Enhance the Interoperability of the Library Search Systems with zSAPN

Michalis Sfakakis
Laboratory on Digital Libraries & Electronic Publishing,
Department of Archive & Library Sciences, Ionian University
72, Ioannou Theotoki Str.
Corfu, Greece, GR-49100

sfakakis@ionio.gr

Sarantos Kapidakis
Laboratory on Digital Libraries & Electronic Publishing,
Department of Archive & Library Sciences, Ionian University
72, Ioannou Theotoki Str.
Corfu, Greece, GR-49100

sarantos@ionio.gr

ABSTRACT

The wide adoption of the Z39.50 protocol from the Libraries exposes their abilities to participate in a distributed environment. In spite of the specification of a unified global access mechanism from the Z39.50 protocol, unsupported Access Points result to query failures and/or inconsistent answers. A challenge to this issue is to substitute an unsupported Access Point with others, so that the most similar semantics to the original Access Point can be obtained. In this paper we present the zSAPN (Z39.50 Semantic Access Point Network), a system which enhance the interoperability of the library search systems, by exploiting the semantics from the Bib-1 Access Point official specification of the Z39.50 information retrieval protocol. zSAPN substitutes each unsupported Access Point with a set of other supported ones, whose appropriate combination would either broaden or narrow the initial semantics, according to the user's choice.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation*

General Terms

Standardization, Measurement.

Keywords

Z39.50, Interoperability, Semantic Access Point Substitution.

1. INTRODUCTION

The Z39.50 client/server information retrieval protocol [1] is widely used in Libraries electronic communication for searching and retrieving information from a number of diverse, distributed, heterogeneous and autonomous sources. According to Z39.50

architecture, every client can communicate with multiple servers (in parallel or sequentially) and every server can publish many sources, not necessarily with the same structure and search capabilities.

For the implementation of the search primitives, the protocol utilizes the concept of the abstract Access Point (i.e. what entities represent the search terms, like Title, Author, etc.), which the client can only use to query the sources. A server can supply access to many sources and for every source a different subset of the global set of Access Points could be supported. The Bib-1 Attribute Set is the only one required from the conformance requirements of the Z39.50 protocol, which supplies the *Use* attribute type for the specification of the valid Access Points.

Despite of the protocol specification of a unified global access mechanism, query failures and/or inconsistent answers are the pending issues when a query contains an unsupported Access Point. The inconsistent answers are generated from the arbitrary substitution of the unsupported Access Point with a supported one from the server. Unsupported Access points are very common, especially when searches are addressed to more than one source due to the variant or poor implementations of the protocol. Statistical figures -provided by IndexData1- presenting the "Ten most commonly supported Access Points" indicate that no single Access Point is universally supported by Z39.50 sources. The most commonly supported Access Point is the Title (91%) and then is the Subject heading (89.5%).

Our approach copes with the problem of the unsupported Access Points by suggesting the substitution of the unsupported Access Point with a set of others. When substituting the unsupported Access Point we avoid the failure of the query. Furthermore, and of grater importance, we are aware how the substitution affects the semantics of the original query, so we are able to post process the results if it is desirable. The proposed substitution methods produce the minimal set of Access Points that have as similar semantics to the original Access Point as can be obtained by exploiting the semantics of the Bib-1 Access Points. We describe the Access Points semantics using the RDF Schema language [11].

---

1 http://targettest.indexdata.com/stat.php. The Z39.50 servers were last tested on Friday 9th of November 2007.

The RDF Schema provides a mechanism for describing groups of related resources and the relationships between these resources, enforcing the description of domain specific semantics. In addition, this description could form a basis for the deployment of the Z39.50 search primitive semantics to the Semantic Web [2].

Traditional information integration architectures and their extensions [10, 5, 6] deal with the query rewriting problem, which could be considered as a more general case to the Access Point substitution problem. Their rewritings are based on the mapping rules between the global schema and the local schemas of the underlying sources without any further exploitation of the local schema semantics. Furthermore, our described case allows for further optimization.

The remaining sections of this study are presented as follows: section 2 highlights the Z39.50 protocol, its access model and describes the issues derived from the existence of unsupported Access Points. Section 3, describes the Bib-1 semantic-based Access Point Network. Thereafter, section 4 describes and presents zSAPN system for transforming Z39.50 queries using the Access Point substitution policies, and finally, section 5 concludes and presents a number of interesting issues arising from this work for further research.

2. THE Z39.50 INFORMATION RETRIEVAL PROTOCOL

The Z39.50 is a state-full protocol based on the client/server model [1, 3]. It defines a standard way for the communication between the client and the server, and provides them with the ability to interoperate regardless of the underlying source structure, search procedures and computer systems. The system level interoperability is approached by the definition of a set of specific services, and it is accomplished by the exchange of specific messages, between the client and the server. For the content semantics of the published sources (databases), the protocol defines a standard model in a record-based abstract view hiding the logical structure of the underlying source.

The query mechanism of the Z39.50 [1] protocol utilizes sets of predefined Access Points combined with specific attributes (i.e. Attribute Sets) in a number of different query language specifications (i.e. query types). The protocol specifies many different query types (called Type-0, Type-1, etc.) mainly affecting the syntax of the query. For every search term we have to define its characteristics by declaring the Attribute Set which it belongs to. The Attribute Set defines the valid Access Points (i.e. what entities represent the search terms) from a specific set of attribute types, the way the system will match them against the underlying data and the form in which the terms have been supplied.

For the accomplishment of the standard search primitives the general conformance requirements of the protocol specify that at least the Access Points defined in the attribute set Bib-1 and the query Type-1 for the query formulation have to be recognized (although not necessarily implemented). The Attribute Set Bib-1 consists of the following attribute types: *Use* (e.g. Title, Author, etc.), *Relation* (e.g. Equal, less than, etc.), *Position* (e.g. First in field, any position in field, etc.), *Structure* (e.g. phrase, word, etc.), *Truncation* (e.g. right, do not truncate, etc.) and *Completeness* (e.g. complete field, etc.).

The official semantics of the Bib-1 Access Points are defined in [12], which represents a consensus among the members of the Z39.50 Implementors Group (ZIG). The document defines the semantics of the Access Points using the tag values of representative MARC bibliographic format fields. As an example, the definition for the Access Point *Author-name-Personal* (or *use* attribute with value 1004) includes the data from the fields with MARC Tags {100, 400, 700, 800}.

According to the protocol, if a target does not support a given attribute list, the target should fail the search (i.e. query failure) and supply an appropriate diagnostic message, or the target will substitute it according to the 'Semantic Action' value. In most cases, the vast majority of the running Z39.50 servers ignores the 'Semantic Action' value and makes an arbitrary substitution of the unsupported attributes without informing the client.

When searching a Z39.50 source, the various implementations of the protocol result to query failures and/or inconsistent answers, despite of the unified access mechanism of the protocol. The varying implementations mostly differ either to the subsets of the supported attribute types, or to the supported query language characteristics. Our work is focused on the substitution for the unsupported *use* attribute type, which is also referred to as Access Point.

To summarize the problem of unsupported Access Points, consider the source Z_1 that supports the Access Points (*use* attributes) *Any, Subject* and *Title*. According to the Bib-1 semantics specification, the *Any* Access Point can be considered as the union of all the supported Access Points. Any query that refers to other Access Point than the aforementioned will either fail, or the source will arbitrary substitute the Access Point with one of the three supported. The arbitrary substitution will result to an answer with undefined semantics.

Instead of allowing the source to decide for the failure of the query or for the arbitrary substitution of the unsupported Access Point, our approach uses the semantic information from the Access Points specification and discovers which Access Points could be combined to produce a query with the nearest possible semantics to the query that uses the unsupported Access Point. For example, a commonly used Access Point is the *Author-Title-Subject*, requesting to match the searched term against the data from the Author or the Title or the Subject data fields. As we present in the next section (see figure 1), the Access Point *Author-Title-Subject* could be considered as a super set (or a super class) of the Access Points, *Name, Title, Subject, Music-key,* and *Identifier-thematic*. Also, the *Author-Title-Subject* could be considered as a subset (or a subclass) of the Access Point *Any*. For any query to the source Z_1 containing the unsupported Access Point *Author-Title-Subject* we can substitute this Access Point with the Boolean *AND* combination of the supported Access Points *Subject* and *Title*. This substitution for the Access Point *Author-Title-Subject* will restrict the results but if the demand is for increased results, we could substitute the Access Point to an as-well-supported general one, like the *Any* in the Z_1 source.

3. Z39.50 SEMANTICS ACCESS POINT NETWORK

We semantically associate the Access Points using their subset relationships. We consider an Access Point to be a subset

of another if the set of the data fields used to create the first one is a subset of the set of the data fields used to create the second one. As an example, consider the Access Point *Author-name* which, according to its definition [12], includes the data from the fields with MARC tags included in the set {100, 110, 111, 400, 410, 411, 700, 710, 711, 800, 810, 811}, and also, the Access Point *Author-name-conference* which includes the data from the set of fields {111, 411, 711, 811}. The Access Point *Author-name-conference* is considered being a subset of the *Author-name*.

We represent the relationships between the Access Points with a directed graph G in which the vertices represent Access Points and the arcs represent subset relationships. This graph has an arc <*i, j*> if and only if Access Point *i* is a subset of Access Point *j*. The Access Points *Author-name* and *Author-name- conference*, used in the previous example, will be represented by two vertices of the graph and their subset relationship from the arc <*Author-name- conference, Author-name*>. The construction of the Semantics-Based RDF Schema of the Bib-1 Access Points is based on our previous work described in [7].
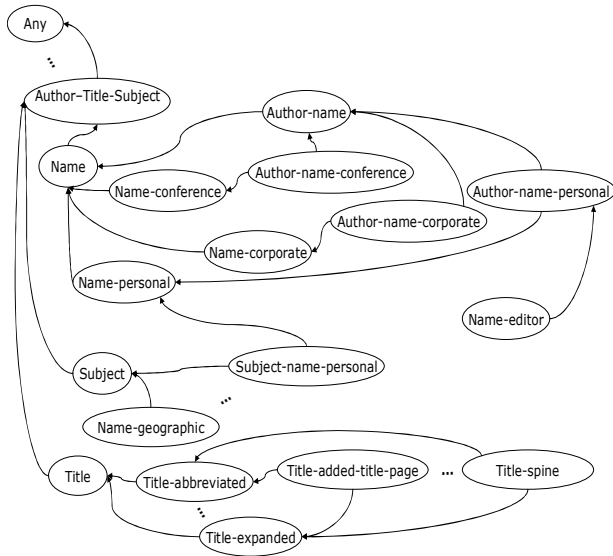

Figure 1. A representative simplified sample of the Semantics-Based RDFS graph of the Bib-1 Access Points

For the construction of the Bib-1 Access Points RDF Schema, we map the vertices of the graph to rdfs:Class and the arcs of the graph to rdfs:subClassOf properties. Figure 1 depicts a representative sample of the resulting RDFS graph. The complete version of the graph can be found at the address http://dlib.ionio.gr/lab/graph.htm. For readability reasons, we simplify the graph excluding the core RDFS classes and properties.

The similarity of the semantics of an Access Point with the others is expressed from their positions in the graph. The direct subclasses or supper classes of an Access Point have the nearest semantics with the class which represents the Access Point. The similarity among the semantics of the Access Points influences the result sets of the queries with the Access Points. Thus, when a query uses an Access Point near the top of the hierarchy, the answer contains more results than a query contains using an Access Point near the leafs of the hierarchy. Thereafter, an ancestor Access Point of another broadens the semantics of the Access Point and increases the resulting hits in the answers, while a descendant of an Access Point narrows the semantics of the Access Point and decreases the resulting hits in the answers.

4. zSAPN PRESENTATION
4.1 Access Point Substitution Policies
Exploring the Access Points similarity from the Bib-1 RDFS, zSAPN provides two alternative substitution policies for the unsupported Access Point. The formal description of the substitution methods and the algorithms are described in our previous work [8]. The first policy broadens the semantics of the unsupported Access Point and substitutes it with the Boolean *AND* combination of all the nearest supported ancestors of the Access Point, from every ancestor path hierarchy starting from the unsupported Access Point. The final set of Access Points from the broad substitution policy is the minimal set of supported Access Points having the nearest semantics to the unsupported Access Point, after an optimization step. This is the case when there are more than one ancestor path hierarchies containing a supported Access Point, while the selected Access Point from one path is also a member of another path at a higher level position than the selected AP from this path

In contrast, the second policy narrows the semantics of the unsupported Access Point and substitutes it with the Boolean *OR* combination of all the nearest supported descendants of the Access Point, from every descendant path hierarchy starting from the unsupported Access Point. Similarly, as for the broad substitution is, the final set of Access Points from the narrow substitution policy is the minimal set of supported Access Points having the nearest semantics to the unsupported Access Point, after an optimization step. This is the case when there are more than one descendant path hierarchies containing a supported Access Point, while the selected Access Point from one path is also a member of another path at a lower level position than the selected Access Point from this path.

4.2 System Architecture
Figure 2 depicts the overall process of the system we have implemented which can be accessed at http://dlib.ionio.gr/zSAPN. For every source, which does not support the Access Point, a new query is produced using the Bib-1 RDFS and its source description for the supported Access Points. The core of our system which implements the substitution algorithms is built on top of the ICS-Forth RDFSuite[2]. Initially, we import the Bib-1 RDFS into the database using the RSSDB module of the suite, and for the exploitation of the Bib-1 RDFS we use the RQL [4], a query language for RDF bases. For the Z39.50 services we use the PHP/YAZ toolkit from the IndexData[3].
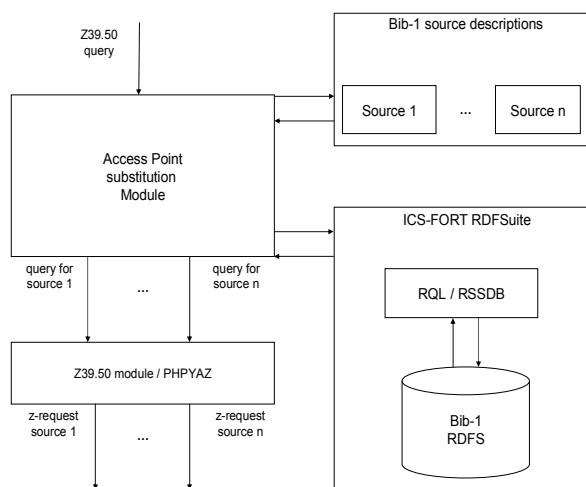
---

[2] http://athena.ics.forth.gr:9090/RDF/

[3] http://www.indexdata.com/phpyaz

Figure 2. zSAPN Architecture and the Process of a Z39.50 query

4.3 Querying the System

proceedings from the IEEE's conferences and only these. It is worth mentioning that we are not interested for other types of IEEE's editions such as technical reports, technical standards etc. Thereafter, we are selecting the Z39.50 sources of the *Library of Congress*, the *COPAC Academic & National Library Catalogue* (from United Kingdom) and the *University of Crete Library* (from Greece). The most appropriate Access Point for our query is the *Author-name-conference* according to the Bib-1 specification, which is rarely available from the majority of the known meta-search or local systems.

Table .1 depicts the summary of the search results of the example query. The *COPAC* source supports the query thus no substitution is required. The *Library of Congress* source does not support the Access Point[4] *Author-name-conference* and the system substitute the unsupported Access Point with the Boolean *AND* combination of the supported Access Points *Author-name* and *Name-conference*. Finally, for the source of the *University of Crete*, the zSAPN substitutes the unsupported Access Point with the supported *Author-name*. We observe the deferent substitutions for the sources of the *Library of Congress* and the *University of Crete*, influencing the differences into their sets of supported Access Points.

Before the comparison of the semantics similarity between the requested Access Point and the Access Points resulted from the

Figure 3. zSAPN first page for starting a session

The start page of the zSAPN system is appearing into the figure 3. The Current version of the zSAPN at http://dlib.ionio.gr/zSAPN gives access to some representative world wide most popular Z39.50 sources. For the composition of a query, a user has to select the sources, the search terms, the Access Point and the substitution policy in case when a source does not support the selected Access Point. As we see in figure 3, zSAPN offers also an option to broadcast the query directly to the selected sources as it is, without substitution, and to leave the sources to reply with their default behaviour.

For the clarification of the broad substitution policy let's consider the following example. We assume that we are searching for the

broad substitutions for the previous example, let's examine the semantics of the used Access Points, as depicted in figure 4. We observe that both the Access Points *Author-name* and *Name-conference* are direct ancestors of the Access Point *Author-name-conference*. Moreover, the semantic sets of data fields of the *Author-name-conference* Access Point is the intersection of the semantic sets of data fields of the *Author-name* and *Name-*

---

[4] According to the document "LC Z39.50/SRW/SRU Server Configuration Guidelines" at: http://www.loc.gov/z3950/lcserver.html

conference Access Points, thus the semantics of the *Author-name-conference* can be considered as being equivalent with the semantics of the Boolean AND combination of the Access Points *Author-name* and *Name-conference*. Finally, the semantic set of data fields of the Access Point Author-name consists of more fields than the *Author-name-conference* but less than the *Name* Access Point.

Table 1. zSAPN Broad substitution search results summary

Source	Hits	Query
COPAC	2542	Author-name-conference_1006 = IEEE *The Access Point is supported by the source.*Last
Library of Congress	1675	Author-name-conference_1006 = IEEE From the Access Point Substitution... --*The Minimal Set is:* Author-name_1003 Name-conference_3
University of Crete	340	Author-name-conference_1006 = IEEE From the Access Point Substitution... --*The Minimal Set is:* Author-name_1003

Concluding the comparison between the semantics of the zSAPN broad substitutions and the semantics of the requested access point, we observe that the substitution for the *Library of Congress* produces equivalent results with the requested Access Point and its answer has the same precision as the *COPAC's* answer which supports the Access Point. For the *University of Crete*, we receive an answer with similar semantics, but with less precision. In this way we manage to exclude records having as subject the conferences of the IEEE, although the answer still contains also other types of editions of the IEEE (e.g. technical standards, etc.).
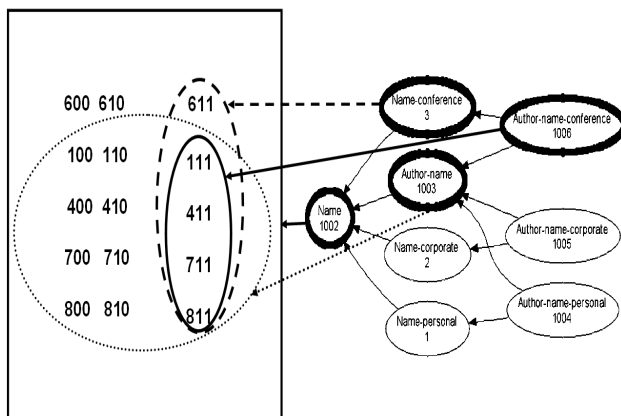

Figure 4. Comparison of the Access Points semantics

Ending the zSAPN presentation for the broad Access Point substitution, we compare the number of hits resulted from the answers using the zSAPN with the no substitution and with the broad substitution options. As we observer from the Table 2, for the *COPAC* it makes no difference due to the support of the Access Point from the source. The *Library of Congress* does not inform the user that the Access Point is not supported and does not also give any further information concerning the arbitrary substitution of the unsupported Access Point. Moreover, for the *Library of Congress*, we see a big difference in the number of matching records, about 5 times more hits approximately. We remind that the substitution for the *Library of Congress* produces equivalent semantics with the original request. For the *University of Crete*, the user receives a similar answer with less precision instead of a query failure.

Table 2. Resulting hits without substitution and with broad substitution

Source	No Substitution	Broad Substitution
COPAC	2542	2542
Library of Congress	8007	8007
University of Crete	Error: Unsupported attribute combination	340

The second substitution policy narrows the semantics of the requested Access Point. Let's consider again another example for the presentation of the narrow substitution policy. We assume that we are searching the *Library & Archives Canada* Z39.50 source, for all the metadata records containing the term *"Malinowski"* as either Author, or Subject, or in the Title. The most appropriate Access Point from the bib-1 Access Points set is the *Author-Title-Subject*, which is also rarely offered for use.

Table 3 depicts the search results summary from the *Library & Archives Canada* source. This version of the summary results table contains in the last column both the initial and the final optimized minimal set of Access Points for the substitution. We note that the source does not support the Access Point *Author-Title-Subject* and we select the narrow substitution policy. The zSAPN substitutes the unsupported Access Point *Author-Title-Subject* with the Boolean *OR* combination of the supported Access Points *Title*, *Subject* and *Author-name* thus avoids the query failure from the source.

As we already mentioned, for both the substitution policies the final set of Access Points for the substitution is the minimal set of supported Access Points having the nearest semantics to the unsupported Access Point. We observe in the table 3 that the initial set for the substation of the unsupported Access Point consists of the *Title*, *Subject*, *Author-name-corporate*, *Author-name-personal*, *Author-name* and *Author-name-conference*, while the minimal set consist of the Access Points *Title*, *Subject* and *Author-name*. Whereas the Access Points *Author-name-corporate*, *Author-name-personal* and *Author-name-conference* are subsets of the Access Point *Author-name*, see figure 2.1, their semantics are already included into the Access Point *Author-name*. Therefore, the optimization step excludes the Access Points *Author-name-corporate*, *Author-name-personal* and *Author-name-conference* from the minimal set of the narrow substitution policy.

In the previous example, the user receives an answer with similar results to the original query and avoids the query failure. We remind that the *Library and Archives Canada* source fails the query when it does not support a requested Access Point, such as the *Author-Title-Subject*. By this approach the user does not retrieve results while he knows that the Access Point is not supported. Thereafter, the user has two alternatives in order to search the *Library and Archives Canada* source: either to broadens the semantics using a more general Access Point like the *Any*, or to replace the Access Point with the union of other Access Points with narrower semantics, similar to what zSAPN would do. The first, the easiest one, will increase the results by approximately 57% due to receiving 30 additional non relevant matching records, while the second requires a deep knowledge of the semantic relationships of the Access Points and the available search services from the source.

Table 3. zSAPN Narrow substitution search results summary

Source	Hits	Query
Library and Archives Canada	52	Author-Title-Subject_1036 = Malinowski From the Access Point Substitution... *-The AP could be substituted with the AP(s):* Title_4 Subject_21 Author-name-corporate_1005 Author-name-personal_1004 Author-name_1003 Author-name-conference_1006 *--The Minimal Set is:* Title_4 Subject_21 Author-name_1003

5. CONCLUSIONS AND FUTURE RESEARCH

We found that semantics based substitution methods could really improve the effects from the unsupported Access Points and also to enhance the interoperability between the Z39.50 library search systems. The current implementation of the zSAPN system, improves the search consistency and lowers the query failures exploiting the semantic information of the Access Points from an RDFS description. Moreover, zSAPN substitutes the unsupported Access Point with a set of others whose proper combination either broadens or narrows the semantics of the unsupported Access Point.

The proposed substitution policies enable the meta-search system, or any other mediator, to decide how to modify, if it is necessary, the semantics of an unsupported query prior to initiating the search requests. In addition, a source using the zSAPN underlying methodology could expand its functionality instead of making arbitrary or general substitutions, as is the case of the Library of Congress source. Finally, the RDFS description of the Bib-1 Access Points could be a basis for the deployment of the library community search primitive semantics to the Semantic Web [2].

A number of interesting points are derived from this work for future development and research. First, we have to evaluate the implications of the proposed Access Point substitutions into the recall/precision of the initial queries. Also, another interesting point is to measure and describe the degree of the similarity of the semantics among the Access Points. Up to this point we have only studied the subset relationship between the Access Points. A number of questions remain to be answered like: Are there any substitutions that produce equivalent semantics to the original Access Point? How tight are the subset relationships and what parts of these semantics are used in practice? How is it possible to expand the current RDFS in order to contain all the needed subset semantics? Is it possible to automate the process of the Access Point substitution so that the system will be able to decide the best substitution policy under some criteria?

Another motivating study would be the application of our methodology and algorithms to the corresponding concepts of the Search and Retrieve Web Services (SRW) protocol (Sanderson 2004) and to measure the degree of improvement. An essential point for the deployment of the SRW would be the creation of a meta-RDF schema expressing the correlations of the underlying semantics contained in every context set defined in the protocol.

6. REFERENCES

[1] ANSI/NISO. 1995. Z39.50 Information Retrieval: application service definition and protocol specification: approved May 10, 1995.

[2] Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Semantic Web. In *Scientific American* (May 2001).

[3] Finnigan, S., Ward, N. *Z39.50 Made Simple*. http://archive.dstc.edu.au/DDU/projects/Z3950/zsimple.html.

[4] Karvounarakis, G., Alexaki S., Christophides, V., Plexousakis D., and Scholl, M. 2002. RQL: A Declarative Query Language for RDF. In *Proceedings of the 11th International World Wide Web Conference (WWW'02)*. Honolulu, Hawaii, USA 2002.

[5] Motro, A. 1987. Superviews: Virtual Integration of Multiple, Databases. In IEEE Transactions on Software Engineering, 13(7), 1987.

[6] Papakonstantinou Y., Gupta, A. Garcia-Molina, H. Ullman, J. 1995. A Query Translation Scheme for Rapid Implementation of Wrappers. *In Proceedings of the Conference on Deductive and Object Oriented Databases, DOOD-95*.

[7] Sfakakis, M., Kapidakis, S. 2006. A Semantics-Based Graph for the Bib-1 Access Points of the Z39.50 Protocol. *Proceedings. of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*. Alicante (Spain), 17-22 September. Lecture Notes in Computer Science 4172. Berlin-Berlin-Heidelberg: Springer. 445 – 448.

[8] Sfakakis, M., Kapidakis, S. 2007. A Semantics-Based Substitution for Unsupported Z39.50 Bib-1 Access Points. *Proceedings. of the 2nd International Conference on Digital Information Managemant (ICDIM07)*. Lyon (France), 28-31 October.

[9]   Sanderson, R. 2004 *A Gentle Introduction to SRW*. www.loc.gov/z3950/agency/zing/srw/introduction.html.

[10] Ullman, J. 1997. Information Integration Using Local Views. *LNCS, Proceedings of the 6th International Conference on Database Theory, pages 19-40, 1997*.

[11] W3C. 2000. *Resource Description Framework Schema Specification 1.0*. W3C Candidate Recommendation 27 March 2000. http://www.w3.org/TR/2000/CR-rdf-schema-20000327/.

[12] ZIG. 1995. *Attribute Set BIB-1 (Z39.50-1995): Semantics*. http://www.loc.gov/z3950/agency/bib1.html.