

# Semantification of Query Interfaces to Improve Access to Deep Web Content

Arne Martin Klemenz, Klaus Tochtermann

ZBW – German National Library of Economics  
Leibniz Information Centre for Economics,  
Düsternbrooker Weg 120, 24105 Kiel, Germany  
{a.klemenz,k.tochtermann}@zbw.eu  
<http://www.zbw.eu/>

**Abstract.** This position paper as part of a PhD thesis is a contribution to an automatic retrieval of information from the *Deep Web*. Addressing current limitations of the *Deep Web Information Retrieval* leads to the prevailing lack of semantics regarding the retrieval process. Focusing this problem from the information providing services perspective, indicates the significant potential of additional semantic annotations provided by websites. *Web query interfaces*, the interfaces to the majority of available information on the Deep Web, are interpreted as *Semantic Deep Web Services* (SDWS). The introduction of a SDWS annotation leads to great potential for *Information Retrieval* services based on the large variety of information available on the Deep Web.

**Keywords:** Deep Web, Semantic Deep Web Service, web query interface, semantic annotation

## 1 Introduction

A continuously increasing amount of content on the web is not directly accessible and indexable by search engines. The content might, for example, be hidden in non-public, inaccessible areas or might be stored in background databases and therefore only accessible through *web query interfaces*. This part of the web is known as the *Deep Web* (or *Hidden Web*) in contrast to the *Surface Web* which can be easily accessed and indexed by common search engines [1].

The Surface Web consists of mostly static content, which is directly inter-linked with static hyperlinks. "Search engines rely on hyperlinks to discover new webpages [...]" [11], but static websites are outnumbered by dynamic websites on an extremely large scale and the web has been rapidly deepened [6]. The content as part of dynamic websites is mostly not accessible through static hyperlinks, as this content is dynamically enwrapped into web pages as the response to a query submitted through a web query interface. These are intended to be used by human users to retrieve content from a background database often containing highly relevant content of a specific domain. Common search engines do not

reach this part of the web. This is caused by the fact, that search engines "[...] typically lack the ability to perform form submissions" [11].

Considering current arising services on the web like the Google Knowledge Graph, "we can use [...] [these services] to answer questions you never thought to ask and help you discover more"<sup>1</sup>. These services are related to *Knowledge Discovery*, but in general the benefit from the automatic discovery of new knowledge from existing information on the web is depending on an excellent *Information Retrieval*. As the retrieval of information from the Deep Web is still limited, Knowledge Discovery services are also still limited in their potential. Therefore, more efficient and targeted retrieval mechanisms for the Deep Web are needed to achieve full potential of Knowledge Discovery services.

The usage of semantic annotations for information on the web play a crucial role "to assimilate information from multiple knowledge sources" [13]. This challenge has been addressed, resulting in standards like *Resource Description Framework in attributes* (RDFa) and Microdata markups like *schema.org* initiated by the search engine big players Bing, Google, Yahoo! and Yandex. Therefore, this paper addresses the improvement of accessing this semantically annotated content on the Deep Web.

## 2 Related Work

The retrieval and indexing of Deep Web content have been addressed from different perspectives in the past. The effort has mostly focused specific applications to discover, retrieve and index structured data from the Deep Web. This includes special emphasis on the automatic web query interface interpretation.

Common approaches focusing on exposing Deep Web content can be classified to *surfacing* and *virtual integration* approaches. The surfacing approach focuses a search engine initiated process to index the search result pages for pre-computed (randomized) queries to discover Deep Web content on large scale [10]. The virtual integration approach follows the data integration paradigm, using a mediator system to map queries to relevant web query interfaces [10]. The content, that is retrieved, is brought to the user by the virtual integration to the search result page. Both of these approaches have been approved as useful in some cases. But in general the virtual integration approach is related to a lot of manual effort setting up query mapping rules for each Deep Web query interface in the mediator system. Furthermore, the surfacing approach is too imprecise or ineffective and therefore not scalable regarding the pre-computation of queries for domain independent sets of Deep Web websites.

Regarding the discovery and cataloging of Deep Web sources Hicks et al. [8] highlight the challenges and demonstrate via prototype implementation, that their Deep Web discovery framework can achieve high precision using domain dependent knowledge for probing web query interfaces. Wenye et al. [15] focus "Manufacturing Deep Web Service Management [...] [by] Exploring Semantic

<sup>1</sup> <http://www.google.com/insidesearch/features/search/knowledge.html?hl=en>

Web Technologies” by semantically annotating the Deep Web Services to reflect their hidden, dynamic, and heterogeneous contents while the relevance of semantic annotations for the Deep Web has already been identified in 2003 by Handschuh et al. [5]. Whereas these publications as well as Chun et al. [3] discuss these challenges from the information retrieving services perspective this paper will set the focus to the information providing services perspective.

Furche et al. [4] introduced a promising automated form understanding ontology based approach, which is far beyond heuristics to fill out search forms [12], combining “[...] signals from the text, structure, and visual rendering of a web page”. But according to Li, Xian et al. in “Truth Finding on the Deep Web: Is the Problem Solved?” [9] the challenges arising from the Deep Web are regarded as not yet solved. In general, current approaches are still limited either in being domain specific or limited in their efficiency.

Until today, there still exists no general domain independent solution for the Deep Web Information Retrieval problem. Just a fraction of total available data in background databases may be covered by common state of the art approaches. This is particularly due to the fact, that for large data sets there exist nearly endless possible permutations of search results. This especially applies to the retrieval of dynamic content. Therefore, it seems to be improbable to improve retrieval and indexing mechanisms towards reaching a 100% coverage of all available Deep Web content. Consequently, this is not the focus of our current research. Currently still limited mechanisms have already “[...] succeeded largely by targeting narrow domains where a search application can be fine-tuned to query a relatively small number of databases and return highly targeted results” [14]. For that reason, we focus e.g. on the reduction of manual effort regarding the query mapping on the one hand and more precise query generation or pre-computation for the targeted retrieval from broader domains on the other hand. Therefore, this paper is intended to improve access to Deep Web content by providing great potential for new Information Retrieval mechanisms and for the significant improvement of previously existing mechanisms.

### 3 Approach

#### 3.1 Research Focus

To step forward towards a *Semantic Deep Web*, which is the superordinated long-term objective, it is necessary to focus on additional research questions resulting from previously identified limitations. For the targeted retrieval especially of dynamic Deep Web content, the need of an efficient and in an ideal case fully automatic approach is essential. Therefore, the focus needs to be set to these challenges: content providing service *Discovery, Invocation & Execution* and *Composition*. Addressing these challenges will ensure the discovery of appropriate web query interfaces providing access to relevant content ( $\rightarrow$  *Discovery*), the appropriate query mapping and query submission ( $\rightarrow$  *Invocation & Execution*) and the service interoperability ( $\rightarrow$  *Composition*). Common approaches for Deep Web Information Retrieval focus these challenges from the information

retrieving services perspective. The conceptual idea being introduced in this section focuses these challenges from the information providing services perspective.

Common semantic annotation standards like *RDFa* and *schema.org* microdata address particularly the annotation of web content and do not have means for the prevailing *lack of semantics* at the crucial point of the Deep Web Information Retrieval process. This crucial point is regarding the web query interfaces. To improve common crawling, indexing and content retrieval mechanisms and to ensure new mechanisms, a semantic annotation for query interfaces is suggested. This will reuse the query interfaces originally intended for human users in a combined computer and human readable format. The abstract concept, to describe query interfaces in a computer readable format, is derived from the semantic annotation of *Web Services*. Standards like *Semantic Annotations for WSDL and XML Schema* (SAWSDL) provide a machine readable Web Service annotation describing the functionality and retrievable data. A semantic annotation for query interfaces will provide machine readable information for henceforth called *Semantic Deep Web Services* (SDWS).

### 3.2 Semantic Deep Web Service annotation

An implementation of the SDWS annotation should meet the following fundamental criteria: SDWS interface semantics, providing a generalization of SDWS interfaces ( $\rightarrow$  *abstract*) with the ability to include own vocabularies as for example thesauri ( $\rightarrow$  *extendable*). Going more into details, a SDWS annotation prototype should provide information about general properties regarding the content that is provided by the SDWS ( $\rightarrow$  *content* properties) and concrete interface field properties to describe the semantic structure and internal structural dependencies of the SDWS interfaces ( $\rightarrow$  *field* properties).

The prototype SDWS *content properties* describe the content *domain* of the retrievable information, the content *language*, as well as the content *type*. The content type attribute may be described based on *schema.org* microdata and the supplementary usage of other vocabularies. An additional content property might provide information about the amount of available data (property: *count*). These content properties are just the extendable basis for this prototype providing general information about the retrievable content. Further ideas for the extension of SDWS content properties will be discussed in the following section.

A simple example for the SDWS content properties is provided in Fig. 1, describing the basic SDWS field properties for the interface of the subject portal *EconBiz*<sup>2</sup>. *EconBiz* provides highly relevant content for the domain of *economics* and *business studies* and access to specific content types (various types of *CreativeWork* and information about *Events*).

The prototype SDWS *field properties* describe the field *type* (e.g. *selectField*, *inputField*), as well as the input *domain* and output *range* of each particular SDWS interface field. The input *domain* attribute describes valid input values of a specified SDWS field. Furthermore, it is a trigger for the output *range* attribute,

<sup>2</sup> <http://www.econbiz.de/en/>

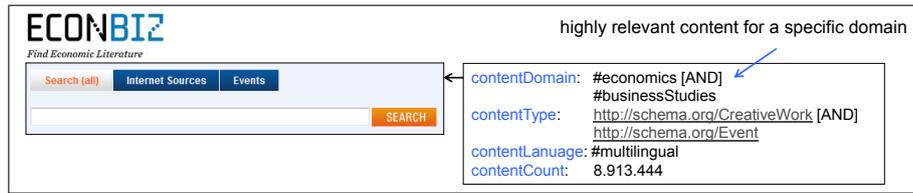


Fig. 1. SDWS content properties (example)

as its input value defines the restriction set for the retrieval process at time of form submission (examples, Fig. 2-4). Additionally, a *vocabulary* attribute may reference for instance a thesaurus that can be used as suggest-value vocabulary for the particular domain to ensure a targeted retrieval.

The basic SDWS field properties example in Fig. 2 refers to a standardized vocabulary, the *STW Thesaurus for Economics*. The STW provides "vocabulary on any economic subject" containing "[...] more than 6,000 standardized subject headings and about 19,000 entry terms to support individual keywords"<sup>3</sup>. This thesaurus is the basis for the annotation on metadata level in EconBiz and will therefore ensure a targeted retrieval. The benefit of the vocabulary property will especially apply to digital libraries but also to other domains. Therefore, regarding simple SDWS interfaces, this might be one of the most appropriate use cases for SDWS field properties as there is no complex interface structure.

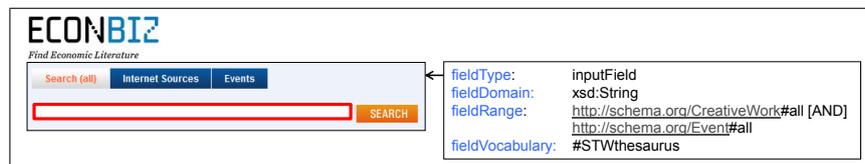


Fig. 2. SDWS field properties (basic example)

Focusing on more complex SDWS interfaces, the example in Fig. 3 contains *chunks* of related fields that affect each other. The first *selectField* as part of the marked chunk defines the relation to the other chunks. The second *selectField* as part of this chunk defines the input field domain and restricts the input field range of the *inputField* that is part of the focused chunk. Furthermore, Fig. 4 considers some exemplary effects triggered by the selection of different select values of the second *selectField* within the focused chunk in Fig. 3.

More complex examples as illustrated in Fig. 3 and 4 demonstrate that the semantic meaning behind a SDWS interface might be quite complex and automated form understanding approaches will quickly reach their limits. Especially

<sup>3</sup> <http://zbw.eu/stw/versions/latest/about.en.html>

6 Arne Martin Klemenz, Klaus Tochtermann

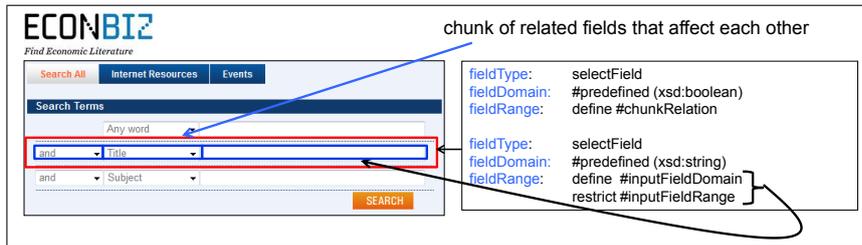


Fig. 3. SDWS field properties (chunk relation)

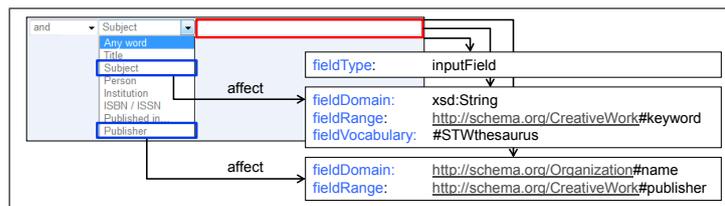


Fig. 4. SDWS field properties (triggered effects)

the automated detection of related fields and the detection of complex relations within chunks might be the most difficult part, where common approaches fail.

In addition to the SDWS interface annotation, it is advisable to link every SDWS interface from the websites root index. This will ensure a targeted SDWS discovery and can be realized by using XML Sitemaps to define a SDWS *retrieval index*. The SDWS annotation itself is suggested to be embedded directly to each particular SDWS interfaces.

#### 4 Benefit and further Use Cases

The introduced SDWS annotation will lead to great potential for new information retrieval mechanisms and plays a significant role for the improvement of current mechanisms. Queries might for example be automatically mapped to various SDWS at the same time based on the SDWS annotation (→ *abstract semantic querying*). In accordance with Heath et al. the vision of "users [...] interacting with the Web as a data space" [7] will therefore also benefit from the introduced SDWS annotation. In general, ensuring new user oriented services especially new Knowledge Discovery services based on the large variety of available information on the Deep Web is one of the major purposes. Furthermore, the SDWS annotation might also be used for purposes, not directly focusing on the retrieval itself, as for example reasoning processes for client side query interface input validations. Another use case focuses content licensing issues as these are a problematic topic for digital libraries. These may be addressed by adding licensing information directly to the SDWS interfaces by extending the introduced annotation prototype. This will be an appropriate possibility to pro-

vide licensing information exactly at that point where the information itself is being provided e.g. based on the *Creative Commons* licensing model.

Overall, this approach will make webmasters aware of their responsibility to add SDWS annotations to SDWS interfaces in addition to current semantic content annotations. This process requires additional effort on the one hand, but on the other hand it also enables the webmasters to control the information content that may be retrieved by various retrieving services like search engines. For now webmasters may only use common HTML attributes like *nofollow* or *noindex* and the *Robots Exclusion Standard* to control the crawling behavior on their websites. The SDWS annotation ensures the targeted influence of the webmaster. Furthermore, only the webmaster knows the exact semantic statement intended by the implemented SDWS interface. Regarding web content, search engines rely on semantic content markups as it is more reliable than current automatic content interpretation approaches. Therefore, it is obvious, that this will also apply to the annotation of SDWS interfaces.

## 5 Conclusion

This paper addressed the lack of semantic information regarding web query interfaces in the process of Information Retrieval from the Deep Web. Transferring the concepts of semantic web content annotations on the one hand and Semantic Web Service Descriptions on the other hand, leads to the great potential of semantic annotations for SDWS interfaces. Equivalent to semantic web content annotations, the SDWS annotation provides an unambiguous semantic interpretation of the SDWS interface. A variety of current information retrieval mechanisms and form understanding systems try to analyze SDWS interfaces automatically by focusing the Deep Web Information Retrieval challenge from the retrieving services perspective. Instead of relying on these, the introduced SDWS interface annotation is focusing this challenge from the information providing services perspective.

In general, this approach follows the open knowledge sharing paradigm as part of the *Semantic Web* vision from Berners-Lee et al. [2]. This is based on the assumption, that the information provided on websites is intended to be retrieved by various services. Any additional licensing issues restricting the retrieval and further usage of the retrievable information have also been addressed.

This approach will contribute to domain independent and automatic Information Retrieval mechanisms based on the introduced SDWS annotation. Manual effort for currently still limited Deep Web Information Retrieval mechanisms will be reduced or even eliminated. Furthermore, these retrieval mechanisms will benefit regarding their efficiency and can be adapted targeting broader domains.

## 6 Future Work

Future work will especially focus on the critical evaluation based on further research studies. The definition of a concrete SDWS annotation syntax based

on the usage of existing annotation standards will concern the challenge how retrieving services will learn to understand the SDWS annotation. Reduction of manual effort for the annotation process also requires further effort. A semi-automatic generation process may provide support for the definition of SDWS annotations. This process may be based on sampling and probing the background database utilizing promising automated form understanding approaches. This may lead to semi-automatic generation approaches for the SDWS annotation.

## References

1. BERGMAN, M. K. White paper: The deep web: Surfacing hidden value. *the journal of electronic publishing* 7, 1 (2001).
2. BERNERS-LEE, T., HENDLER, J., LASSILA, O., ET AL. The semantic web. *Scientific American* 284, 5 (2001), 28–37.
3. CHUN, S. A., AND WARNER, J. Semantic annotation and search for deep web services. In *E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008 10th IEEE Conference on* (2008), IEEE, pp. 389–395.
4. FURCHE, T., GOTTLÖB, G., GRASSO, G., GUO, X., ORSI, G., AND SCHALLHART, C. Opal: Automated form understanding for the deep web. In *Proceedings of the 21st international conference on World Wide Web* (2012), ACM, pp. 829–838.
5. HANDSCHUH, S., AND STAAB, S. *Annotation for the semantic web*, vol. 96. IOS Press, 2003.
6. HE, B., PATEL, M., ZHANG, Z., AND CHANG, K. C.-C. Accessing the deep web. *Communications of the ACM* 50, 5 (2007), 94–101.
7. HEATH, T., AND BIZER, C. Semantic annotation and retrieval: Web of data. *Handbook of Semantic Web Technologies* (2011).
8. HICKS, C., SCHEFFER, M., NGU, A. H., AND SHENG, Q. Z. Discovery and cataloging of deep web sources. In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on* (2012), IEEE, pp. 224–230.
9. LI, X., DONG, X. L., LYONS, K., MENG, W., AND SRIVASTAVA, D. Truth finding on the deep web: Is the problem solved? In *Proceedings of the 39th international conference on Very Large Data Bases* (2012), VLDB Endowment, pp. 97–108.
10. MADHAVAN, J., AFANASIEV, L., ANTOVA, L., AND HALEVY, A. Harnessing the deep web: Present and future. *4th Biennial Conference on Innovative Data Systems Research (CIDR)* (Jan. 2009).
11. MADHAVAN, J., KO, D., KOT, L., GANAPATHY, V., RASMUSSEN, A., AND HALEVY, A. Google’s deep web crawl. *Proceedings of the VLDB Endowment* 1, 2 (2008), 1241–1252.
12. MASANÉS, J. Archiving the hidden web. In *Web Archiving*. Springer, 2006, pp. 115–129.
13. MUKHERJEA, S. Information retrieval and knowledge discovery utilising a biomedical semantic web. *Briefings in Bioinformatics* 6, 3 (2005), 252–262.
14. OGRAPH, T., AMANCA, Y., AND MAAHS, Y. Searching the deep web. *Communications of the ACM* 51, 10 (2008).
15. WENYU, Z., JIANWEI, Y., MING, C., JIAN, W., AND LANFEN, L. Manufacturing deep web service management: Exploring semantic web technologies. *Industrial Electronics Magazine, IEEE* 6, 2 (2012), 38–51.