Semantic Enrichment of Web Query Interfaces to Enable Dynamic Deep Linking to Web Information Portals

Arne Martin Klemenz(✉) and Klaus Tochtermann

ZBW – Leibniz-Information Centre for Economics, Kiel, Germany
{a.klemenz,k.tochtermann}@zbw.eu

Abstract. This article addresses how to improve the automated accessibility and visibility of information from *Web Information Portals* and in particular virtual library systems. Information from web information portals could provide great value to satisfy information needs. But most of this information stays hidden in data silos which are part of that section of the web that is not indexable by common search engines and is therefore called *Deep Web*. Shared vocabularies like *Schema.org* helped to increase machine readability of structured information on the web in general, but markup vocabularies didn't increase the accessibility and visibility of information from data silos. This article addresses the limitations regarding the accessibility of information from data silos on the Deep Web and proposes an extension to Schema.org to fill the identified gaps. The extension improves the automated accessibility and visibility of information provided in web information portals by providing *Dynamic Deep Linking* capabilities to Deep Web data silos by lifting web forms of web information portals to the level of machine understandable semantic *Web Query Interfaces*.

Keywords: Dynamic Deep Linking · Schema.org extension · Web query interface · Web information portals · Virtual library systems

1 Introduction

The Web is a continuously growing diverse set of information. Various kinds of web-accessible data sources like search or information portals provide access to vast amounts of information and can be classified into service-oriented web services on the one hand and user-oriented web portals on the other hand. Service-oriented web services like web APIs (Application Programming Interfaces) provide information in a machine readable and accessible way to ensure its retrievability. In contrast to this, user-oriented web portals like information portals and in particular virtual library systems provide information in a way that is machine readable in terms of displaying purpose but primarily intended to be human readable. Information in user-oriented web portals – hereafter referred to as *Web Information Portals* – is usually retrieved based on user interactions like user-initiated web form submissions. These web form submissions become the focal point of interest in this article as it makes automated access and retrieval of information difficult and therefore causes this information to be hidden in *data silos* of the so called *Deep Web* [1].

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 553–559, 2017.
DOI: 10.1007/978-3-319-67008-9_45

The Deep Web refers to the part of the web that is not indexable by common search engines due to its limited accessibility. Web crawlers basically rely on hyperlinks to discover new information on the web [6]. The limited automated accessibility and therefore limited visibility of information from web information portals is caused by the likewise high complexity of web form submissions that rely on user input.

In contrast, the *Surface Web* [1] consists of information that can be easily accessed by common search engines. Most users even rely on search engines when searching information from known information sources like a well-known subject portal [9]. In contrast to classical library *OPAC*s (Online Public Access Catalogues) which provide a service for local users, modern virtual library systems aim to provide global services. Whereas classical OPACs often exclude search engine crawlers on purpose [2], virtual libraries rely on search engine optimizations to reach their targeted user group.

This article proposes the semantic enrichment of web forms based on an extension to the open vocabulary standard *Schema.org*. The extension investigates the potential of semantic annotations for web forms to provide *Dynamic Deep Linking* capabilities and therefore increase the automated accessibility of information from web information portals with a focus on the special characteristics of virtual library systems.

2 Related Work

The above described user behavior in information gathering corresponds with the search engines intention to refine their search results based on the usage of structured data. For example, this applies to rich snippets [11] as well as embedded additional information in search results pages. This was one of the major reasons for launching the Schema.org initiative. Schema.org schemata supported the semantic description of static entities with a focus on optimized ranking and rich snippet generation. The schemata had a wider range and less specialization in contrast to previous vocabularies and extremely improved machine readability of information that is covered by the provided schemata. But this just applies to information that was already accessible by search engines and so far just machine readable in terms of displaying purpose.

As the web is not just about static descriptions of entities, the Schema.org community announced Schema.org *Actions*[1] to describe actions performed on entities. One of these indicates search actions based on form submissions to ease access to annotated websites for web crawlers and provide *Sitelinks Search*[2] functionality. As lots of information on the web still cannot be reached due to the semantic complexity of web forms, this work investigates further potential of semantic annotations for web forms with the application of *Linked Data* principles. These principles refer to “best practices for publishing and connecting structured data […] to create typed links between data from different sources” [3] and led to improvements regarding the retrievability of information on the web, e.g. based on *Semantic Search* capabilities.

[1] Schema.org Actions: http://blog.schema.org/2014/04/announcing-schemaorg-actions.html
[2] Sitelinks Search: https://developers.google.com/structured-data/slsb-overview

The challenge of improving the automated accessibility of information on the web has been addressed from several perspectives in the past. On the one hand, Semantic Web Services play an increasingly important role in web data integration processes. In particular, *Hydra* a lightweight vocabulary has been published aiming to create hypermedia-driven Web APIs [8]. In contrast to this article, the Hydra W3C community puts its focus on service-oriented web services. Moreover, the *OpenURL* framework for open reference linking in the scholarly information environment provides linking capabilities to library services going beyond the classic notion of a reference link [10]. It "provides a standardized format for transporting bibliographic metadata […] between information services" [10]. The format is mainly used for link resolvers and has the basic concept of deep linking information in target services in common with this work. But in contrast to the service-oriented OpenURL format this work provides a more generic vocabulary with Dynamic Deep Linking capabilities to user-oriented web information portals.

On the other hand, previous research focuses on the extraction of information from Deep Web data sources. Special emphasis has been placed on surfacing information respectively the information extraction from the Deep Web [4] based on automated web form discovery, understanding and classification approaches [12, 13]. Additionally, general Deep Web harvesters have been developed [5, 14]. While most of these approaches address the Deep Web data integration challenges from the retrieving services perspective and have to deal with either strict limitations regarding their application domain or their efficiency, this work switches the perspective and addresses these challenges from the information providing services perspective [7].

3 Semantic Deep Search Extension

Semantic annotations provide further potential for the improved access to Deep Web data silos and in particular web information portals. Virtual library systems rely on expert generated bibliographic metadata to describe their provided content. This bibliographic metadata is usually based on authority data and vocabularies defined as thesauri which play a significant role in the targeted search. Whereas Linked Data principles have a widespread use in service-oriented information provision and access, they still lack recognition when accessing user-oriented web information portals. We investigate the further potential of expert generated bibliographic metadata based on additional semantic annotations for web forms following the Linked Data principles. Web forms are hereafter referred to as *web query interfaces* to indicate their relevancy for the automated accessibility of information in web information portals.

The *Semantic Deep Search Extension* to Schema.org should meet the following challenges to enable better automated access to Deep Web data silos:

1. Identify web query interfaces of web information portals (*Service Discovery*)
2. Select web query interfaces for specific information need (*Service Classification*)
3. Generate service-specific query URLs (*Dynamic Deep Linking*)

According to the introduced switch in perspective in contrast to approaches from the retrieving services perspective, *Deep Web Service Endpoints* (underlying retrieval

service of a web query interface) need to be self-describing in terms of general classification purpose and in terms of their detailed URL parameter assignment. As a result, service endpoints serve as semantic APIs to the primarily just user-accessible search functionality of web information portals. The overall objective of this approach is a framework for *Semantic Deep Search* based on *Dynamic Deep Linking*.

In contrast to Schema.org's syntactical *SearchAction* markup, the extension provides a vocabulary that is capable of describing the detailed web query interface semantics. General information for the discovery and classification of the service is provided based on so called *content* properties: e.g. content domain, language, content type and licensing information. Furthermore, this extension specifies detailed *WebFormElement* properties which correspond to the service parameters or sets of related parameters of a web query interface to describe *semantic parameter constraints* like input domain (valid parameter values) and output range (output restrictions triggered by input values) as well as *structural* or *semantic parameter dependencies*.

This model is based on the formalization of a web query interface. A web query interface can be formally described as a service with a set of input variables X which correspond to the service parameters. The result set of a query submission is based on several restrictions to the underlying dataset defined by each input variable $x_i \in X$. Each variable $x_i \in X$ has a specified input domain inp_i. An assigned variable value specifies a restriction out_i regarding the result dataset. The input domain inp_i and output restriction out_i of x_i can be formalized as graph patterns. Furthermore, there might exist semantic dependencies between related variables or sets of variables. For example, the value of a variable $x_i \in X$ (e.g. select field) might restrict or redefine the input domain inp_j and output restriction out_j of a related variable $x_j \in X$ (e.g. input field). The altered input domain inp_j' is a restricted subset or redefinition of inp_j and the output restriction out_j' is a restricted subset or redefinition of out_j. Therefore, the overall result of a query submission for a valid variable assignment $\mathrm{X} = \{x_1, \ldots, x_n\}$ can be specified by a conjunctive query: $\mathrm{CQ}(X) = T_0 \cap \left(\bigcap_{x_i \in X} T_{out_i} \right) = T_r$.

The result set can described as graph pattern T_r which is the conjunctive intersection of all graph subsets T_{out_i} defined by formal restrictions out_i of variables $x_i \in X$ and the provided dataset of a web information portal described by graph pattern T_0.

The Semantic Deep Search Extension for Schema.org provides the expressiveness to define these introduced formal parameter constraints and semantic dependencies. The full vocabulary extension is publicly available on the extension website[3].

In the following, selected aspects of a prototype annotation utilizing the implemented extension will be introduced to illustrate some of the main concepts like semantic parameter dependencies and parameter constraints. The prototype implementation is based on a web query interface from EconBiz[4]. The described example web query interface consists of three parameters: input field, select field and checkbox. The specified

[3] Semantic Deep Search Extension website: http://semdeepsearch.vocab-ext.appspot.com/
[4] EconBiz, subject portal for economics and business studies: https://www.econbiz.de/

select field defines restrictions to the whole range of attributes of the specified input field depending on its selected value. As an example, the restriction caused by the selection of the predefined value "Author" can be described based on the following annotation (in Notation 3; _:elem1 references input field):

```
@prefix schema: <http://schema.org/> .
@prefix gndo: <http://d-nb.info/standards/elementset/gnd#> .
... a schema:WebFormParameterValue ;
      schema:webFormParameterValue "Author" ;
      schema:definesWebFormElementRestriction
          [ a schema:WebFormElementRestriction ;
            schema:restrictedWebFormElement _:elem1 ;
            schema:webFormParameterVocabulary
               "http://d-nb.info/gnd/" ;
            schema:webFormParameterInputDomain
                [ a schema:ValueInputDomain ;
                  schema:valueInputDomainClass "gndo:Person" ;
                  schema:valueInputDomainProperty
                     "gndo:preferredNameForThePerson" ] ;
            schema:webFormParameterOutputRange "schema:author" ] ...
```

With ?val_elem1 as value of the input field element, the corresponding output restriction can be formalized as RDF statement:

```
?person a gndo:Person .
?person gndo:preferredNameForThePerson ?val_elem1 .
?publication a schema:CreativeWork .
?publication schema:author ?person .
```

The Deep Search URL which can be generated based on the semantic annotation is: https://econbiz.de/Search/Results?lookfor=[?val_elem1]&type=Author.

In addition, search operators have special relevance for virtual library systems. It is common to refine a search query based on classifications, descriptors or authority data. To be capable of describing specifications like these, the extension introduces *WebFormPrefixSearchOperator*s as shown in the following example annotation:

```
... a schema:WebFormPrefixSearchOperator ;
    schema:prefixSearchOperatorPrefix "gnd" ;
    schema:prefixSearchOperatorPrefixNamespace
       "http://d-nb.info/gnd/" ;
    schema:prefixSearchOperatorVocabulary "http://d-nb.info/gnd/" ;
    schema:prefixSearchOperatorInputDomain "gndo:Person" ;
    schema:prefixSearchOperatorRange "schema:author" . ...
```

This annotation will allow parametrized Linked Data based Deep Links like: https://econbiz.de/Search/Results?lookfor=gnd:[GND-IDENTIFIER]

Overall, automated systems are capable of generating service-specific query URLs for a specific information need to perform a *Semantic Deep Search* based on *Dynamic Deep Linking*. This means that automated systems, which understand the web query interface markup, are able to provide links to any search results page for any specific

information need. The entire markup example applying the full vocabulary range of the extension is available for review on the vocabulary extension website.

4 Summary and Future Work

This article introduced the *Semantic Deep Search Extension* to Schema.org to improve the automated accessibility of web information portals and in particular virtual library systems. These are part of the so called Deep Web and thus not indexable by common search engines. The extension adds further expressiveness to Schema.org *SearchAction*s to provide a semantic markup for *WebForm*s. As a result, the information providing web information portal is able to provide a self-describing semantic annotation that enables automated access. In contrast, previous research focused these challenges from the retrieving services perspective, e.g. based on form understanding. The semantic markup allows automated systems to lead users to the search results page of an annotated web information portal that is the most expedient according to their information need. The generation of query URLs is introduced as *Dynamic Deep Linking* and has great potential in combination with *Semantic Search* strategies.

Our future work will focus on the distribution of the introduced extension to the Schema.org community with the intention to enter the routine for official extension candidates. The widespread acceptance of the extension is fundamental to achieve its full potential and finally its acceptance by the search engines that are part of the Schema.org initiative. In addition, future work will concern evaluation studies.

References

1. Bergman, M.K.: White paper: the deep web: surfacing hidden value. J. Electron. Publish. **7**(1), 1–17 (2001)
2. Blandford, A.: Google, public libraries, and the deep web. Dalhousie J. Interdiscip. Manage. **11** (2015)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. In: Semantic Services, Interoperability and Web Applications: Emerging Concepts, pp. 205–227 (2009)
4. Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R.: Web data extraction, applications and techniques: a survey. Knowl.-Based Syst. **70**, 301–323 (2014)
5. Furche, T., Gottlob, G., Grasso, G., Guo, X., Orsi, G., Schallhart, C.: The ontological key: automatically understanding and integrating forms to access the deep Web. VLDB J. **22**(5), 615–640 (2013)
6. Henzinger, M.R.: Hyperlink analysis for the web. IEEE Internet Comp. **5**(1), 45–50 (2001)
7. Klemenz, A.M., Tochtermann, K.: Semantification of Query Interfaces to Improve Access to Deep Web Content. SDA, pp. 104–111 (2013)
8. Lanthaler, M., Gütl C.: Hydra: a vocabulary for hypermedia-driven web apis. In: LDOW, vol. 996 (2013)
9. Purcell, K., Brenner, J., Rainie L.: Search engine use 2012 (2012)
10. Van de Sompel, H., Beit-Arie, O.: Open linking in the scholarly information environment using the OpenURL framework. New Rev. Inf. Netw. **7**(1), 59–76 (2001)
11. Steiner, T., Troncy, R., Hausenblas, M.: How Google is using linked data today and vision for tomorrow. In: Proceedings of Linked Data in the Future Internet, vol. 700 (2010)

12. Wang, L., Hawbani, A., Wang, X.: Focused deep web entrance crawling by form feature classification. In: Wang, Yu., Xiong, H., Argamon, S., Li, X., Li, J. (eds.) BigCom 2015. LNCS, vol. 9196, pp. 79–87. Springer, Cham (2015). doi:10.1007/978-3-319-22047-5_7
13. Zhang, Z., He, B., Chen-Chuan Chang, K.: Understanding web query interfaces: best-effort parsing with hidden syntax. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 107–118. ACM (2004)
14. Zhao, F., Zhou, J., Nie, C., Huang, H., Jin, H.: SmartCrawler: a two-stage crawler for efficiently harvesting deep-web interfaces. IEEE Trans. Serv. Comput. **9**(4), 608–620 (2016)