Using Semantic Searching for Web Portal Interoperability

Francisco Pinto[*], Cláudio Baptista[†], Nick Ryan[‡]
<fqp1@ukc.ac.uk>, <baptista@dsc.ufpb.br>, <N.S.Ryan@ukc.ac.uk>

Abstract

The Web can be viewed as an open heterogeneous information space indexed by search engines and directories based on keywords. The indexing techniques fall short of providing satisfactory efficiency/effectiveness and recall/precision parameters which are required by information search and retrieval systems. The use of metadata for resource discovery would improve these search and retrieval parameters. However, the incompatibility of formats for exchanging metadata, metadata spam or even lack of widespread knowledge about metadata has inhibited this evolution. To improve these parameters and obtain a higher level of performance for searching and retrieving over the Web, interoperability and trust-ability are some of the fundamental factors to be improved. To solve these problems some projects try to create an infrastructure for interoperability based on metadata to give more semantics and organization to the Web. However, there is almost no agreement about the metadata element sets to be used, they do not catalogue all the information and they do not enter into the places where the information really lives, the databases. This paper raises issues of semantics based on metadata, ontologies and Z39.50. Additionally, it presents an architecture for an interoperable and distributed infrastructure over Thematic Portals, providing searching and retrieval services from the Web.

1 Introduction

The Web can be viewed as an heterogeneous information space is composed by resources from different domains, structured in different formats and distributed over several different software and hardware platforms. This information space is very dynamic, locally placed in a disperse geographical distribution and managed/organized by individual Web publishers. The information resources are mainly available as static and dynamic HTML documents and its inter-relationships are obtained by hyperlinks forming the Web. These features are perhaps some of the Web strengths, but they raise problems including how to search this information space efficiently, and how to retrieve this information coherently.

Some approaches to solving these problems have been implemented by current search engines (*altavista*[1], *google*[2]) and directories (*yahoo*[3], *open directory*[4]). Search engines index potential keywords in a very large database. This approach can be efficient due to the huge processing power involved, but it is not effective as most of the hits obtained in a search are poorly related to what the user is looking for (*altavista*). The reason for this low effectiveness is related with the absence of semantics in the indexing. Even using sophisticated methods such as to index just the citations associated with links, instead of the entire text

[*]The Computing Laboratory/University of Kent at Canterbury/United Kingdom, sponsored by the Portuguese Fundação para a Ciência e a Tecnologia.
[†]Departamento de Sistemas e Computação/Universidade Federal da Paraiba/Brazil.
[‡]The Computing Laboratory/University of Kent at Canterbury/United Kingdom.

and ranking the documents by number of citations, the effectiveness is still poor as there are no meaning and context associated with the resources (*google*).

The approach carried out by directories seems to be more effective and more efficient when compared to that of search engines. However, its effectiveness is far from fulfilling the user requirements due to the small quantity of resources indexed. In the case of *yahoo* as the information is registered and organized manually following a given ontology in a centralized database, problems such as scalability can arise. In contrast with *yahoo*, the *open directory* follows a distributed approach. There is an hierarchy of subjects following a given ontology and maintained by a community of editors responsible for the subjects who evaluate and include the sites submitted. The Web sites are cataloged by their publishers using the subjects adopted by the ontology in RDF files. These RDF files support metadata elements organized as structure and content files, where the first support the hierarchy and the second describe the information. Descriptive metadata elements such as *Title* and *Description* and relation elements as *Link* are used for this end. This meta-information is then available for internal navigation from the *open directory* web site or externally for use by partners. These are the search engines such *altavista* and *google* which use the directory as a starting point to index the Web documents under the site published. However, they just index static documents and the problem of access to dynamic documents based on databases or based on other information spaces still persists.

An ideal Web would be like a conventional database in which it would be possible to have semantics associated with every static or dynamic resource located in every information space, through their schemas. Applications such as search engines and directories could use these for indexing and provide a better service to their users. Even in this case agreed semantics would have to be used to achieve interoperability. These semantics could be achieved through agreed metadata elements sets which give meaning and context to resources. However, this would bring more problems as on the Web there are no constraint rules and access control as in DBMS to restrict who and what can be done, or simpler, what has to be done just to publish a document and make it *visible* from outside. On the one hand, most of the publishers would not provide metadata as this means additional work and knowledge, and on the other hand, there would be Web publishers producing too much metadata (metadata spam) to improve the chances of their resources appearing at the top of any search result. Additionally, it would be very difficult or even impossible in some cases to generate dynamically metadata for dynamic contents.

The Web needs an infrastructure where all resources could be accessed virtually in the same way, and interoperability supported by metadata is the key factor for it. Several technologies including mediators and federated architectures to provide an infrastructure for interoperability have been studied and applied in related projects[5]. In addition, with the evolution of metadata, standards for specific domains and Cross Domain (XD) have been proposed[6]. Recent initiatives such as the Universal Description, Discovery and Integration (UDDI)[7] and the Open Archives Initiative (OAI)[8] are proposed as solutions for these needs. They specify a central authority where the publishers can register their information and services. UDDI uses XML and SOAP, while OAI use well defined XML and Dublin Core metadata as the mechanisms to provide interoperability. In this way, the existing and new services will always have to use this infrastructure as the entry point for any harvesting of information.

Metadata seems to be the common denominator for all the solutions. To overcome some of the problems associated with metadata, there are closed and trusted communities adopting standard metadata element sets suitable for their information domain needs[9]. Additionally, more generic metadata element sets, have been developed to provide the Web with a framework for XD searching[10].

This paper discusses the problem of using metadata for Web resource discovery and

presents an architecture for an infrastructure to provide interoperability using trusted Portals. The remainder of this paper is structured as follows. Section 2 discusses an *Infrastructure for Interoperability*. Section 3 presents the *Implementation* of such an infrastructure based on Thematic Portals. Section 4 presents an *Application Portal* based on a *Historic Environment* domain information. Finally, section 5 concludes with a *Summary* of the paper and discusses further work.

2 An Infrastructure for Interoperability

The Web lacks an infrastructure for *Interoperability*. With one, it would be possible to virtually integrate the existing information spaces in a normalized and linear structure. The approach followed by the *open directory* seems to be better designed than the one provided by *yahoo* as it de-centralize the authority. However, it does not solve the problem of access to dynamic Web documents generated from other information sources apart from the Web space. The infrastructure offered by either UDDI or OAI creates the needed mechanism to provide interoperability and achieve access to these information sources. However, it is centralized which can have problems in terms of scalability not only on the registering process, but also on performing the work it proposes to do. A distributed infrastructure similar to the one provided by the *X.500 Directory*[11], which de-centralizes the authority control, but without its overhead of management and functionality would be a better solution for the Web needs.

This infrastructure could be achieved by Searching Portals using *semantic access points* based on metadata, for searching with more precision the resources associated with the potential sources of information. These access points would be accessible transparently from the Web by users or applications through these Portals which have a set of associated targets containing access to simple applications indexing resources on file systems, to front-end applications for heterogeneous database management systems or even to Web server applications indexing the documents, all acting as digital libraries for their specific resources.

2.1 Technologies and Standards

Some technologies and standards are potential tools with a fundamental role in providing an infrastructure for interoperability offering uniform access to information in an heterogeneous and dynamic environment. Z39.50 is an ISO standard defining application services and a protocol specification for search and retrieval of information in distributed environments. Z39.50 specifies access points available on Z39.50 targets with semantics defined by information domain profiles. These profiles are agreed pre-defined rules of how to achieve interoperability by adopting well defined attribute sets for searching, and abstract schemas and record syntaxes for retrieving the records found in a search[12]. Dublin Core (DC) is a *de facto* metadata standard intended for resource description. DC has been applied to several domains such as libraries, museums, geo-spatial projects, etc. It defines a simple metadata element set which consists of 15 generic elements with well defined semantics and common to all information domains, with potential to be used for XD searching[10]. XML is a standard developed by W3C which can provide a framework to annotate the information with tags and to exchange information in an independent way. XML is a potential technology to support interoperability, as its tags can be the metadata elements to structure the resources. If different metadata element sets are used, XML can support a framework such as RDF to enrich the resource description and provide interoperability for applications. Ontology mappings to metadata, thesaurus and gazetteers can be represented using RDF and exchanged using XML. Finally, XML is one of the registered Z39.50 record syntaxes, and in this way can be used the exchange and presentation of metadata records when they are retrieved, using technologies such as a

publishing framework[13, 14]. Bath Profile is a set of rules to be adopted by information providers to obtain conformance in Z39.50 targets. It specifies the mandatory access points and how they should be set up. Bath profile requests DC, XML and RDF as the elected standards to provide interoperability[15].

2.2 Searching Portals

Each Portal is supported by technologies and standards mentioned in section 2.1 to provide access to heterogeneous information from the Web. The components of a possible Portal implementation are a *Web Interface*, the Web face of the Portal, consisting of a set of parameters. These parameters can be used by applications (Web Robots, Applets or Web Forms) through GET or POST HTTP requests; a *Web/Z Gateway*, which is an implementation of a Web Application, an HTTP and a Z39.50 Client as Servlets; and a *Z39.50 Target*, an implementation of a Z39.50 Server and respective specific information profiles. These components form the Portal architecture, as shown in Figure 1.

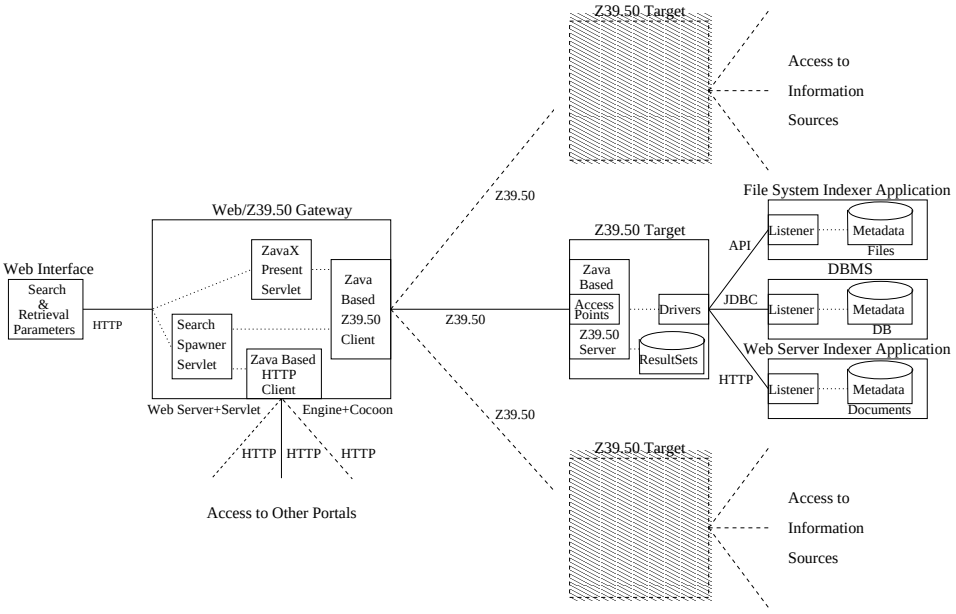

Figure 1: Portal Architecture.

2.3 Access Points

Access points have well defined semantics created by trusted communities of specific information domains with knowledge and interest in having that information accessible for specific and XD searching. Searching Portals could then use these target access points to access the virtually integrated information. Each target maps the accessible access points to the real fields at the information sources available under its management. Therefore, a target is responsible for converting the Z39.50 queries to native queries (SQL) for searching and converting the native record syntax (text table) to one of the Z39.50 record syntaxes (SUTRS, GRS1, XML) for retrieving.

Figure 2 shows examples of generic access points suitable for XD searching, including *Title*, *Subject*, *Author* and *Any* or the more abstract *Who*, *What*, *When* and *Where*, giving generic meaning about the resources the users are looking for. Specific information spaces such as those used in Environment Information Systems, Geographical Information Systems or E-Business could have more specific access points such as *Date*, *Temperature*, *Referencing System*, *X-Coordinate*, *Y-Coordinate*, *Product* and *Price*, *Etc.*

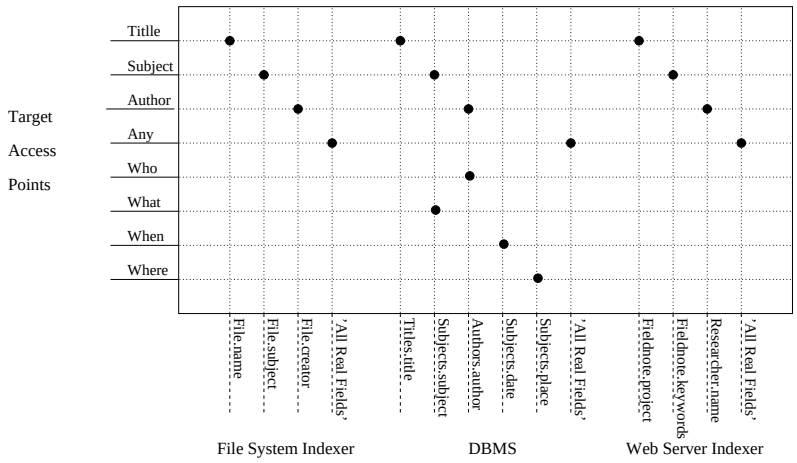

Figure 2: Access Point Mappings.

2.4 Directory of Portals

In order to create the whole infrastructure, an hierarchy of Portals organized by *themes* can be mounted where the root can be viewed as a Web front-end for a set of federated digital libraries, as shown in Figure 3. This hierarchy follows a given ontology, organized by theme, forming a directory of Portals. Portals have their own target(s) and can access any target belonging to their Portal children. At the bottom, leaf Portals have the targets with specific information sources. The non-leaf Portals are more generic. However, they behave in the same way as the leaf Portals, as their targets provide infrastructure information such as the conceptual coverage of their children.

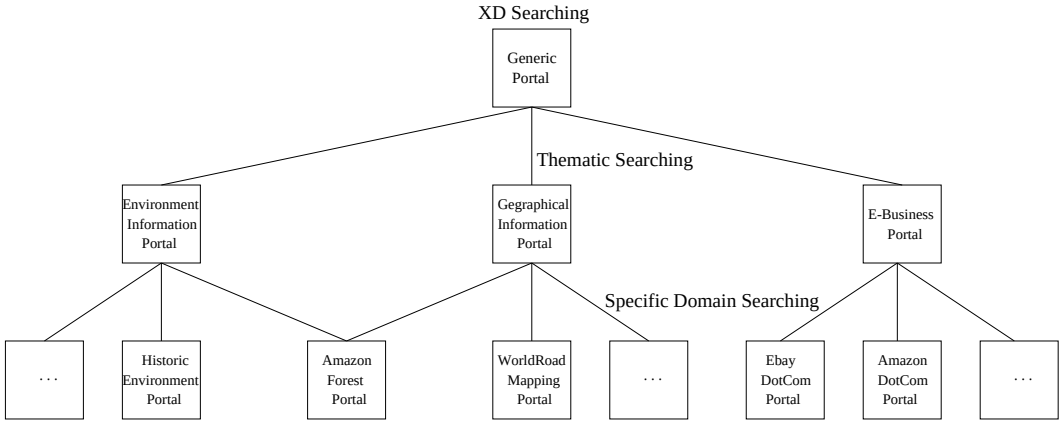

Figure 3: Portal Organization.

At higher levels there are Portals with generic access points for XD searching and concepts for browsing through the directory to select Portals by information domain over which searches will be launched. To provide compatibility between Portals each target belonging to a leaf Portal should have access points for XD searching and domain specific access points. These Portals are the most abstract and just support XD searching over all the Portals. However, having more information the user can browse concepts available at lower levels. These concepts are obtained with ontologies describing the lower level Portals. This information is implemented with metadata and available under the child Portals in their Z39.50 targets. Portals have to be created by an entity which manages and specifies the rules for the Directory. One of these tasks is to register Portals and add them to the directory connecting them to their parent(s) in a position which best covers the concepts they provide.

2.4.1 Access Methods

Any Portal in the directory can be directly accessed via HTTP and can perform Z39.50 accesses to its targets and to the targets belonging to its child Portals. To perform the Z39.50 accesses, the Portal has to know which targets to use. By default all the targets associated with the child Portals are chosen. If the user browses by concepts, more specialized Portals can be selected, and the same principle is recursively applied.

Each Portal has a set of associated concepts. They catalog the Portal with the information they manage (Environment, Geographic, E-Business, Historical). These concepts are placed in a given target and can be accessible through HTTP (from RDF files) or through Z39.50 (via Explain or via a common database). For browsing, a cascade of HTTP requests is triggered over child Portals to obtain the concepts they provide. In turn, each Portal triggers HTTP or Z39.50 requests over their targets to obtain the available concepts. Knowing the concepts, at each level the user can select the child Portals over which the searches will be launched or choose to go down to the next more specialized Portal. If neither of the options is chosen, the search will be launched over all child Portals.

2.4.2 Directory Usage

Starting at the root Portal, the user can launch XD searches over all the possible targets associated with all Portals. Alternatively, the user could browse concepts driving the user through the Portals with the information they want. Having selected the Portals, a domain specific Portal or group of Portals could be reached, narrowing the information space for searching and enhancing the system scalability.

For instance, at the higher level, the user browsing the available Portals could choose the Historic Environment (HE) concept, restricting the information domain. At this point, more related information objects could be presented, such as semantically related access points (spatial reference systems and coordinates), *gazetteers* (continent/countries, time period diagrams) or *click-able maps* for each of the available geographical zones. The user could then launch searches in parallel over one or more HE targets associated with this Portal. Finally, the resources could be retrieved as soon as the results arrive from the different targets.

3 Implementation

In this section a proposed solution is discussed by presenting the architecture of an infrastructure for interoperability and highlighting implementation issues based on the application that has been developed.

Two packages were developed to support this infrastructure, Zava and ZavaX. Zava is a Z39.50 client/server Java API offering basic Z39.50 features, such as Init, Search, Present and Close. Internally, Zava has a XML parser and an RDF processor to implement the profiles and to exchange the resources. ZavaX is based on a Publishing Framework and on Zava to provide the presentation level of Z39.50. It retrieves the records from the Z39.50 targets and converts them to XML. Using XSL, they can be transformed to the format that best suits the Web client. In Figure 1, three components are essential for this implementation: a Web Interface; a Web/Z39.50 Gateway; and Z39.50 Targets. The Web Interface is used by Web applications, which access to the target using the Web/Z39.50 Gateway. It can also be used by Web Robots to feed their databases if they use the available access points. The Web/Z39.50 Gateway is a Web application located on the Web Server as a Servlet. It is based on Zava and ZavaX packages in order to inherit all the implemented technologies and provide the required functionality. This launches threaded searches over the Z39.50 targets and presents the records found in each target to a Web browser. Additionally, it can perform HTTP requests over the

child Portals to obtain their concepts. A Z39.50 Target is any implementation of a Z39.50 server and the profiles providing access points to a given information domain. For local use a Z39.50 target was developed using Zava to implement a basic stand alone Z39.50 server. This searches a DC conformant database via JDBC. Additionally, it uses XML as a platform to describe the profiles in RDF and to exchange the result sets between the server and the client, using record syntaxes such as simple text to display the records, or structured tagged records for later processing. Each Z39.50 server is responsible for searching and retrieving according to the mappings implemented. At search time, it maps the access points to real queries on the databases. At retrieval time, the abstract schemas are used with another set of mapped queries to build the records and exchange them with the client using a negotiated record syntax with very precise rules.

Although Zava and ZavaX involve technologies usually not present on Z39.50 targets, the Web/Z39.50 gateway is totally interoperable with any other Z39.50 target. The only requirement for the targets is to implement the same profiles as the Web/Z39.50 gateway. However, even if the profiles are different, but some access points are the same, basic search and presenting can be done with costs in terms of interoperability. Finally, the records obtained from the targets are always converted to XML for later processing by the publishing framework, even if the targets does not implement the XML record syntax. For the presentation a publishing framework is used to transform the XML records on the fly to the desired format (HTML, WML, PDF), and according to the capabilities of the Web clients, show different selected information (Browser in a PC using HTTP or in a Mobile Phone using WAP).

4 Application: An Historic Environment Portal

In order to validate the ideas mentioned earlier, an HE Portal has been developed. This Portal is supported by a group of Z39.50 targets which provide access to databases with more than 400,000 indexed records of spatio-temporal DC conformant metadata[16]. This Portal aims to provide the users with a better tool to find resources either by the usual access points or by more specialized ones such as spatial coordinates. These searches could have be done directly at the native databases with better performance, however not every user would know the schema and the native language of the database. Four targets are associated with this Portal, each one with HE information stored in a database. The Portal launches the searches over these targets. Each one has a different DBMS and thus different access methods to the databases (JDBC, ODBC, Private API). Having the same access points for the different native schemas, one mapping is needed for each target (Figure 2).

For XD searching, generic access points such as *Title*, *Subject*, *Author* or even *Any* (for full-searching) are being used. As the HE emphasis is on spatial and temporal searching there are also *Where* and *When* access points. This is achieved using the DC *Coverage* element and feeding it with temporal and spatial descriptors based on thesauri of places and time period terms. Other access points include: *Who* to search by creator or publisher; and *What* to search by subject keywords. Additionally, coordinate access points based on three different spatial referencing systems, OSGB (Ordnance Survey of Great Britain), OSIRL (Ordnance Survey of Ireland) and LL (Longitude and Latitude), can be combined with all the other access points in a complex boolean expression to perform a precise, effective and efficient searching.

A possible query based on the generic access points on this Portal could be achieved taking a boolean option to search by *Title=king*, and *Subject=fort* at the selected targets. A more specialized query could be based on other access points such *Who=royal*, *What=castle*, *When=medieval* and *Where=scotland*. All these examples can always be complemented with queries on coordinate access points based on the OSGB, OSI or LL referencing systems.

5 Summary

This paper covered the technology, functionality, conformance and research potential of semantic and interoperable searching Portals. The importance of interoperability, technologies and standards that have been used were discussed. Based on this, an HE Portal was implemented. The main consideration to take from this Portal is that by using agreed semantics through metadata it allows the users to find efficiently and effectively specialized information. Using Z39.50 as a leverage for interoperability and the Web for accessibility a high level of resources are available to their users. Reasons for good performance in terms of search and retrieval parameters in this Portal can be related with several factors such as the use of: Information resources and indexes located on their native places; DC metadata for indexing the resources and provide XD and domain specific searching; Z39.50 to specify searching and retrieving remotely on the requested resources; JDBC or other open or private DBMS access methods to locally fetch and build the resources found in a search; RDF to express the mappings and profiles; and XML to exchange the resources.

The HE Portal is based on specialized domain information, in a possible directory of Portals. As further work it would be interesting to study how such a system would perform in terms of a large number of hierarchical Portals from different information domains forming a real infrastructure for interoperability. This study could cope with the search and retrieval parameters maintaining the system scalability and interoperability. We believe this infrastructure could give good results based on the following principles: Directory of Portals as a distributed directory; Generic to specific levels of abstraction; and Searching and Browsing paradigms for narrowing the information space.

References

[1] AltaVista. AltaVista Search Engine. *http://www.altavista.com.*

[2] Google. Google Search Engine. *http://www.google.com.*

[3] Yahoo. Yahoo Directory. *http://www.yahoo.com.*

[4] Open Directory. Open Directory. *http://www.dmoz.org.*

[5] Michael Lesk. Practical Digital Libraries: Books, Bytes and Bucks. *Morgan Kaufmann*, 1997.

[6] Amit Sheth and Wolfgang Klas. *Multimedia Data Management, Using Metadata to Integrate and Apply Digital Media.* McGraw Hill, 1998.

[7] Ariba Inc. Universal Description, Discovery and Integration. 2000.

[8] Carl Lagoze, Hussein Suleman and Herbert Sompel. The Open Archives Initiative. 2001.

[9] William Moen. The CIMI Profile: A Z39.50 Profile for Cultural Heritage Information. 1998.

[10] Stuart Weibel and Eric Miller. Dublin Core Metadata. 1997.

[11] David Goodman and Colin Robbins. Understanding LDAP and X.500. 1997.

[12] ANSI/NISO. Z39.50-1995/ISO 23950:1998 Information Retrieval: Application Service Definition and Protocol Specification, 1995.

[13] Tim Bray, Jean Paoli and C. M. McQueen. Extensible Markup Language (XML) 1.0. 1998.

[14] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. 1998.

[15] Carrol Lunau, Paul Miller, William E. Moen. The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery. 2001.

[16] Tony Austin, Francisco Pinto, Julian Richards and Nick Ryan. Joined up writing: an Internet portal for research into the Historic Environment. *Forthcoming paper in CAA 2001: Proceedings of Computer Applications and Quantitive Methods in Archeology, ed. G. Burenhult*, 2001.