TOWARDS THE SEMANTIC WEB IN E-TOURISM: CAN ANNOTATION DO THE TRICK?

Hepp, Martin; Siorpaes, Katharina; Bachlechner, Daniel; Digital Enterprise Research Institute, University of Innsbruck, Technikerstrasse 21a, A-6020 Innsbruck, Austria, {martin.hepp|katharina.siorpaes|daniel.bachlechner}@deri.org

Abstract

Semantic Web technology may support more advanced E-Commerce. Namely the representation of products and services in the form of ontologies will simplify the automated extraction and processing of explicit information and will make implicit information available for the discovery and comparison of offerings. One common assumption is that the Semantic Web can be made a reality by gradually augmenting the existing data (mainly HTML/XHTML) by ontological annotations, derived from the non-machine-readable content of today, and that the main limitation of today's Web is the „needle-in-the-haystack" problem: everything is on the Web, but we only have insufficient methods of finding and processing what's already there. In this paper, we (1) evaluate whether this assumption is appropriate for the tourism industry, (2) show, based on a quantitative analysis of Web content about Austrian accommodations, that even a perfect annotation of existing Web content would not allow the vision of the Semantic Web to become a short-term reality for tourism-related E-Commerce, and (3) discuss the implications of these findings for various types of E-Commerce applications that rely on the extraction of information from existing Web resource, and stress the importance of Semantic Web services technology for the Semantic Web.

Keywords: E-Tourism, Semantic Web, Ontologies, Annotation, Information Extraction, Semantic Web services.

1 INTRODUCTION

The tourism domain can especially benefit from sophisticated E-Commerce solutions and Semantic Web technology, due to the significant heterogeneity of the market and information sources, and due to the high volume in online transactions (Werthner and Klein 1999). It can also be observed that E-Commerce has transformed the business models and processes of the tourism industry much quicker and more substantially than those in most other B2C arenas (Werthner and Ricci 2004). However, current Web technology suffers from the following limitations:

1. Search is limited to keywords and cannot be based on concepts. Thus, homonyms (the same word represents different concepts depending on the context, e.g. the car "jaguar" vs. the animal) result in a lack of precision in the search, while synonyms (there are multiple words that represent the same concept, e.g. "car" and "automobile"), and the lack of generalization/specialization relationships (e.g. a search for "accommodation" will not find pages that contain just the word "hotel"), result in a low recall, i.e., not all matching resources are found.
2. The relationship of Web resources to keywords cannot be expressed in queries (e.g. whether a Web page contains an *offer* of a specific product or a *review*). This further reduces the precision of information retrieval on the Web.
3. The support for information extraction from Web pages is very limited and must rely on volatile and error-prone assumptions about the binding between document structure and the semantics of data. This technique is known as "screen-scraping".
4. Facts from multiple sources cannot be automatically combined in order to answer a query.
5. Representational mismatches between queries and facts cannot be resolved.

Inside managed Web portals, some of these limitations can be overcome, e.g. if the portal owner collects and augments the data from individual market participants, and invests effort in the maintenance and cleansing of data. On a Web scale, however, the limitations listed above are a relevant bottleneck. Try to find "a good hostel or a cheap hotel in the North of London" via Google and you get a hands-on impression of the size of this problem.

The Semantic Web aims at making the wealth of information that is available on the Web accessible to more precise search and automated information extraction and processing, based on a machine-readable representation of meaning in the form of ontologies. This general vision is of course not limited to commercial transactions, but the potential for more advanced E-Commerce is very significant and thus receives a lot of attention.

The core components of Semantic Web technology are

1. XML as a generic serialization syntax with mature tool support,
2. RDF as a data model for the representation of Semantic Networks in a distributed fashion,
3. ontology languages like RDF-S, OWL, and WSML, for the representation of a domain of discourse,
4. (TBox) ontologies, like e.g. WordNet, Cyc, TOVE, Dublin Core, FOAF, or Harmonise, and
5. (ABox) data, which can be regarded part of the ontology or not, depending on the respective research community.

In the E-Commerce domain, these parts in combination can be used for the representation of products and services descriptions in a machine-accessible form and will thus simplify the automated extraction and processing of *explicit* information, and they will make *implicit* information available for the

discovery and comparison of offerings. With implicit information we mean such facts that can be inferred from others, for example the fact that since "Finchley Road" is in the North of London, a hotel that is located in Finchley Road is located in the North of London.

The potential for Semantic Web-enabled E-Commerce goes way beyond just transaction support; it will allow the automation of very sophisticated functionality, e.g. personalization, agents performing multi-dimensional trade-off decisions about alternative offerings based on complex user preferences, recommender systems, or mobile E-Commerce applications.

In the context of the Semantic Web, *annotation* means adding machine-readable information to existing content (cf. Lassila 1998, Heflin et al. 2003); it can support a multiplicity of applications, e.g. more precise discovery or adaptation of content (Hori 2003). Annotation can be done entirely manually, or semi-automatically based on Human Language Technology (HLT). The resulting annotations can be either stored in the very same document, in an external repository, or they can be generated on the fly using lightweight Human Language Technology.

One common assumption is that the Semantic Web can be made a reality by gradually augmenting the existing data (mainly HTML/XHTML) by annotations, and that the main bottleneck of today's Web is the „needle-in-the-haystack" problem: everything is there, but we only have insufficient methods of finding and processing what's already on the Web.

In this paper, we evaluate whether this assumption is appropriate for the tourism industry and show that, at least for the Tyrol region in Austria, even a perfect annotation of existing Web content would not allow the vision of the Semantic Web to immediately become a reality for tourism-related E-Commerce.

1.1 Related Work

Related work falls mainly into the following two categories:

Semantics-based Techniques in Tourism: Werthner and Klein (1999) provide an overview of the relationship between ICT innovations and the tourism industry. Fodor and Werthner (2004) present Harmonise, a project that deals with business integration in tourism using ontologies for mediation. Werthner and Ricci (2004) show the strong pickup of E-Commerce transformation in the tourism sector. Maedche and Staab (2002) discuss possible applications of Semantic Web technologies in the tourism domain. Bullock and Goble (1998) describe a Hypermedia system for tourism that is based on Description Logic. Dogac et al. (2004) discuss the benefit of adding semantics to Web services descriptions in the travel industry. Schwabe and Prestipino (2005) show that travel communities provide better travel information quality than vendor-operated systems.

Annotation: Heflin et al. (2003) explain the role of ontologies and annotation in the Semantic Web by describing a prototype that uses the SHOE ontology language. Hori (2003) addresses the role of annotation for Web content adaptation. Popov et al. (2003) present the semantic annotation platform KIM, which uses human language processing techniques to automatically annotate Web sites. KIM is equipped with an upper-level ontology as well as a knowledge base containing a collection of general concepts. A detailed comparison of existing annotation platforms is given in Reeve and Han (2005). Song et al. (2004) propose a system for the mechanized extraction and annotation of instance data from dynamic Web pages. Stojanovic et al. (2002) present a technique for migrating data-intensive, dynamic Web sites to the Semantic Web.

We do not know of any other work that has quantitatively analysed the sufficiency of existing Web content with regard to annotation for the Semantic Web, neither in general nor with the specific focus on E-tourism.

1.2 Our Contribution

In this paper, we provide a quantitative analysis of whether existing Web resources in the E-tourism domain in Austria contain a sufficient amount of information for making the Semantic Web in this sector a reality on the basis of annotating publicly available HTML/XHTML data. This includes the following steps: We (1) develop an evaluation framework of relevant information categories derived from a survey of the Austrian Chamber of Commerce on needs of individuals seeking an accommodation, (2) take a representative random sample of all officially registered accommodations in the state of Tyrol and the respective Web resources, and (3) manually analyze the amount of information per category that is available for each accommodation in the sample. Then, we (4) show that these findings are incompatible with the assumption of achieving the Semantic Web by annotating human-readable content published on the Web for the tourism sector in Austria, (5) discuss the implications of these findings for various types of E-Commerce applications that rely on the extraction of information from existing Web resources, and (6) substantiate the claim that the annotation of functionality by means of Semantic Web services is a core component for making the Semantic Web a reality in E-tourism.

The structure of the paper is as follows: In section 2, we describe our research methodology. In section 3, we present the resulting data and highlight significant observations. In section 4, we discuss the implications of our findings for the implementation of Semantic Web technology in the tourism domain and for E-Commerce research in general. Section 5 summarizes the main points of our work.

2 METHOD

In this section, we describe our research framework and methodology and justify necessary assumptions.

2.1 Approach

First, we identified information categories that are relevant for individual consumers looking for travel accommodation. Our main input for this were the results of a survey by the Austrian Chamber of Commerce in 2001 (Dolnicar and Otter 2001), which describes the information needs of consumers when seeking an accommodation. Since the study refers to the context of individuals looking for an actual accommodation, we added the category “availability”, which is not listed explicitly in the survey because it applies to all individuals. Then, we created an ordinal scale for the amount of information per category, ranging from 0 (no information) to 5 (comprehensive coverage of all aspects).

Second, we obtained the official directory of all legal accommodations located in the state of Tyrol (n=4,665). This can be regarded as a very reliable data source of the full market participants on the supply side, since it a mandatory legal requirement to register accommodations made available to the general public in Austria.

Third, we took a random sample (n=100) of the listed accommodations, and for each entry in this sample, searched the Internet for an official Web page. If we could not find a Web resource or if we had doubts about the identity, we called the owner or operator of the accommodation for clarification.

Fourth, we checked the leading Austrian tourism portal Tiscover (http://www.tiscover.at) for entries covering the very same sample.

Fifth, we manually analyzed the content of both the respective vendor-operated Web resources and the Tiscover entries, and graded the amount of available information using the predefined ordinal scale.

Sixth, we aggregated the results and determined the amount of Web resources and portal entries that provide at least a “sufficient” amount of information in the respective category according to the

grading scheme. Sufficient was defined in the sense that all information is given that an average consumer needs in order to determine his or her perceived utility[1] of an available accommodation, i.e. to make a reservation decision.

2.1.1 Relevant Information Categories

The survey by the Austrian Chamber of Commerce in 2001 (Dolnicar and Otter 2001) reveals that the most relevant information categories for individuals seeking an accommodation are location, price, room features, hotel star rating, breakfast/meals information, access and directions, details about the staff, accommodation features, swimming pool, technical equipment, parking/garage, images, sauna, and fitness room, in this descending order. Figure 1 illustrates this ranking of information needs.

Since the study targets the phase of information gathering during the process of decision making, we can implicitly assume that availability information is needed by 100 % of this group, because the consumers are only seeking available accommodations (the context of the survey was search for available accommodations).

Additionally, we introduced the categories "type of accommodation" (hostel, bed and breakfast, hotel, etc.), since useful information may be inferred from that later on (e.g., a hotel must offer certain services, such as breakfast and a front desk).

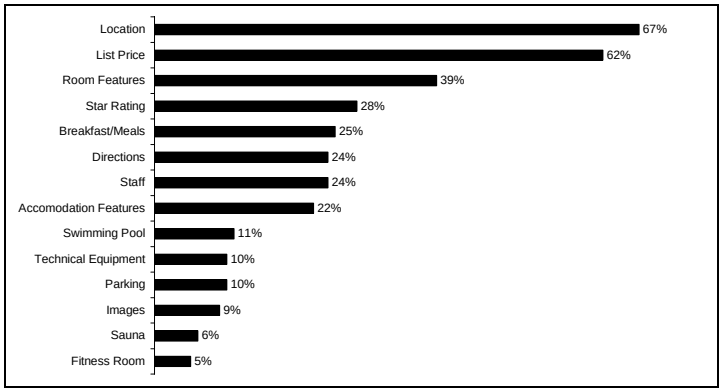

Figure 1. Most relevant information categories, based on Dolnicar and Otter (2001)

Summing up, 67 % of users check the location of an accommodation, which includes the geographical area as well as proximity to infrastructure. Information about the price of an accommodation is requested by 62 % of users (the remaining do likely rely on private estimates). Next follows the category "room features" with 39%. The information about the star rating of an accommodation is relevant for 28% of potential guests. From a Semantic Web perspective, the star rating is especially interesting, because it is based on the availability of well-defined accommodation features (e.g. "hair dryer"). Thus, much implicit information can be inferred from the star rating.

Almost of equal importance are the categories breakfast and meals, access and directions, and staff; they are requested by about 25% of the target group. Less relevant categories are swimming pool, technical equipment of an accommodation, information on parking or a garage, images, sauna, and fitness room.

[1] When we use "utility" in this paper, we mean it in the economic sense.

2.1.2 Scale for the Amount of Information

In order to rate the amount of information that is provided per each category, we used the following ordinal scale, as shown in Table 1.

Points	Amount of Information in this Category
0	No information
1	Some rare information, but not useful
2	Hardly useful information
3	Sufficient information for decision-making
4	Good information
5	Comprehensive information

Table 1. Evaluation scheme for the amount of information per category

0 points mean that the Web site provides no information at all in the respective category. 1 point means that the Web site mentions the category. However, no useful information is offered. When there is a little bit of useful information, 2 points are given. 3 points represent sufficient information in this category. This is a significant step, as we defined 3 points as sufficient for reusing the information in Semantic Web applications. We define "sufficient" in the sense that there is enough information to make a reservation decision without calling the accommodation. 3 points is the minimum requirement to enable the user to check whether the accommodation matches his or her requirements. 4 points mean that there is detailed information and 5 points mean that the Web site provides complete information in a category.

In some information categories, only the minimum (0) and maximum value (5) can be usefully applied. For example, a Web site either lists the star rating or it does not. As we wanted to be able to compare multiple categories, we nonetheless decided to use the same scale for all dimensions.

We did not validate the correctness of given information, i.e. we neither called the hotels to verify that the information on the Web site was still accurate, nor did we use external sources to validate the facts. Such would be separate research questions.

2.2 Data Sources

The basis for our analysis is the official list of all registered accommodations in the state of Tyrol. Using a random number generator, we took a random sample of the listed accommodations, and for each entry in this sample, searched the Internet for an official Web page. If we did not find a Web resource, or had doubts about the identity, we called the owner or operator of the accommodation for clarification. Additionally, we checked for an entry for the respective accommodation in the Tiscover portal

Then, we analyzed the content of both the respective Web site and the Tiscover portal entries manually and graded the information quality using the predefined ordinal scale.

3 RESULTS

In this section, we summarize the results of the analysis and highlight significant observations.

3.1 Availability of Specific WWW Resources

The random sample consists of 100 accommodations out of a list of 4,556 registered accommodations. Using common search engines, we checked each of them for a Web site and their membership in a tourism portal. Out of the 100 accommodations in the sample, 60 maintain a Web site individually for

this accommodation, either operated by the hotel owner or managed by a service provider. Additionally, all of these 60 are members of the Austrian tourism portal Tiscover. 33 are *only* represented in the Tiscover tourism portal. 5 % cannot be found at all on the Web but their existence could be verified by phone, and 2 % do either not exist any longer or could not be found at all. Figure 2 illustrates the findings.

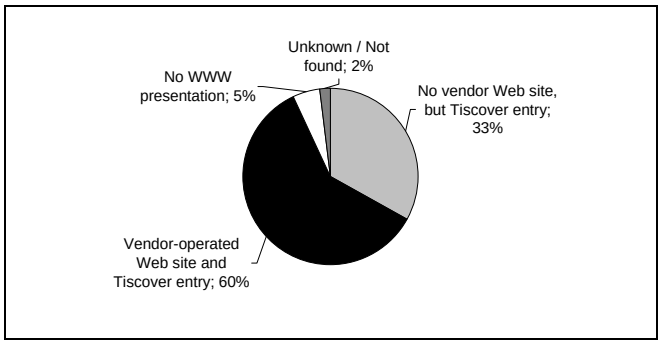

Figure 2. Availability of Web resources for accommodations in the sample (n=100)

3.2 Coverage of Information Categories

In the following, we focus on the analysis of the Web sites operated by or on behalf of the management of the accommodation, i.e. the site-specific Web pages. This distinction between vendor-provided data and portal data is very important, since Tiscover is a managed Web site and does not grant external access to the full internal database. Thus, the data available inside cannot be used for external annotation, and the discovery process is under the control of the Tiscover engine. In other words, Tiscover will constrain access to the data further than the single vendors. Also, the potential contribution of Semantic Web technology for tourism portals is likely smaller as compared to individual Web sites, since portals put a lot of effort into resolving inconsistent data representation. Also, screen-scraping of portal data might be subject to technical and legal constraints.

The analysis of the Web sites was done manually using the scale described in section 2.1.2. We went through the whole site, gathering information about each category, while ignoring the presentation (formatting, highlighting, etc.) of the data. We only took into account explicit information, e.g., we did not infer facts from the images.

In the following we describe the percentage of Web sites that contain at least a sufficient amount of information in the various categories. As detailed above, sufficient means a rating of 3 points or more in the grading scheme

Two categories are covered quite well, the type of accommodation and the list price. 98 % of the 60 individual Web pages contain sufficient information about the type of accommodation (hostel, bed and breakfast, hotel, etc.). 70 % give precise information about the list price, which is, however, often subject to negotiations and thus discounts or surcharges. The information category “price” ranks second in the user profile as it plays a predominant role in the decision making process.

58 % of the Web sites contain images of the building, the rooms, the infrastructure, or any combination of this. This is an important finding, since images usually convey a lot of information for human users but are rather hard to annotate in a semi-automated fashion.

About half of the Web sites contain sufficient information about the location (55 %), room features (57 %), directions and access (55 %), and accommodation features (52 %). 43 % indicate whether a sauna is available.

Poorly covered are star rating (37 %), parking (33 %), availability of a swimming pool, breakfast and meals. Extremely rare is sufficient information about fitness rooms (10 %), the staff (3 %), and technical equipment (2 %).

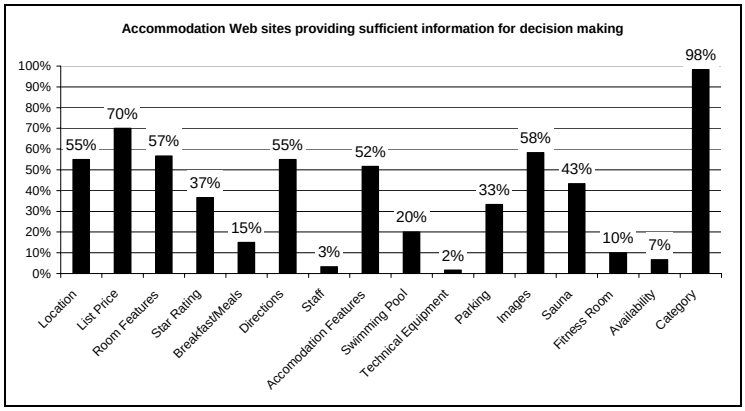

Figure 3. Percentage of specific Web sites with sufficient information for decision making

Only 7 % of the Web sites offer sufficient room availability information. Sufficient in this sense means that there is current data on the availability. General information about the amount of rooms or the duration of the season was not regarded sufficient.

While the direct numbers do not look too bad, it is important to look at the amount of Web sites that do *not* cover any of the core information categories: 93 % do not provide availability information and almost half of the Web sites do not cover most categories in a manner sufficient for decision making.

We also analyzed the distribution properties of the information quality in each of the 16 categories. Due to the limited space available, we can only present part of the results in here. Figure 4 shows the median value for each category in the available Web sites (i.e. the subset of the sample that reflects the specific Web sites, n=60).

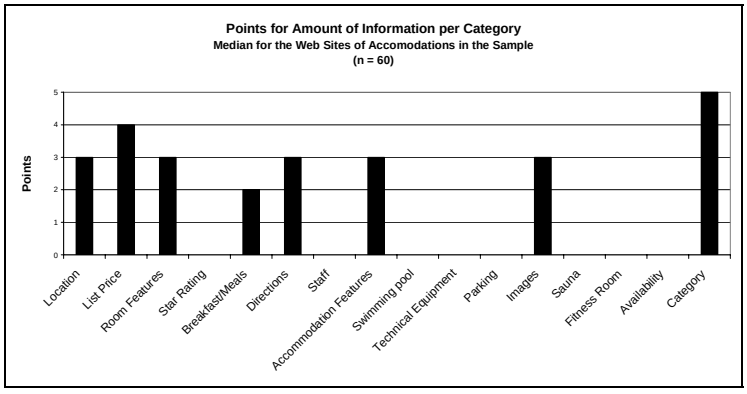

Figure 4. Median of information quantity per category

In only seven out of the 16 categories, the median value is equal to or greater than 3, i.e. sufficient. Since the median value very well reflects the center of a distribution, this is a clear indicator for lack of information on the Web sites. Eight categories have a median of zero, which means that at least half of the Web sites contain no information in the respective category at all. It is important to keep in mind that this is only for the subset of the available Web sites (n=60). If we judge the overall sample (n=100) on the same basis and include those accommodations that have no specific Web site with a rating of zero in all categories, the results would be even more significant.

3.3 Comparison of Information Supply and Demand Profiles

Apart from the absolute amount of information in each category, it is also interesting to compare the profile of provided information (information supply) with the profile of requested information (information demand). For this purpose, we drew a star diagram that compares the patterns of information needs and information supply. Figure 5 shows the resulting diagram. The grey-shaded shape inside represents the information needs, based on the Austrian survey (Dolnicar and Otter 2001). Each axis shows the percentage of potential guests who perceive the respective dimension as relevant for their decision-making. The second shape, indicated by the bold black line, indicates the percentage of individual Web sites in the sample that cover the respective category at least in a sufficient manner (3 points or more). One can see that there is a basic fit between the two shapes; however more information is needed than provided in the categories location, staff, technical equipment, and breakfast/meals.

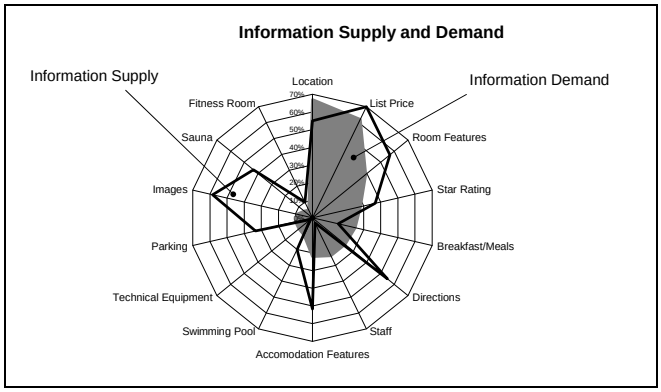

Figure 5. Information supply and demand for available WWW resources

Again, the percentages reflect only the 60 available Web sites. If we judge the categories based on the full sample size, then the supply shape (bold black line) would shrink by the factor of 0.6 (40 % have no specific Web site) in each dimension.

3.4 Availability of Information in Tourism Portals

In this section, we summarize the coverage of information needs among the 93 hotels that are listed in the Tiscover portal. Figure 6 shows a summary of the findings. We were surprised by the following observations:

1. Only 27 % of those hotels from our sample that are listed in Tiscover give price information (whereas 70 % of the hotel Web pages contain at least a list price).
2. Only a third of those hotels have sufficient room feature descriptions in Tiscover, while more than half (57 %) of the vendor-operated Web pages contain such detail.
3. 20 % of all listings in Tiscover give sufficient information about technical equipment, while only 2% of the Web pages contain such detail.
4. Three times as many hotels give current availability information on Tiscover (22 %) as compared to vendor operated Web sites (7 %). Still, the biggest part of all Tyrolean hotels does not provide current availability information anywhere on the Web.

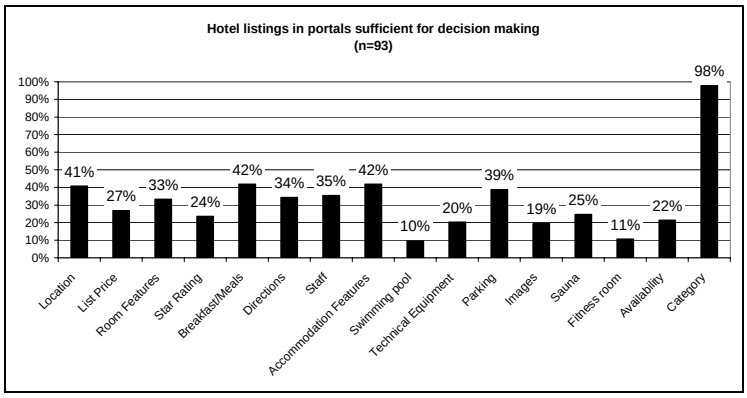

Figure 6. Percentage of specific Web sites with sufficient information for decision making

4 DISCUSSION

In this section, we discuss the implications of our findings for the implementation of Semantic Web technology in the tourism domain and for Semantic Web research in general.

4.1 Significant Findings

Only 60 % of the accommodations in the sample maintain a dedicated Web site, while 33 % rely solely on the tourism portal Tiscover to make their services visible on the Web. All of the 60 % of the accommodations in the sample are also listed in one or more tourism portals. That means that 93 % of all accommodations in the sample (and, since the mean of a sample is a reliable estimate for the mean of the population, the same order of magnitude of all Tyrolean accommodations) are listed in a tourism portal. Without further analysis, we can only speculate about the underlying rationales, but it is very likely that portals are a more feasible and more economic alternative than self-operated Web sites for the type and size of accommodations predominant in Tyrol. Also, portals might act as electronic marketplaces where supply and demand are aggregated and thus improve the efficiency of the market. This is of paramount importance for assessing the potential of Semantic Web technology in this sector.

We can see very clearly that the Web resources operated either by or on behalf of the management of a respective accommodation lack important information. Only 7 % offer room availability information, which is *the* most important fact when searching for a suitable offer. The remaining 93 % of accommodation Web sites require a user to either call or communicate by e-mail with the provider in order to get availability information. This is a serious obstacle for making the Semantic Web a reality in the E-tourism domain.

Only 7 out of 16 categories relevant for decision making are covered in a sufficient degree of detail on at least half of the Web sites.

The situation inside the tourism portal Tiscover is remarkably better, but in several ways still surprisingly insufficient. For eight out of ten hotels, no current availability data is available, and for 73 % of the hotels not even a list price can be retrieved.

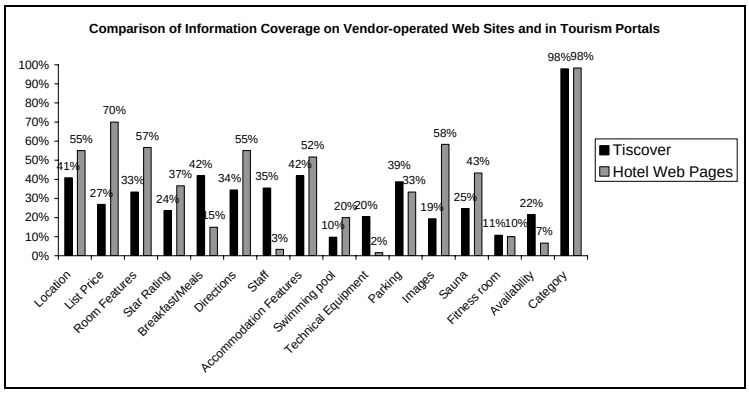

Figure 7. Variation in information coverage between accommodation Web sites and Tiscover

As long as the predominant way of E-Commerce in the tourism industry is end users searching the Web manually, the encapsulation of needed information inside tourism portals has no obvious adverse effects. It might be even more convenient for users to check one portal with a standardized user interface than visiting several Web pages. When transforming the Web into the Semantic Web by annotating available Web content, however, the managed portals are a significant challenge. This is because the internal database of the portal is not exposed to the general public. That means while we can build wrappers to annotate the functionality of the portal, we can not annotate the data contained inside. The discovery and matchmaking of consumer request and available supply is hidden inside the portal.

As a consequence, the Semantic Web cannot be made a reality in the sector we analyzed by annotating information on Web pages ("data-centric Semantic Web"). Rather, turning the Web into the Semantic Web requires annotating exposed functionality, i.e. services. In other words, the current market structures in the Tyrolean tourism industry likely require Semantic Web services technology, e.g. WSMO or OWL-S, in order to make the Semantic Web a reality in this domain. This is insofar interesting as the Semantic Web services research community is currently comparatively small in comparison to the general Semantic Web research community. For example, Google Scholar returns 19,200 scientific documents for "Semantic Web", but only 1,820 for "Semantic Web services".

We assume that a combination of the following facts contribute to the situation that the total amount of human-readable information available on public Web sites is insufficient as a basis for a Semantic Web that supports automated discovery and negotiations:

Cost and benefit of providing information on the Web: It might be that the perceived or actual business gain does not outweigh the cost of providing and updating this information on the Web.

Technical limitations: Especially the lack of current availability data on the Web sites is, at least for smaller hotels and hostels, likely caused by the sheer technical difficulty of linking decentralized, isolated hotel management systems with a Web application.

Insufficient support for strategic behavior: In imperfect markets, revealing information is an important strategic action. E.g. a hotel might not want to make public that they have almost no bookings for a specific date, since this would give potential guests bargaining power. Also, the price quoted might be based on inferences about the willingness to pay of the potential guest; every good sales person on the phone will try to exploit this aspect. Also, market participants do not want to reveal information to the general public including their competitors, but will try to limit this to seriously interested customers. Current Web technology does support all this to only a limited amount.

As a preliminary summary we can observe that it is more attractive for market participants to expose *functionality* on the Web, e.g. in the form of a Web Service, while keeping most of the data internally, instead of publishing all relevant *facts* directly on the Web; and still do the portals not contain

sufficient detail to use the potential of Semantic Web technology for overcoming the limitations of today's Web for searching an accommodation.

At least for the Tyrolean accommodation sector we have convincing evidence that the Semantic Web cannot be made a reality by annotating available Web pages but that this rather requires access to the functionality of systems, i.e. Semantic Web services technology.

4.2 The Swimming Pool Problem: Open World Assumption Revisited

When taking Web content as the input for machine-readable representation, it is important to think about the meaning of the absence of information in a specific category. In the predominant Open World Assumption of the Semantic Web, a property that is not stated is regarded as unknown and nothing else. In other words, when there is no explicit statement about the availability of a swimming pool in a certain hotel, then we only know that there is no such statement. We do not know whether there is a swimming pool or not.

This is a fundamental difference to closed knowledge-based systems, where absence of a statement often meant that the respective fact is false. A detailed discussion about the advantages and disadvantages of the Open World Assumption (OWA) vs. the Closed World Assumption (CWA) is outside the scope of this paper; however, some domain knowledge can be applied when annotating Web content that describes features of offerings. For example, one could assume that for such features that are perceived as *adding* utility to a specific offering for the vast majority of consumers, the absence of information is a good proxy for the lack of availability of the feature itself. If a hotel does not say anything about its swimming pool, we have reason to believe that they do not have one. For characteristics with a negative or varying impact on the perceived utility, this assumption very likely does not hold. The absence of a statement "noisy street next to the hotel" can of course not be used to infer that there is no noisy street next to the hotel. We call this the "Swimming Pool Problem". This partly explains the low values for the categories "swimming pool" and "fitness room" in the analysis.

4.3 Implications

We can see that, contrary to the popular assumption, the problem with today's Web is not only that we have insufficient methods of searching the Web, but that a significant amount of information in still available only inside closed, managed systems, or offline. The available Web resources do simply not contain sufficient information, even from a perspective of a human user. Thus, we cannot make the Semantic Web a reality in the tourism sector by crawling and annotating the content of existing Web pages.

We have no reason to assume that the encapsulation of information inside systems will decrease, because it gives the owner of the system better control over the usage of the encapsulated content. This implies that the Semantic Web in the tourism sector will only become a reality by the use of Semantic Web services technology, i.e. by annotating exposed functionality instead of static content. In short this is a strong indicator that there will be no useful Semantic Web without Semantic Web services.

Also, Semantic Web research should revisit the rather naïve assumptions on the willingness of market participants to persistently reveal availability and price information to a general audience. No sane businessperson will publish its full inventory data to the general public.

5 CONCLUSION

We have presented evidence that even a perfect annotation of existing Web content would not allow the vision of the Semantic Web to immediately become a reality in the domain of tourism-related E-Commerce, as long as the annotation is limited to persistently published information. Our quantitative analysis of the tourism sector in Tyrol has shown that dedicated Web sites do not contain sufficient

information that would allow potential guests to make a reservation decision without additional e-mail or phone communication. In other words, the problem in this sector is not just the lack of machine-access to the Web content, but the lack of content itself.

Inside managed tourism portals, more of the required information is available online. However, the amount of facts stored even in the most popular tourism portal is still not sufficient and covers only a fraction of the information needs. Furthermore, managed tourism portals expose only well-defined functionality and access on their data. Especially, they keep control over search, discovery, and comparison and prevent direct access to the internal database on the Web. 93 % of the accommodations are represented in such tourism portals, which allow, from the technical point of view, already today online reservations and detailed category information. The lack of booking functionality is in here likely because hotel owners do not want to update available rooms or room volumes or do not want to pay fees to the portal operator for successful bookings.

Since we have no reason to assume that the encapsulation of information inside systems will decrease, we can assume that the Semantic Web will only become a reality if it includes the annotation of *functionality* and not just published information. In short, the vision of the Semantic Web will not become a reality without Semantic Web services technology, e.g. WSMO or OWL-S.

ACKNOWLEDGEMENTS

The work presented in this paper is partly funded by the European Commission under the project DIP (FP6-507483) and the TransIT Entwicklungs- und Transfercenter at the University of Innsbruck.

REFERENCES

Bullock, J. and C. Goble (1998). TourisT: The Application of a Description Logic Based Semantic Hypermedia System for Tourism. Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space - Structure in Hypermedia Systems (HYPERTEXT '98), Pittsburgh, PA, USA, ACM.

Dogac, A., Y. Kabak, G. Laleci, S. Sinir, A. Yildiz, S. Kirbas and Y. Gurcan (2004). Semantically Enriched Web Services for the Travel Industry. SIGMOD Record 33 (3), 21-27.

Dolnicar, S. and T. Otter (2001). Marktforschung für die oesterreichische Hotelklassifizierung (Market research for the Austrian hotel classification schema; in German). Vienna, Austrian Chamber of Commerce.

Fodor, O. and H. Werthner (2004). Harmonise: A Step Toward an Interoperable E-Tourism Marketplace. International Journal of Electronic Commerce 9 (2), 11-.

Heflin, J., J. Hendler and S. Luke (2003). SHOE: A Blueprint for the Semantic Web. Spinning the Semantic Web. D. Fensel, J. Hendler, H. Lieberman and W. Wahlster. Cambridge (MA) and London, The MIT Press, 29-63.

Hori, M. (2003). Semantic Annotation for Web Content Adaptation. Spinning the Semantic Web. D. Fensel, J. Hendler, H. Lieberman and W. Wahlster. Cambridge (MA) and London, The MIT Press, 403-429.

Lassila, O. (1998). Web Metadata: A Matter of Semantics. IEEE Internet Computing 2 (4), 30-37.

Maedche, A. and S. Staab (2002). Applying Semantic Web Technologies for Tourism Information Systems. 9th International Conference for Information and Communication Technologies in Tourism, ENTER 2002, Innsbruck, Austria, Springer.

Popov, B., A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff and M. Goranov (2003). KIM - Semantic Annotation Platform. 2nd International Semantic Web Conference (ISWC2003), Sanibel Island, Florida, USA, Springer.

Reeve, L. and H. Han (2005). Survey of semantic annotation platforms. Proceedings of the 2005 ACM symposium on Applied computing. Santa Fe, New Mexico, ACM Press, 1634-1638.

Schwabe, G. and M. Prestipino (2005). How Tourism Communities Can Change Travel Information Quality. Proceedings of the Thirteenth European Conference on Information Systems, Regensburg, Germany.

Song, H., S. Giri and F. Ma (2004). Data Extraction and Annotation for Dynamic Web Pages. 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04), Taipei, Taiwan.

Stojanovic, L., N. Stojanovic and R. Volz (2002). Migrating data-intensive web sites into the Semantic Web. ACM Symposium on Applied Computing (SAC 2002), Madrid, Spain, ACM Press.

Werthner, H. and S. Klein (1999). Information Technology and Tourism: A Challenging Relationship. Berlin etc., Springer.

Werthner, H. and F. Ricci (2004). E-Commerce and Tourism. Communications of the ACM 47 (12), 101-105.