Using Euler Diagrams in Traditional Library Environments

Jérôme Thièvre, Marie-Luce Viaud [1]

INA,
4, avenue de l'Europe
94366 BRY SUR MARNE France

Anne Verroust-Blondet [2]

INRIA Rocquencourt,
B.P. 105,
78153 LE CHESNAY Cedex France

Abstract

In this paper, we present a new graphical interface for traditional library environments, which allows the user to elaborate easily and efficiently new strategies in search processes. This tool is based on two linked interactive Euler diagram representations. The first one is an interactive representation of the structures composing the documentary kernel of the library. The user may navigate and select items, making their own understanding of the database content, structure and access. The second one is a set based visualization of the results of a composed query. This allows the user to validate his search context and to elaborate strategies to go through the results. The association of both interfaces generates a tool that allows the user to elaborate the main search strategies through graphical manipulations.

Key words: Euler diagrams, User Interface, digital libraries

1 Introduction

The role of the "Institut National de l'Audiovisuel" is to manage the preservation of the French audiovisual heritage, to assure its exploitation and to make it more readily available. Moreover, under the terms of the French law of 20 June 1992 relating to legal deposits and the decree for application of 31 December 1993, Inathèque de France is responsible for collecting and conserving radio and television broadcast sound archive and audiovisual documents, taking part in drawing up and distributing national bibliographies and making such documents available to the public for the purposes of research.

[1] Email: jthievre@ina.fr, mlviaud@ina.fr
[2] Email: Anne.Verroust@inria.fr

This is a preliminary version. The final version will be published in Electronic Notes in Theoretical Computer Science
URL: www.elsevier.nl/locate/entcs

To achieve this mission, the INA's information officers index the audiovisual documents received daily. For every received document, this phase of indexing consists in generating the most complete and the least ambiguous textual description possible. Documentary or indexing tools assure the coherence of these descriptions according to a set of authoritative classifications called the documentary kernel. This kernel has been elaborated since the beginning of this database in 1975 and is in constant evolution. It contains several thesauri. A thesaurus is a graph of terms allowed to describe the set of documents and can be viewed as a hierarchical set. For example, the thesaurus of common nouns contains approximately 10000 terms. The documentary kernel constitutes the "added value" of the documentary process. Indeed, it allows controlling the terms of description in order to minimize the "noise" during the search process.
Searching in such a context differs from web searches in the fact that the client usually knows what kind of video sequences he wants. The result is usually a target and the task of the researcher will consist in finding the tiny path in the documentary kernel that leads to a small set of pertinent results. In fact, such an approach usually requires much practice from the user to be effective.
In 2003 INA began providing an online Internet access to part of its database for a selected panel of clients. At this time the documentary kernel is not accessible, making the search process complex for the user.
In order to make archives more available for the general public, it is necessary to elaborate new navigation and search user interfaces. These interfaces must be well suited to future users who have a limited knowledge of the documentary techniques and associated tools. The documentary kernel is composed of large structures such as graphs, trees and lists. A query is composed of a list of terms extracted from the kernel, a boolean expression and a list of results. At this time, there is no unified interface to visualize such objects. This paper is an attempt to propose a unified interface in such a context.
This paper describes a search and navigation user interface based on Euler diagram representations. The Euler diagrams constitute an interactive, unified and intuitive representation that allows users to perform graphically complex queries. The interface is composed of:

- A zoomable representation of the objects of the documentary kernel used to explore and select the efficient semantic fields associated with the database of documents.
- An interactive cartography of the results according to the pre-selected fields.

2 Authorities' explorer

Thesauri are the most complex objects of the traditional documentary kernel. A thesaurus is used in two processes. During the indexing process [7,8,15] of the documents, the thesaurus defines the terms that can be used to describe a document. It is also used during the tasks of document retrieval to perform more precise searches than can be achieved in full text mode.
A thesaurus is an oriented graph in which each node describes a term, its meaning,

its synonyms and its links to other close terms (”see also” relation). Each term can have several children terms. This parent-child relation can be interpreted in several different ways:

- The child term is more specific (hyponym),
- The child term is a facet of the parent term,
- The child term is an element of the parent term.

The structure of the thesaurus is usually viewed as an arborescence from the most generic terms to the most specific ones. It offers an interesting starting point for a semantic navigation in the database. The user must be able to navigate from term to term or look for a particular term and therefore reach the corresponding documents. Dynamic transversal paths associated with the ”see also” relation complete the semantic exploration functionalities.

2.1 Trees and Euler Diagrams

Euler diagrams are a natural and intuitive way to visualize sets [14]. If we consider each tree node as a set that contains its children nodes then it becomes easy to represent a tree as an Euler diagram. Figure 1 shows the same tree visualized as a node-link diagram or as an Euler diagram.

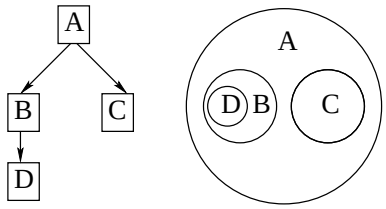

Fig. 1. node-link diagram and Euler diagram.

2.2 Visual Attributes

Visual attributes, as defined by Bertin [3], are the various characteristics of a graphic object. The main attributes are the position in the plane, the shape, the size and the colour. In our system the position and the shape of nodes are automatically fixed. The position of a node means its inclusion in its parent, and all the nodes are represented by circles to preserve the traditional Euler circles representation. We decided to use the sizes of nodes (here the diameter) to code the number of indexed documents or the number of descendants of the term. Colour is used to differentiate the nodes of the same sibling. The presence of a ”see also” link is shown by a pictogram.

2.3 Nodes Layout

The algorithm lays out the node in a breadth first traversal of the tree. The root is arbitrarily placed, and then we calculate the scale factor and the positions of all its children. The scale factor of a node is proportional to the number of its children.

The centers of the children nodes are positioned on a circle and are such that the subcircle distance from one another is constant. The result of the layout for the term ”transport” is shown on Figure 2.

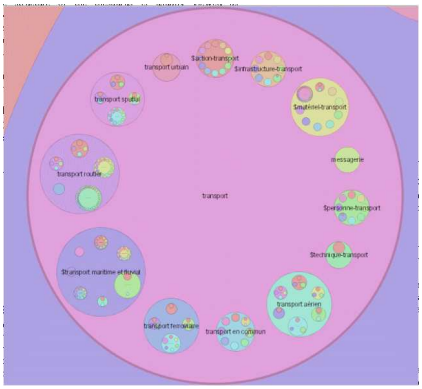

Fig. 2. the term ”transport” and its descendants.

2.4 Navigation, Search Text Field and Selection

Navigation is based on focus change and zoom [13,2]. The user can give the focus to any visible node by clicking on it. Then the view scales until it shows the focus node in detail. When the focus changes from a node A to a node B, an animation decomposes the minimal path from A to B: the view zooms out to the smallest common ancestor (if it exists) then zooms in towards B.
A text fi eld allows users to perform textual searches on the name and synonyms of terms. Nodes that do not correspond to the query are then greyed. Hit nodes are also displayed in a list. During the navigation, terms may be selected and used to formulate a search query.

2.5 Discussion

The navigation through the hierarchical links and the ”see also” transverse links allow a semantic exploration of the documents database. The functionality of textual search enables users to quickly locate interesting nodes.
The coding of the number of documents by the size of nodes allows a quick tracking down of the most important terms, and thus gives a kind of map of the database. Furthermore, full-text search is available and may be combined with any predefi ned terms selection.

3 Document searcher

The DocumentSearcher is an interface which allows the user to visualize the results of n queries (n from 1 to 5) in parallel. Each of those queries is associated with a contour and will be called a contour query in the following. A contour query can be a term or a boolean query. Our goal is to provide the user a set based cartographic representation of the database content with respect to those contour queries.

3.1 *Venn diagrams*

A Venn diagram [14] will contain all possible intersection zones between the n sets, even if a zone does not contain any document. To make the representation more readable we draw only diagrams having an elliptic curve for each contour. In this case, the maximum drawable number of contour queries is 5, thanks to [14].
More precisely, our Venn diagram representation is such that (cf. Figure 3):

- A unique Venn diagram corresponds to a given n. This diagram is such that the zones are ordered from the center to the border according to the number of contour queries defi ning it (the central zone corresponds to the set of documents indexed by the n contour queries).
- The colours of the zones correspond to the number of documents associated to them.

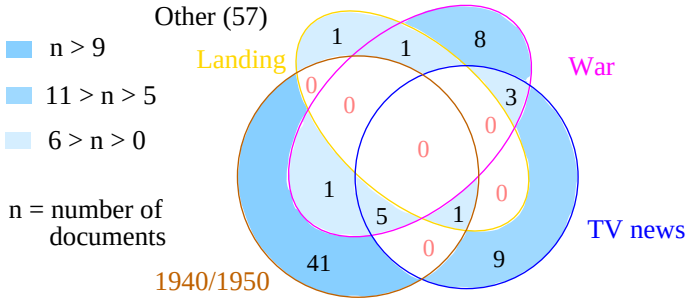

Fig. 3. a Venn diagram on 4 terms.

3.2 *Extended Euler diagrams*

According to [16], a planar extended Euler diagram exists for $n < 9$. The notion of "extended Euler diagram" is less restrictive than the concrete Euler diagrams defi ned in [9]: in a planar extended Euler diagram, sets are represented by connected regions of the plane that may have holes and each non empty intersection of a subset of the n sets corresponds to a unique zone of the plane. Note that Euler diagrams having a region of common intersection are always planar as indicated in [4].
We use extended Euler diagrams to generate a cartography which eliminates empty zones. In our context, many exclusive indexes have been generated to classify the documents with a minimum of ambiguity. As the region corresponding to the intersection of all terms is often empty, the corresponding diagram can not be generated using Chow and Ruskey's method [4]. However in those cases, extended Euler diagrams are interesting because they provide representations containing many fewer zones than Venn diagrams. We expect that these representations will be more readable (cf. Figures 3 and 4).
We are currently implementing the extended Euler diagram representation.

3.3 *Discussion*

The use of Venn or Euler diagrams simplifi es boolean formulations: the formulation of a query is done interactively by selecting the zones of interest [12]. An

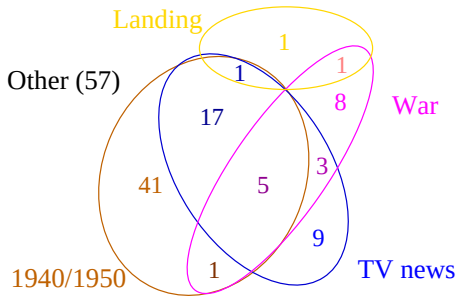

Fig. 4. an extended Euler diagram.

interesting property is that each document belongs to a unique zone on this cartography.
This interactive tool may also lead to new search strategies:

- Any selection of zones may be used to be the set to perform a new search, allowing easy progressive formulation;
- This representation gives a meta classifi cation, which allows the user to elaborate a strategy to go through the results.
- This cartography is a local map of the database according to the set of contour queries and provides a tool for the user to validate the pertinence of the terms used in the contour queries.

4 Analysis of search strategies

While analyzing a set of about 100 real queries, several types of situations come out, leading to different search strategies:

- The user gets too many documents. He reduces the set of result by choosing children terms of any term.
- The user gets too many documents. He performs an iterative query while formulating a new query on the set of results. We observe that users do not usually use more than 5 queries in parallel and that iterative formulation is the most current strategy.
- The user knows exactly the video sequence he is looking for because he has already seen it. He chooses semantic terms to formulate his queries but doesn't get the desired sequence in the results. He generalizes the chosen terms of the queries while choosing a set of documents corresponding to the parent term in the thesaurus or several sibling nodes.
- The user explores the database on a subject. He formulates composed queries and tries to elaborate a road map of classifi ed documents on his subject.
- The user proceeds to the analysis of the database on a subject. The resulting classifi cation, volumes and correlations on specifi c and chosen terms will constitute the new set of data for the researcher.
- The user has a description of the desired sequence but the choice of pertinent terms is diffi cult because an extra knowledge is needed. For example, one of

the studied queries was formulated as follows: ”President Chirac with a helmet”. In this case, as the term helmet” does not exist in the thesaurus, the user has to guess the contexts where President Chirac would wear a helmet (an official visit of a factory or a construction site).

For many of those tasks, our graphic tool provides easy graphical manipulation. Indeed, extending or focusing the meaning of terms according to the documentary kernel can be done by selecting parent, sibling or children nodes in the Authority Explorer. The distribution of the results in the document searcher allows the user to identify and modify quickly an inappropriate term (an inappropriate term returns too many or not enough documents). Iterative strategies can be performed easily just by selecting zones of the diagram in the Document Searcher. The graphic attributes allow the user to analyze and perceive quickly the distribution of the documents relative to the documentary kernel or the current query. However, if terms must be extrapolated, the knowledge and strategy of the user stays the main factor of a quick success.

5 Scenarios

The following scenarios illustrate the utility of our tool through several problematic cases.

5.1 Easy Boolean Query Formulation

Some studies [12,5,1] show that textual formulation of complex boolean queries may be difficult to elaborate. The main problems are:

- Most of users don’t know the concept of operators precedence, in consequence the use of parenthesis is a great source of mistakes ;
- The logical ”OR” and the coordinating ”OR” have different semantics, the second can be understood as a logical ”XOR” (exclusive), these problems may lead to misinterpretations of some queries that can be really frustrating.

The Venn diagram tool of the DocumentSearcher can be used to compose graphically any complex boolean query in order to produce syntactically valid queries. Moreover, this process of set and subset interactive selection avoids most of the semantics misunderstanding as shown in [12,11,5,10,17].

Scenario: Formulation of a complex boolean query

The task is to find the query for the retrieval of documents about music and song of the 1950’s. The textual boolean query is formulated as follows: ”(music OR song) AND period:$[1950 - 1960]$. The most common error in this example is to forget the brackets. In this case, the meaning and the results of the query will be different. The formulation of this query can be simplified with the interactive Venn diagram tool (cf. Figure 5), because the user has to select within the time period contour the zones that belong to music or song.

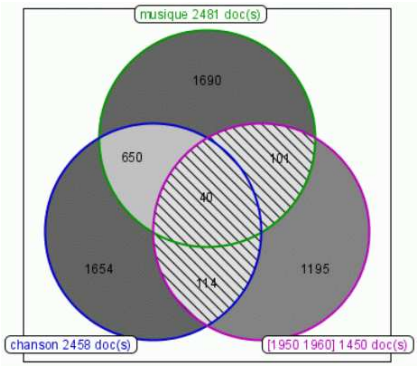

Fig. 5. Venn diagram on “music”, “song” and “period:$[1950 - 1960]$.”

5.2 Analysis of query relevance for reformulation

Within a Venn diagram we show the numbers of documents corresponding to each contour and area. This information gives a good feedback of the relevance of each contour query. Indeed, if a contour query is associated to a low number of documents, two possibilities appear: either the database does not contain documents relative to these query terms, or the terms are not used as indexes. This may lead to a reformulation of the query, using the authorities' explorer, in order to find terms forming the initial contour query. This process becomes important for databases whose indexing terms are dependent of the period. News databases are mostly concerned by such vocabulary evolution.

Scenario: Evaluate contour queries relevance

The task here is to find documents about De Gaulle's foreign policy from 1950 to 1970. In French, the terms for foreign policy have changed over time. As documents have mainly been indexed while being broadcasted, the three following terms may have been used over time to describe ”foreign policy”: ”politique internationale” (international policy), ”politique extérieure” (external policy) and ”politique étrangère” (foreign policy). Then, the query ”politique internationale” AND ”De Gaulle” AND date:$[1950 - 1970]$ will give a surprising Venn diagrams with an empty set for ”politique internationale” (cf. Figure 6). We can easily see that this

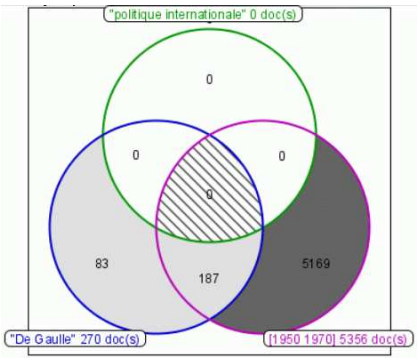

Fig. 6. Venn diagram on “politique internationale”, “De Gaulle” and “date:$[1950 - 1970]$.

term is never used and the strategy here will be to try and combine the other terms.

Figure 7 shows more realistic results.

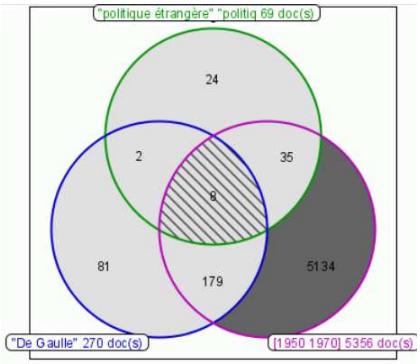

Fig. 7. Venn diagram on “politique ´etrang`ere”, “De Gaulle” and “date:$[1950 - 1970]$.

5.3 Identify the best results

Conjunctive queries (”AND” queries) are really common, but a recurring problem is that the result set may be empty or contain only few documents, especially when the query is composed of several sub-queries. In the other hand, dual disjunctive queries produce huge results sets. The repartition of the documents inside the Venn diagram regions allows the user to iteratively go through the sets of results while choosing first the regions corresponding to the most relevant combinations of terms.

Scenario: Identify most relevant documents

The task here is to find documents about wars and conflicts that involve France, United Kingdom, Germany and USA.

Let’s define each contour query:

WAR: war conflict army military

FRANCE: France

UK: England ”United Kingdom” ”Great Britain” UK GB

GERMANY: Germany GRD FRG

USA: USA ”United States” ”United States of America”

For the disjunctive query WAR OR FRANCE OR UK OR GERMANY OR USA we get 15204 documents, and for the conjunctive query we have no document. With the corresponding Venn diagram (cf. Figure 8) we can find two interesting documents. These documents are indexed by WAR and 3 of the 4 specified countries.

5.4 A Database Cartography

We call database cartography any Diagram representing a database repartition and such that there is a surjection from the set of documents to the set of terms used in the contour queries. Moreover, the terms should be semantically coherent.

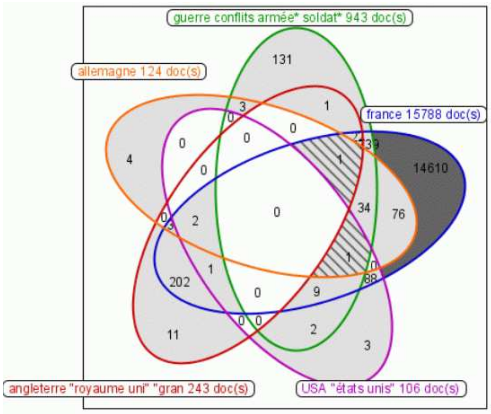

Fig. 8. Cartography for “WAR”, “FRANCE”, “UK”, “GERMANY”, “USA”.

Temporal Cartography

Figure 9 shows the map of repartition of the documents according to periods, from 1930 to 2004. This map gives a good view of the evolution of French TV and INA’s digitalization strategy. We observe that the period 1930-1950 contains fewer documents than the other contour queries. Indeed programs of this period are fi lms and are very expensive to digitize. Moreover, this map illustrates the policy of digitization at INA.

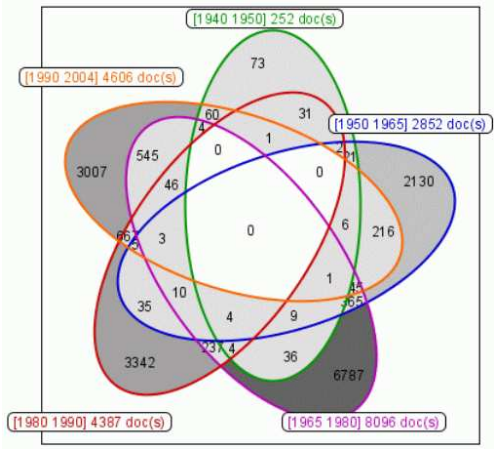

Fig. 9. Cartography for $[1940-1950]$, $[1950-1965]$, $[1965-1980]$, $[1980-1990]$, $[1950-2004]$.

Programs Type Cartography

Figure 10 shows cartography for the main programs types:

Song and music ;

Sport ;

Fictions, movies and series ;

News ;

Documentaries.

Those categories cover two thirds of this database, which contains mainly news and

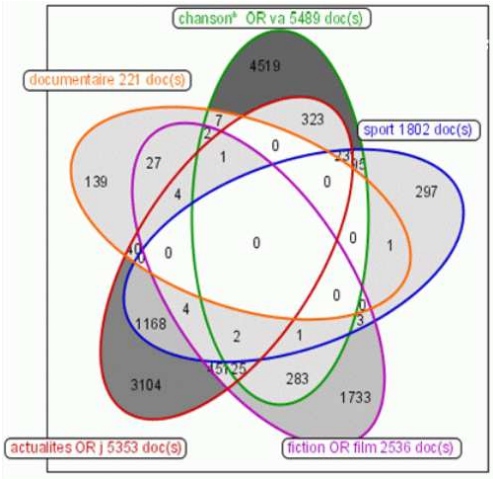

Fig. 10. Cartography for “song and music”, “sport”, “fictions, movies and series”, “news” and “documentaries”.

”song and music” documents. There are much fewer documentaries than fi ctions. Sport is mainly diffused during news programs.

6 Conclusion

The use of Venn or Euler diagrams in traditional library environments improves the search process. As noticed in [6], users have serious and sophisticated information needs. But they may be unable to express them in order to build effi cient queries because of a lack of knowledge of the documentary kernel. The authority explorer gives them a simple and rapid way to get appropriate terms to compose their queries. The document searcher proposes a representation to analyze and to go through the results of a composed query. Furthermore, this interactive tool helps the user to elaborate strategies and provides an ”all graphic” interface to realize them.
This work is being tested and validated with users. Graphical attributes and additional functions are under discussion in order to satisfy users’ needs and wishes.

References

[1] J. Avrahami and Y. Kareev. What do you expect to get when you ask for ”a cup of coffee and a muffin or croissant? On the interpretation of sentences containing multiples connectives. *International Journal of Man- Machine Studies*, No 38, pages 429-434, 1993.

[2] B. Bederson, J. Grosjean and J. Meyer. Toolkit Design for Interactive Structured Graphics. *IEEE Transactions on Software Engineering*, 30 (8), pages 535-546.2004

[3] J. Bertin. *Semiology of Graphics.* The University of Wisconsin Press, 1983.

[4] S. Chow and F. Ruskey. Towards a general solution to drawing area-proportional Euler diagrams. In *Euler Diagrams 2004*, ENTCS, 2004.

[5] M. Chui and A. Dillon. Speed and accuracy using four boolean query systems. In *Tenth Midwest Artificial Intelligence and Cognitive Science Conference*, pages 36-42, 1999.

[6] R. K. France, L.T. Nowell, E.A. Fox, R.A. Saad and J. Zhao. *Use and usability in a digital library search system.* Unpublished manuscript, 1999.

[7] ISO 999: *Information and documentation - Guidelines for the content, organization and presentation of indexes*, 1996.

[8] ISO 5963: *Documentation - Methods for examining documents, determining their subjects, and selecting indexing terms*, 1985.

[9] J. Flower and J. Howse. Generating Euler diagrams. In *Diagrams 2002*, pages 61-75, LNAI 2317, Springer Verlag, 2002.

[10] M. Hertzum and E. Frokjaer. Browsing and Querying in Online Documentation: A Study of User Interfaces and the Interaction Process. *ACM Transactions on Computer-Human Interaction*, Vol. 3, No 2, pages 136-161, 1996.

[11] S. Jones. Graphical Query Specification and Dynamic Result Previews for a Digital Library. In *ACM symposium on User Interface Software and Technology*, pages 143-151, 1998.

[12] A. Michard. A new database query language for non-professional users: Design principles and ergonomic evaluation. *Behavorial and Information Technology*, vol. 1,3, pages 279-288, 1982.

[13] K. Perlin and D. Fox. Pad: an alternative approach to the computer interface. In *Proceedings of SIGGRAPH'93*, pages 57-64 1993.

[14] F. Ruskey. A survey of Venn diagrams. *The electronic journal of combinatorics*, 2001 website: http://www.combinatorics.org/Surveys/ds5/VennEJC.html

[15] G. Salton. On the use of term association in automatic information retrieval. In *Coling'86, 11th international conference on computational linguistics*, pages 380–386, 1986.

[16] A. Verroust and M-L. Viaud. Ensuring the drawability of extended Euler diagrams for up to 8 sets. In *Diagrams 2004*, pages 128-141, Cambridge, 2004.

[17] D. Young and B. Shneiderman. A Graphical Filter/Flow Representation of Boolean Queries: A Prototype Implementation and Evaluation. *Journal of the American Society for Information Science*, Vol. 44, No 6, pages 327-339, 1993.