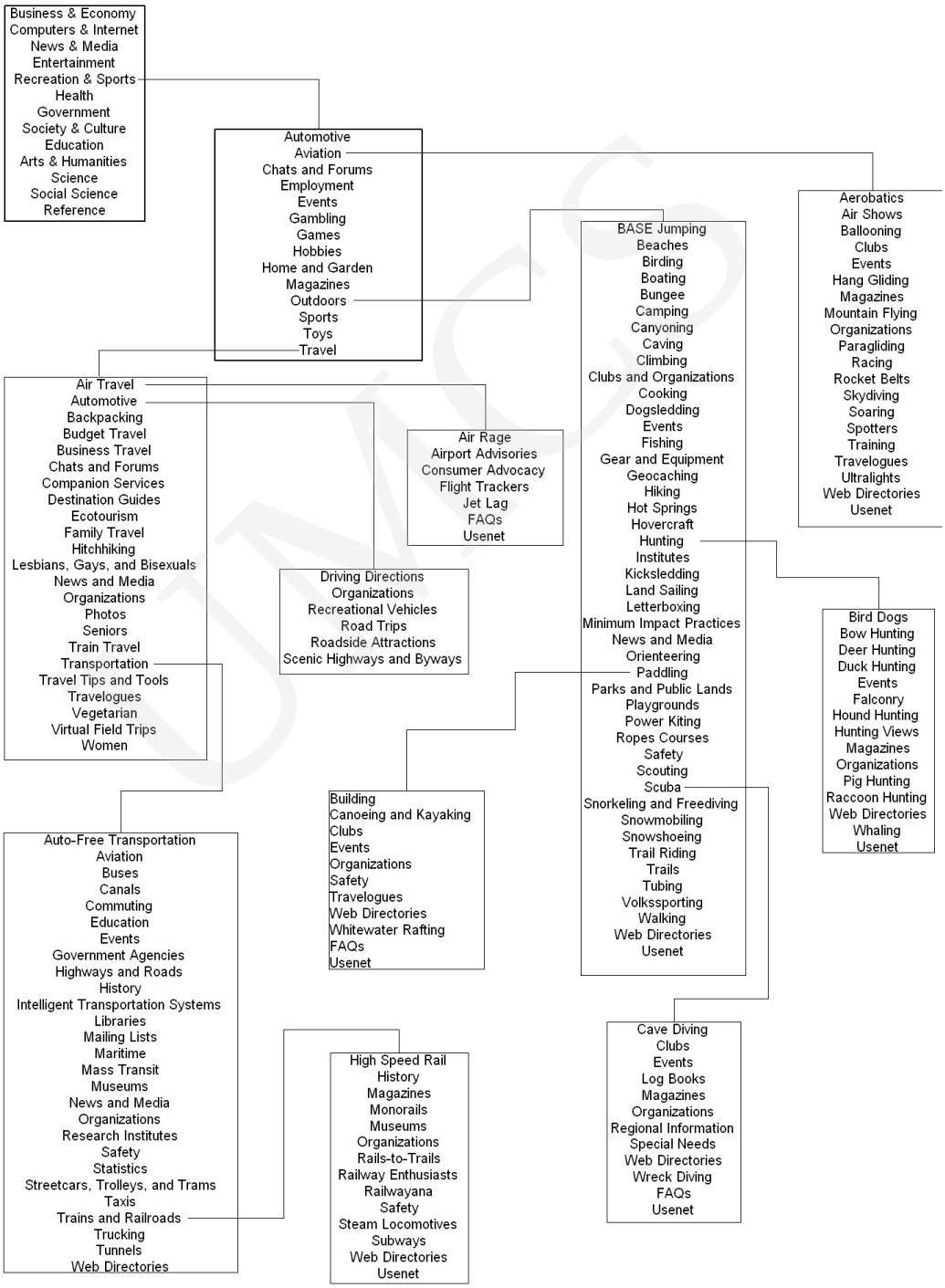

Fig. 1. Partial depiction of the world of travel according to Yahoo!

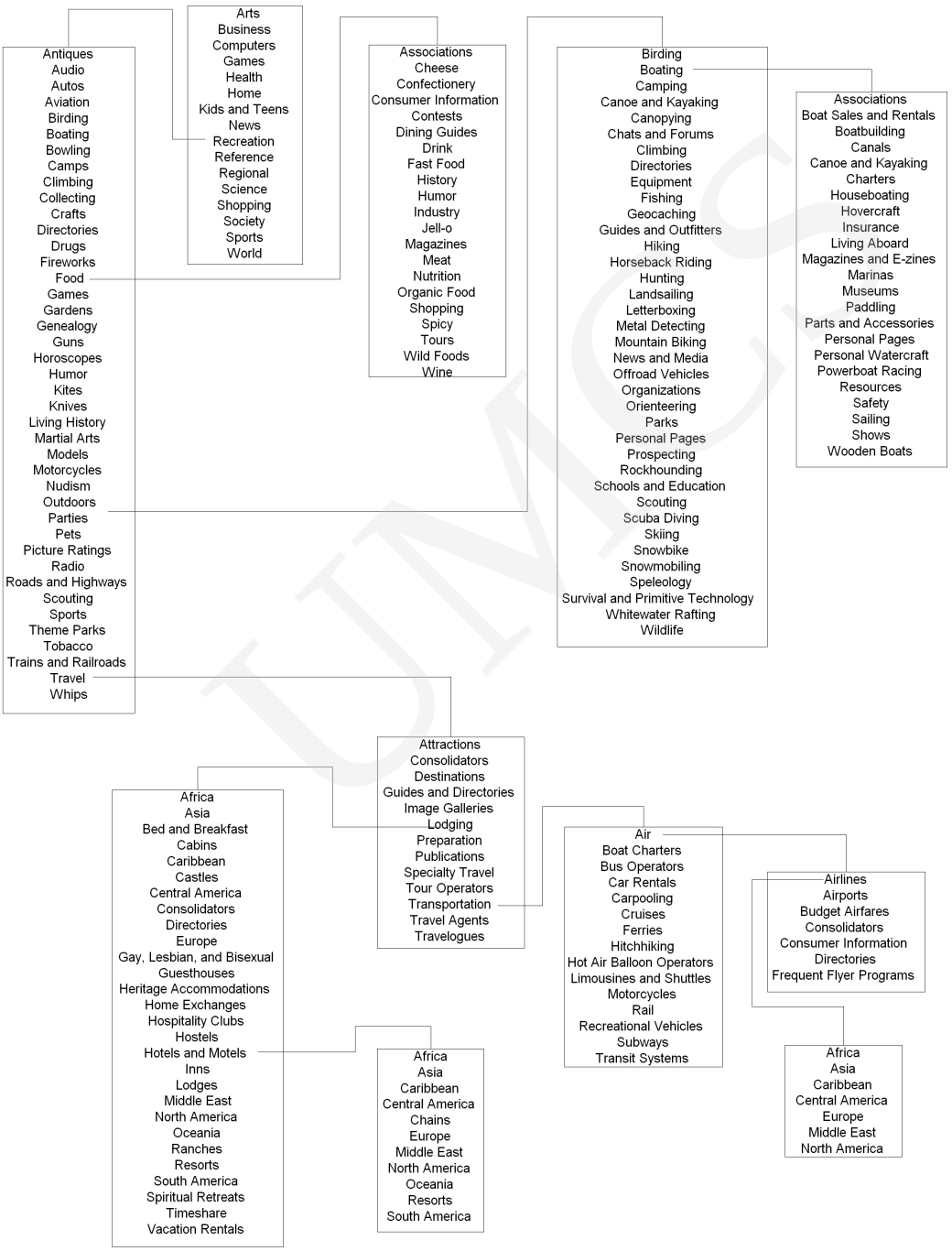

Fig. 2. Partial categorization of the world of travel according to Google

We observe that the Google “world of travel” suffers from problems similar to those of the Yahoo! directory: the categories represented “together” do not seem to have much in common (see for instance: *Attractions*, *Consolidators* and *Destinations*). What is interesting, however, is the fact that Google intermixes two important aspects of travel: services and locations. We find for instance: *Castles*, *Central America* and *Consolidators* as three subsequent subcategories. In this context, let us make two observations. First the question "Is really appropriate for the human consumers to have these two aspects of travel intermixed together?" (a detailed analysis of this point falls outside of the scope of this paper). Second, because it does mix category and location - two very conceptually different classification schemes - the Google directory is decidedly more difficult for machine consumers than Yahoo!

2.2. OTA

A very different approach to classifying and describing information about travel resources has been developed by the travel industry itself. The Open Travel Alliance (OTA) specifications are part of the travel industry's answer to the migration of travelers to Internet booking sites such as Travelersadvantage and Expedia. They also represent a movement in the industry to replace the proprietary EDI approach with an open, XML-based standard. The OTA specifications define a common language for members of the travel industry to exchange information about and execute transactions on travel resources - thus defining interactions between suppliers and consumers of travel services. The OTA specifications are based on ebXML [22], a set of standards for describing business processes and relationships in XML. The development of the Open Travel Alliance specifications is significant in the context of this paper in that it (1) clearly represents an attempt to facilitate machine-to-machine communication of travel information, and (2) is originated by the experts, and designed to be comprehensive. These specifications have the potential to vastly improve the ease of communication between cooperating travel industry partners over the Internet. The specifications are detailed to such an extent that any business conforming to the standard should easily be able to understand and communicate with other participating parties. The following is a list of travel resource types defined by the OTA:

Air Working Group (Air Travel)
- OTA_Air Flifo RQ/RS (Flight Information Request/Response)
- OTA_Air Schedules RQ/RS (Air Schedules Request/Response)
- OTA_Veh Loc Search RQ/RS

Hotel Working Group
- OTA_Hotel Rooming List RQ/RS
- OTA_Hotel Descriptive Info RQ/RS

OTA_Hotel RFP RS/RQ (Hotel Request for Proposal)
OTA_Hotel Reservation Modify
OTA_Pkg Avail RQ/RS
OTA_Pkg Book RQ/RS

Travel Itinerary Messages
OTA_Travel Itinerary RQ/RS

Rail Messages
OTA_Rail Avail RQ/RS
OTA_Rail Book RQ/RS
OTA_Rail Retrieve RS

Loyalty Messages
OTA_Loyalty Account Create RQ
OTA_Loyalty Account RS
OTA_Read RQ
OTA_Loyalty Account RS
OTA_Loyalty Certificate Create RQ/RS
OTA_Loyalty Certificate Create Notif RQ/RS
OTA_Loyalty Certificate Redepmtion RQ/RS

Generic Messages
OTA_Ping RQ/RS
OTA_Cancel RQ/RS
OTA_Delete RQ/RS
OTA_Update RQ/RS
OTA_Read RQ
OTA_Create Profile RQ/RS

Before we proceed, a methodological clarification is in order. The OTA specifications do not define an explicit categorization of resources. However, in describing the business processes and communication between players in the travel industry, the specifications do define a set of travel resource types, which implicitly form a classification scheme. For the purposes of our analysis we have extracted this implicit taxonomy from the set of specifications. Thus we observe that the world of travel according to the Open Travel Alliance currently consists of:

- Air travel
- Hotel/Accommodations
- Itineraries
- Railroad travel
- Loyalty-oriented services

Note that this taxonomy contains at least one category, *Loyalty-oriented services* that is missing from the categorizations suggested by Yahoo! and Google.

2.3. Further observations

One of the inherent problems with directories like Yahoo! and Google is that they organize sites on the Internet, which do not necessarily represent the real world. Unfortunately, travel services are one domain in which the connection between the Internet representation of a resource and the resource itself is especially important. The organization of travel information cannot be disconnected from the "organization" of the physical world (e.g. geographical, cultural, corporate).

This weakness in web directories is countered by one of the strengths of the Open Travel Alliance's approach, which is based on a one-to-one mapping of travel providers (hotel companies, airlines, etc.) to travel information. Unfortunately, this advantage also has a flaw, in that the set of travel providers (communicating peers) in the OTA network represent only a fraction of the travel resources available. In order to discover smaller resources (such as restaurants) the traveler must turn to the web directories, where the problem of information disconnected from resources must be dealt with once again.

3. Toward an integrated solution

Though the two approaches – directories for human users and B2B machine consumption for partners in the travel industry – each serve a specific purpose and have certain strengths and weaknesses, we believe there is a need for a compromise solution that takes advantage of the best aspects of both approaches while avoiding some of their weaknesses. In order to arrive at an integrated solution for classifying and organizing information about travel resources, however, we must first examine some of the assumptions made in the previous attempts: namely, what are we trying to organize? More precisely, what exactly is an Internet travel resource? Let us consider the following example:

Room 703 of the Marriot Dallas/Ft. Worth (DFW) Airport South, 4151 Centreport Drive, Fort Worth Texas 76155, that may be reserved for personal use on Wednesday, April 16, 2003 for $99.00 US via Expedia.com's (or some other comparable service's) implementation of the OTA Hotel reservation interface.

By applying a simple analysis to this description, we can derive some generalizations and develop a framework that defines all types of travel data. First, we know that this resource is a room. But it is not just any room. There is an entire hierarchy of classification that helps to describe it. For example, this room is a member of the class of hotel rooms. Hotels are members of a class of things called Lodging. By recognizing this classification, we are conceptualizing the very nature of the resource. We call this conceptualization the resource's

Type. There are many Types of travel resources; each has its own niche in the taxonomy of travel resources.

The next aspect of our example resource (room 703) is its *Location*. This aspect allows us spatially relate the room and the rest of the world. Knowing that it is room 703 of a particular hotel allows us to know (or learn) the exact position of the room within the hotel. But this is only the base of the hierarchy. We also know that the hotel is in Fort Worth, which is a place in Texas, which is a state in the United States, which is a country in North America, which is part of planet Earth (and so on). A geographical frame of reference is obviously very necessary for travel.

The final aspect of our example resource describes how we may interact with (or reserve) it. There are many flavors of interaction that we may be interested in with respect to this hotel room. For example, we may want to know its cost or its availability, information that is best obtained directly from the provider. Alternatively, we may be interested in "reserving" the hotel room, or in canceling a reservation we previously made; both of these transactions require a granular knowledge of how to communicate with the provider of the room.

A directory-like structure may serve as the basis for organizing travel resources, and additionally incorporate references to the above-indicated aspects of a travel resource. This system is capable of providing the fine-grained semantics necessary for machines to navigate the directory structure and communicate with providers about travel resources, and thus represents a compromise between the breadth and ease-of-use of the Yahoo! and Google directories and the well-defined types of the OTA specifications.

4. Concluding remarks

In this note we have analyzed two approaches to organizing and describing travel resources: the schemes employed by two popular web directories: Yahoo! and Google, and the specifications designed by the Open Travel Alliance for B2B communications between travel industry partners. Furthermore, we have outlined a method of classifying travel resources that integrates the best aspects of both approaches, while avoiding some of their pitfalls.

Our investigation into classifying travel resources was motivated by the need to efficiently organize information about resources on the Internet in our own agent-based travel support system [8, 14, 16]. The use of software agents for discovering and delivering travel content requires an effective system for storing and recovering information, one that allows agents to navigate and manipulate content on a semantic level.

The proposed solution, based on triplets describing travel resources is currently being implemented using the ebXML Registry/Repository [23] as its centerpiece [14, 17], as well as agent-based extraction of information from the

Web [24-25]. We will report on the progress of our investigations in the near future.

Refereces

[1] Labrou Y., Finin, T.W., Yahoo! As an Ontology: *Using Yahoo! Categories to Describe Documents*, CIKM, (1999) 180.
[2] Galant V., Jakubczyc J., Paprzycki M., *Infrastructure for E-Commerce*, in: Nycz M., Owoc ML (eds.), Proceedings of the 10th Conference on Knowledge Extraction from Databases, Wrocław University of Economics Press, (2002) 32.
[3] Levy A.Y., Weld D.S., *Intelligent Internet Systems*, Artificial Intelligence, 118 (2000) 1.
[4] Hendler J, *Agents and Semantic Web*, IEEE Intelligent Systems Journal, 16 (2001) 30.
[5] Sandholm T., Lesser V., *Issues in Automated Negotiation and Electronic Commerce: Extending the Contract Net Framework*, Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95), (1995) 328.
[6] Walla V., Byde A., Cliff D., *Evolving Market Design in Zero-Intelligence Trader Markets*, Proceedings of the IEEE Conference on Electronic Commerce, (2003).
[7] Erdur R.C., Dikenelli O., *A Multi-Agent System Infrastructure for Software Component Market-Place: An Ontological Perspective*, ACM SIGMOD, 31 (2002) 55.
[8] Angryk R., Galant V., Gordon M., Paprzycki M., *Travel Support System – an Agent-Based Framework*, Proceedings of the International Conference on Internet Computing (IC'02), CSREA Press, (2002) 719.
[9] Galant V., Paprzycki M., *Information Personalization in an Internet Based Travel Support System*, Proceedings of the BIS'2002 Conference, (2002) 191.
[10] Gordon M., Paprzycki M., Galant V., *Knowledge Management in an Internet Travel Support System*, in: Wiszniewski B (ed.), Proceedings of ECON2002, ACTEN, (2002) 97.
[11] Jakubczyc J., Galant V., Paprzycki M., Gordon M., *Knowledge Management in an E-commerce System*, Proceedings of the Fifth International Conference on Electronic Commerce Research, Montreal, Canada, October, CD (2002).
[12] Paprzycki M., Angryk R., Kołodziej K., Fiedorowicz I., Cobb M., Ali D., Rahimi S., *Development of a Travel Support System Based on Intelligent Agent Technology*, in: Niwiński S (ed.), Proceedings of the PIONIER 2001 Conference, Technical University of Poznań Press, (2001) 243.
[13] Paprzycki M., Gordon M., Gilbert A., *Knowledge Representation in the Agent-Based Travel Support System*, in: Yakhno T (ed.), Advances in Information Systems, Springer-Verlag, (2002) 232.
[14] Paprzycki M., Gordon M., Harrington P., Nauli A., Williams S., Wright J., *Using Software Agents to Index Data for an E-Travel System*, Proceedings of the ABC Symposium, Orlando, Florida, July, (2003), to appear.
[15] Paprzycki M., Kalczyński P.J., Fiedorowicz I., Abramowicz W., Cobb M., *Personalized Traveler Information System*, in: Kubiak BF, Korowicki A (eds.), Proceedings of the 5th International Conference Human-Computer Interaction, Akwila Press, (2001) 445.
[16] Wright J., Williams S., Paprzycki M., Harrington P., *Using ebXML Registry/Repository to Manage Information in an Internet Travel Support System*, Proceedings of the 6th BIS Conference, Colorado Springs, June (2003), to appear.
[17] http://www.disa.org/.
[18] http://www.wedi.org/.
[19] *Open Travel Alliance Schema Descriptions and Examples, Specification: 2002B, 2003A*, http://www.opentravel.org/2003a.cfm, (2002).
[20] http://www.daml.org/
[21] http://www.semanticweb.org/
[22] http://www.ebxml.org/

[23] *OASIS/ebXML Registry Services Specification v2.0*, http://www.oasis-open.org/ committees/regrep/documents/2.0/specs/ebrs.pdf
[24] Arjona L., Corchuelo R., Ruiz A., Toro M., *A Practical Agent-Based Method to Extract Semantic Information from the Web*, Lecture Notes in Computer Science, 2348 (2002) 697.
[25] Knoblock CA, Lerman K, Minton S, Muslea I, *Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach*, IEEE Data Engineering Bulletin, 23(4) (2000) 33.