

# Modelling and Use of Domain-Specific Knowledge for Similarity and Visualization

TROELS ANDREASEN & HENRIK BULSKOV &  
RASMUS KNAPPE<sup>1</sup>

## 1. Introduction

The use of ontologies can contribute significantly to the organization of concepts, structure and relations within a knowledge domain. Incorporation of ontologies in tools for information access provides foundation for enhanced, knowledge-based approaches to surveying, indexing and querying of document collections.

We introduce in this paper the notion of an *instantiated ontology* as a subontology derived from a general ontology and restricted by the set of instantiated concepts in a target document collection. This instantiated ontology represents a conceptual organization reflecting the document collection, and reveals domain knowledge, for instance about the thematic areas of the domain which in turn facilitates means for a topic-based navigation and visualization of the structure within the domain.

Modeling and use of ontologies is the major focus in this paper. We introduce, in section 2, to a formalism for representation of ontologies. Section 3 describes the modeling of general and instantiated ontologies respectively. In section 4 it is discussed how to establish a measure of domain-specific similarity from a so-called domain-specific ontology. Section 5 introduces the use of an instantiated ontology for domain and query visualization.

Concept expressions, that are the key to modelling and use of ontologies, are explained in more detail below, we refer, however, to (Andreasen et al., 2004b) for a discussion of general principles behind parsing for concepts.

<sup>1</sup> Department of Computer Science, Roskilde University, P.O. Box 260, DK-4000 Roskilde, Denmark {troels,bulskov,knappe}@ruc.dk

## 2. Representation of ontologies

The purpose of the ontology, in this context, is to define and relate concepts that may appear in the document collection or in queries to this. We define a generative ontology framework where a basis ontology situates a set of atomic term concepts  $A$  in a concept inclusion lattice. A concept language (description language) defines a set of well-formed concepts, including both atomic and compound term concepts. The concept language used here, Ontolog (Nilsson 2001), defines a set of semantic relations  $\mathbf{R}$  that can be used for “attribution” (feature-attachment) of concepts to form compound concepts. The set of available relations may vary with different domains and applications. We may choose  $\mathbf{R} = \{\text{WRT,CHR,CBY,TMP,LOC,...}\}$ , for with respect to, characterized by, caused by, temporal, location, respectively.

Expressions in Ontolog are concepts situated in the ontology formed by an algebraic lattice with concept inclusion (ISA) as the ordering relation. Attribution of concepts can be written as feature structures. Simple attribution of a concept  $c_1$  with relation  $r$  and a concept  $c_2$  is denoted  $c_1[r:c_2]$ .

Given atomic concepts  $A$  and relations  $\mathbf{R}$ , the set of well-formed terms  $L$  of the Ontolog language is defined as follows.

- if  $x \in A$  then  $x \in L$
- if  $x \in L$ ,  $r_i \in \mathbf{R}$  and  $y_i \in L, i = 1, \dots, n$   
then  $x[r_1 : y_1, \dots, r_n : y_n] \in L$

It appears that compound terms can be built from nesting, for instance,  $c_1[r_1 : c_2[r_2 : c_3]]$  and from multiple attribution as in  $c_1[r_1 : c_2, r_2 : c_3]$ . The attributes of a term with multiple attributes  $T = x[r_1 : y_1, \dots, r_n : y_n]$  are considered as a set, thus we can rewrite  $T$  with any permutation of  $\{r_1 : y_1, \dots, r_n : y_n\}$ .

## 3. Modeling ontologies

One objective in ontology modeling, is for the domain expert or knowledge engineer to construct a knowledge base ontology over atomic or multi-word concepts.

Ontology modeling in the present information retrieval context consists of two parts. The inclusion of knowledge from available knowledge sources into a general ontology and a restriction to the part of the general ontology covering the instantiated concepts in the document collection. The first part involves modeling of concepts in a generative ontology and the second part the

so-called domain-specific ontology is retrieved as a subontology of the general ontology. The restriction to this subontology is build based on the set of concepts that appears (is instantiated) in the document collection and the result is called an instantiated ontology.

### 3.1 The general ontology

Sources for knowledge base ontologies may have various forms. Typically a taxonomy can be provided with supplements, e.g. word and term lists as well as dictionaries for definition of vocabularies and for handling of morphology. We will not go into details on the modeling here but just assume the presence of a taxonomy in the form of a simple taxonomic concept inclusion relation  $ISA_{KB}$  over the set of atomic concepts  $\mathbf{A}$ .  $ISA_{KB}$  and  $\mathbf{A}$  expresses the domain and world knowledge provided.  $ISA_{KB}$  is assumed to be explicitly specified – e.g. by domain experts – and would most typically not be transitively closed.

Based on  $ISA_{TRAN}$  the transitive closure of  $ISA_{KB}$ , we can generalize into a relation over all well-formed terms of the language  $\mathbf{L}$  by the following:

- if  $x ISA_{TRAN} y$  then  $x \leq y$
- if  $x[...] \leq y[...]$  then also  
     $x[..., r : z] \leq y[...]$ , and  
     $x[..., r : z] \leq y[..., r : z]$ ,
- if  $x \leq y$  then also  
     $z[..., r : x] \leq z[..., r : y]$

where repeated ... in each inequality denote zero or more attributes of the form  $r_i : w_i$ .

The general ontology  $\mathbf{O} = (\mathbf{L}, \leq, \mathbf{R})$  thus encompasses a set of well-formed expressions  $\mathbf{L}$  derived in the concept language from a set of atomic concepts  $\mathbf{A}$ , an inclusion relation generalized from an expert provided relation  $ISA_{KB}$  and a supplementary set of semantic relations  $\mathbf{R}$ , where for  $r \in \mathbf{R}$  we obviously have that  $x[r : y] \leq x$  and that  $x[r : y]$  is in relation  $r$  to  $y$ . Observe that  $\mathbf{L}$  is infinite and that  $\mathbf{O}$  thus is generative.

### 3.2 The domain-specific ontology

Apart from the general ontology  $\mathbf{O}$ , the target document collection contributes to the construction of the domain ontology. We assume a processing of the target document collection, where an indexing, formed by sets of concepts from  $\mathbf{L}$ , of text in documents is attached. In broad terms the domain on-

tology is a restriction of the general ontology to the concepts appearing in this indexing.

More specifically the generative ontology is, by means of concept occurrence analysis over the document collection, transformed into a domain specific ontology restricted to include only the concepts instantiated in the documents covering that particular domain.

(The final paper will contain a formal definition of the instantiated ontology, as well as examples showing the modelling of an example instantiated ontology, based on figure 1)

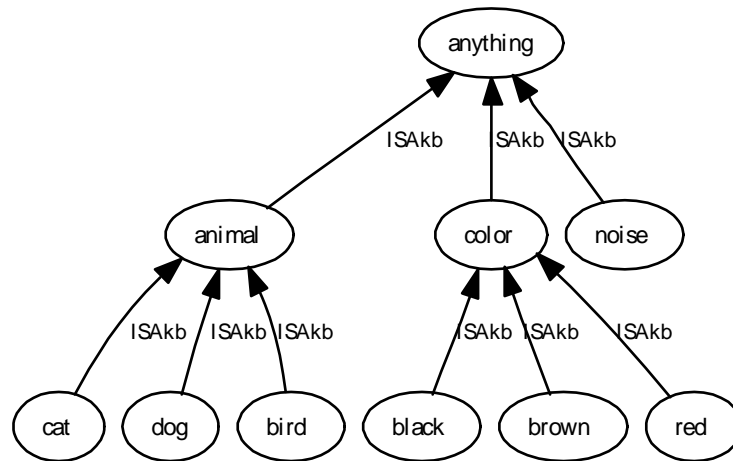


Figure 1 An example knowledge base ontology  $ISA_{KB}$

#### 4. Deriving similarity

The domain-specific ontology may provide an excellent means to survey and give perspective to the collection. However as far as access to documents is concerned ontology reasoning is not the most obvious evaluation strategy and it may well entail scaling problems. Applying measures of similarity derived from the ontology is a way to replace reasoning with simple computation still influenced by the ontology. A well-known and straightforward approach to this is the shortest path approach (Bulskov et al. 2002, Rada et al. 1989), where closeness between two concepts in the ontology implies high similarity. A problem with this approach is that multiple connections are ignored. In the ontology in figure 2 we thus have that the shortest path similarity between *cat* and *dog* would be equal to or greater than the similarity between *cat*[CHR:*black*] and *dog*[CHR:*black*] (depending on whether CHR-edges are included or not), while intuitively the former should be less than the latter because we have two concepts that meet in *animal* AND share the *black*-property.

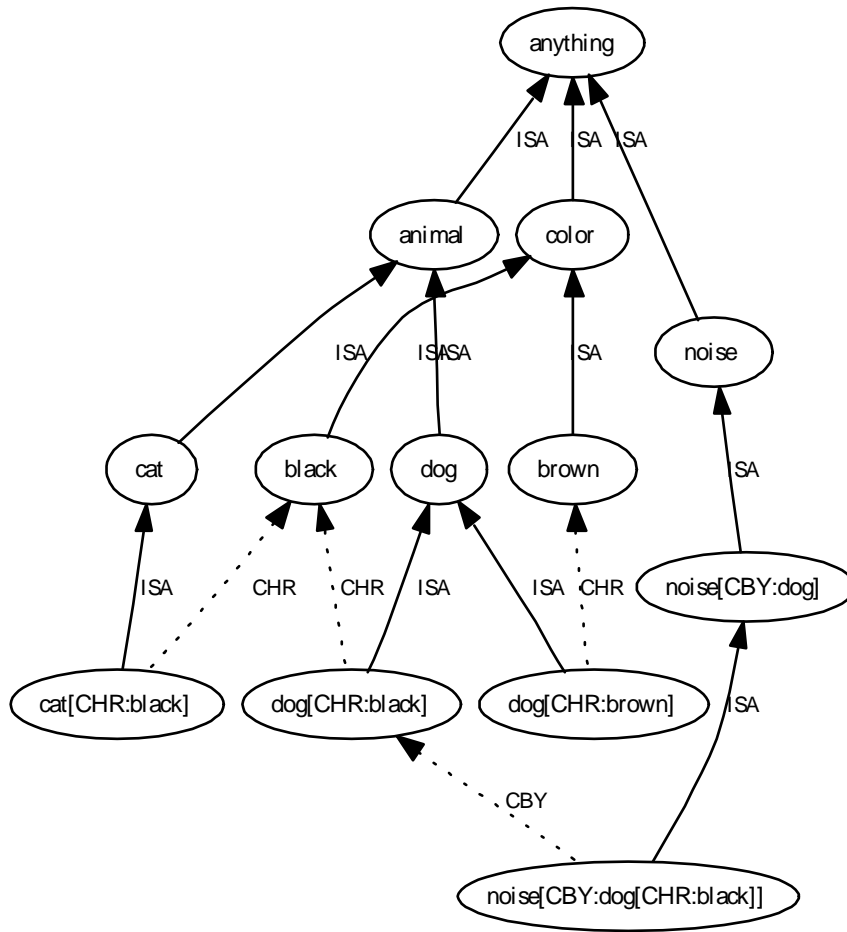


Figure 2 An example instantiated ontology

To differentiate here an option is to consider all paths rather than only the shortest path. A “shared nodes” approach that reflects multiple paths, but still avoids the obvious complexity of full computation of all paths is presented in (Andreasen et al. 2004a). In this approach the basis for the similarity between two concepts  $c_1$  and  $c_2$  is the set of “upwards reachable” concepts (nodes) shared between  $c_1$  and  $c_2$ . This is, with  $\alpha(x) = \omega(\tau(x))$ , the intersection  $\alpha(x) \cap \alpha(y)$ .

Similarity can be defined in various ways, one option being, as described in (Andreasen et al., 2003), a weighted average, where  $\rho \in [0,1]$  determines the degree of influence of the nodes reachable from  $x$  respectively  $y$ .

$$sim(x, y) = \rho \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|} + (1 - \rho) \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(y)|} \quad (1)$$

As it appears the upwards expansion  $\alpha(c)$  includes not only all subsuming concepts  $\{c \mid c \text{ ISA } c_i\}$  but also concepts that appears as attributes to  $c$  or to any

subsuming concept of attributes. The latter must be included if we want to cope with multiple connections and want to consider for instance two concepts more similar if they bear the same color. However, a further refinement seems appropriate here. If we want two concepts to be more similar if they have an immediate subsuming concept (e.g.  $cat[CHR:black]$  and  $cat[CHR:brown]$  due to the subsuming  $cat$ ) than if they only share an attribute (e.g.  $black$  shared by  $cat[CHR:black]$  and  $dog[CHR:black]$ ) we must differentiate and cannot just define  $\alpha(c)$  as a crisp set. The following is a generalization to fuzzy set based similarity.

First of all notice that  $\alpha(c)$  can be derived as follows. Let the triple  $(x, y, r)$  be the edge of type  $r$  from concept  $x$  to concept  $y$ ,  $E$  be the set of all edges in the ontology, and  $T$  be the top concept. Then we have:

$$\begin{aligned}\alpha(T) &= \{T\} \\ \alpha(c) &= \{c\} \cup \left( \bigcup_{(c, c_i, r) \in E} \alpha(c_i) \right)\end{aligned}$$

A simple modification that generalizes  $\alpha(c)$  to a fuzzy set is obtained through a function  $weight(r)$  that attaches a weight to each relation type  $r$ . With this function we can generalize to:

$$\begin{aligned}\alpha(T) &= \{1/T\} \\ \alpha(c) &= \{c\} \cup \left( \bigcup_{(c, c_i, r) \in E} weight(r) \alpha(c_i) \right) \\ &= \{c\} \cup \left( \bigcup_{(c, c_i, r) \in E} \sum_{\mu(c_{ij})/c_{ij} \in \alpha(c_i)} weight(r) \mu(c_{ij})/c_{ij} \right)\end{aligned}$$

$\alpha(c)$  is thus the fuzzy set of nodes reachable from the concept  $c$  and modified by weights of relations  $weight(r)$ . For instance from the instantiated ontology in figure 2 and assuming relation weights  $weight(ISA)=1$ ,  $weight(CHR)=0.5$  and  $weight(CBY)=0.5$  we have:

$$\begin{aligned}\alpha(noise[CBY:dog[CHR:black]]) &= \\ 1/noise[CBY:dog[CHR:black]] &+ 1/noise + \\ 0.5/dog[CHR:black] &+ 0.5/dog + 0.5/animal + 0.25/black + \\ 0.25/color &+ 1/anything\end{aligned}$$

For concept similarity we can still use the parameterized expression (1) above, applying minimum for fuzzy intersection and sum for fuzzy cardinality:

$$\begin{aligned}\alpha(cat[CHR:black]) \cap \alpha(dog[CHR:black]) &= \\ 0.5/black &+ 0.5/color + 1/animal + 1/anything \\ |\alpha(cat[CHR:black]) \cap \alpha(dog[CHR:black])| &= 3.0\end{aligned}$$

The similarities between *dog*[CHR:*black*] and other concepts in the ontology are, when collected in a fuzzy subset of similar concepts (with  $similar(x) = sim(x, y) / y$ ) and  $\rho = \frac{4}{5}$  the following:

$$\begin{aligned} similar(dog[CHR:black]) &= 1.00 / dog[CHR:black] + \\ &0.68 / dog + 0,6 / cat[CHR:black] + \\ &0.6 / noise[CBY:dog[CHR:black]] + 0,52 / animal + 0,45 / black \\ &+ 0,45 / cat + 0,39 / color + 0,36 / anything + 0,34 / brown + 0.26 / noise \end{aligned}$$

## 5. Applications

As the instantiated ontology is a restriction of a general ontology with respect to a set of concepts, it can be used for providing structured descriptions. The restriction could be with respect to the sets of concepts in a particular target document collection, as described earlier, but it could also comprise the set of concept of a query, the set of concept in a complete search result, or part of the search result as in relevance feedback, or any set of concept selected by the user or the querying system.

The notion of instantiated ontologies has as such applications with respect to navigation and surveying of the topics covered by the domain in question, where the domain could be the instantiated ontology of any of the suggested restrictions above, not only a particular target document collection.

### 5.1 A simple prototype system

The examples are constructed using a prototype system based on a general ontology constructed from the WordNet lexicon (Miller 1990, Miller 1995), the Suggested Upper Merged Ontology (SUMO) (Niles 2001), and the mid-level ontology (MILO) (Niles 2001) designed to bridge the high-level ontology SUMO and WordNet. The knowledge base ontology contains approximately 100.000 concepts (synsets). The ontology relation  $ISA_{KB}$  is based on the hypernym relation individually from WordNet, MILO, and SUMO and the three relations, equivalence, subsumed by, and instance of, from the mapping between WordNet and MILO/SUMO.

### 5.2 Navigating and surveying

One of the difficulties users have to overcome when querying information systems, concerns the transformation of their need of information into descriptions used by the system. As the concepts instantiated in the document collection expresses the information available, the instantiated ontology therefore provides a structuring of this information.

Consider the example in figure 2, where we have a document collection with the following four instantiated concepts,  $I = \{palisade, stockade[CHR:old], rampart[CHR:old], church[CHR:old]\}$ .

The instantiated ontology reveals two different aspects covered by the document collection, 1) different kinds of fortifications and 2) a place of worship. On a more general level the instantiated ontology describes buildings and the abstract notion of something dated back in time.

As the size and complexity of the instantiated ontology increases, it can be difficult for the user to form a general view – mainly due to the complexity and volume of the ontology.

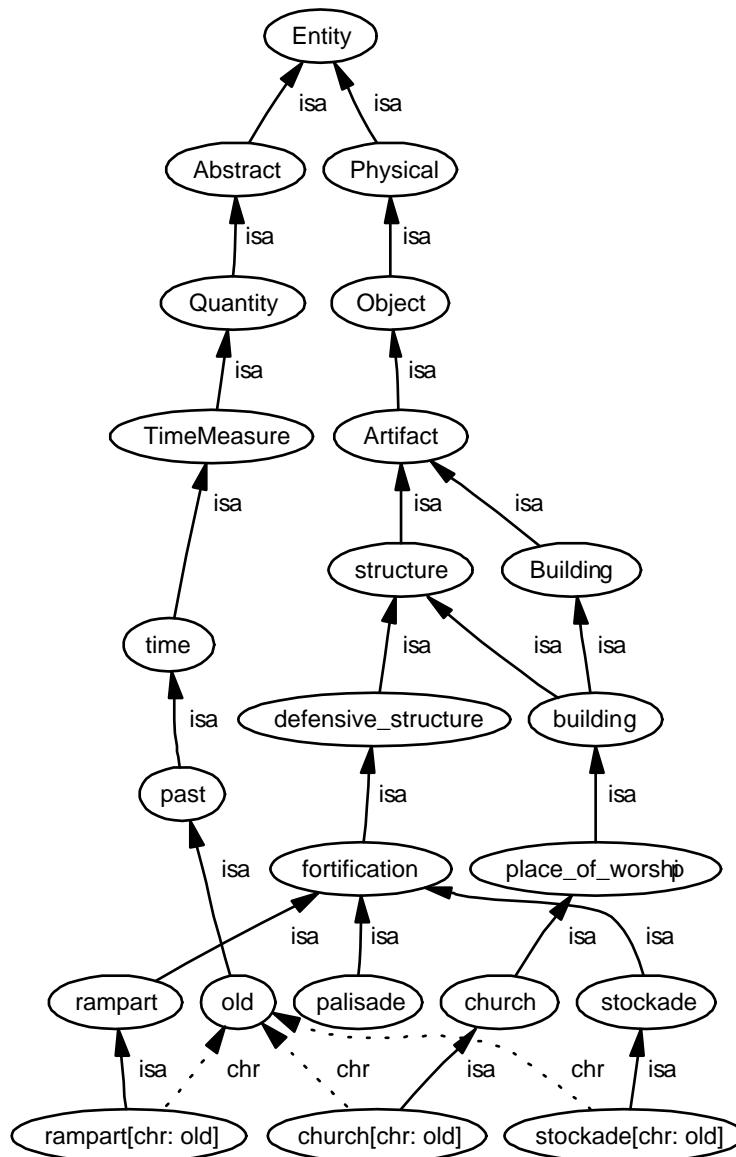


Figure 2: A simple instantiated ontology, based on WordNet, SUMO and MILO and the four concepts *stockade[chr:old]*, *rampart[chr:old]*, *church[chr:old]*, *palisade*

One possibility in the direction of overcoming this challenge is to remove the concepts which do not contribute with significant information and have minor influence on the overall structure.

As one can see in figure 2, there are concepts present in the instantiated ontology that are either very abstract or are not part of a everyday vocabulary. These kinds of concepts could possibly contribute to confusion and could therefore be candidates for exclusion from the visualization. By utilizing the notion of *familiarity* as described in (Beckwith and Miller), these concepts can be selected. Familiarity is defined using the correlation there exist between frequency of occurrence and polysemy. Associated with every word form in the lexicon, there is an integer that represents a count, using the Collins Dictionary of the English Language, of the number of senses that word form has when it is used as a noun, verb, adjective, or adverb (Beckwith and Miller).

One very simple way to utilize familiarity is to eliminate all concepts from the visualization of the instantiated ontology, having a familiarity lower than a certain threshold.

### 5.3 Visualizing queries

Another use of instantiated ontologies is for visualizing user queries. When users pose queries to the system using polysemous concepts, the instantiated ontology constructed from the query can be used to visualize the different senses known to the system. If for example a user poses a query  $Q=\{bank,huge\}$ , then the system cannot use the concept *huge* to disambiguate *bank*, since *huge* can be used in connection many different senses of *bank*.

One possible way to incorporate the knowledge visualized is to ask the user to identify the correct senses of the concepts used in the query, and use the disambiguated concept in the query evaluation.

## 6. Conclusion

Firstly, we have introduced the notion of a domain-specific ontology as a restriction of a general ontology to the concepts instantiated in a document collection, and we have demonstrated its applications with respect to navigation and surveying of a target document collection.

Finally, we have presented a methodology for deriving similarity using the domain-specific ontology by means of weighted shared nodes. The proposed measure incorporates multiple aspects when calculating overall similarity between concepts, but also respects the structure and relations of the ontology.

## References

Andreassen, T., Bulskov, H. and Knappe, R.: *On Querying Ontologies and Databases, Flexible Query Answering Systems*, 6th International Conference, FQAS 2004, Lyon, France, June 24-26, 2004, Proceedings

T. Andreassen, P. Anker Jensen, J. Fischer Nilsson, P. Paggio, B.S. Pedersen, H. Erdman Thomsen: *Content-based Text Querying with Ontological Descriptors*, in *Data & Knowledge Engineering* 48 (2004) pp 199-219, Elsevier, 2004.

Andreassen, T., Bulskov, H., and Knappe, R.: *Similarity from Conceptual Relations*, pp. 179–184 in Ellen Walker (Eds.): 22nd International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003, Chicago, Illinois USA, July 24–26, 2003, Proceedings

Beckwith, R.; Miller, G. A. & Teng, R. (Eds.): *Design and implementation of the WordNet lexical database and searching software*, <http://www.cogsci.princeton.edu/wn/5papers.ps>.

Bulskov, H.; Knappe, R. & Andreassen, T.: *On Measuring Similarity for Conceptual Querying*, LNAI 2522, pp. 100–111 in Andreassen T.; Motro A.; Christiansen H.; Larsen H.L.(Eds.): *Flexible Query Answering Systems 5th International Conference, FQAS 2002*, Copenhagen, Denmark, October 27–29, 2002, Proceedings

Miller, George: *WordNet: An On-line Lexical Database*, *International Journal of Lexicography*, Volume 3, Number 4, 1990

Miller, George: *WordNet: An Lexical Database for English*, *Communication of the ACM*, Volume 38, Number 11, pp. 39–41, 1995

Niles, I.; Pease, A.: *Towards a Standard Upper Ontology*, in Chris Welty and Barry Smith (eds.): *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, October 17-19, 2001.

Nilsson, J. Fischer: *A Logico-algebraic Framework for Ontologies – ONTOLOG*, in Jensen, P. Anker & Skadhauge, P. (eds.): *Proceedings of the First International OntoQuery Workshop – Ontology-based interpretation of NP's*,

*Modelling and Use of Domain-specific Knowledge*

Department of Business Communication and Information Science, University of Southern Denmark, Kolding, 2001

Rada, Roy; Mili, Hafeedh; Bicknell, Ellen & Blettner, Maria: *Development and Application of a Metric on Semantic Nets*, IEEE Transactions on Systems, Man, and Cybernetics, Volume 19, Number 1, pp. 17–30, 1989