Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

SERUBA – A New Search and Learning Technology for the Internet and Intranets

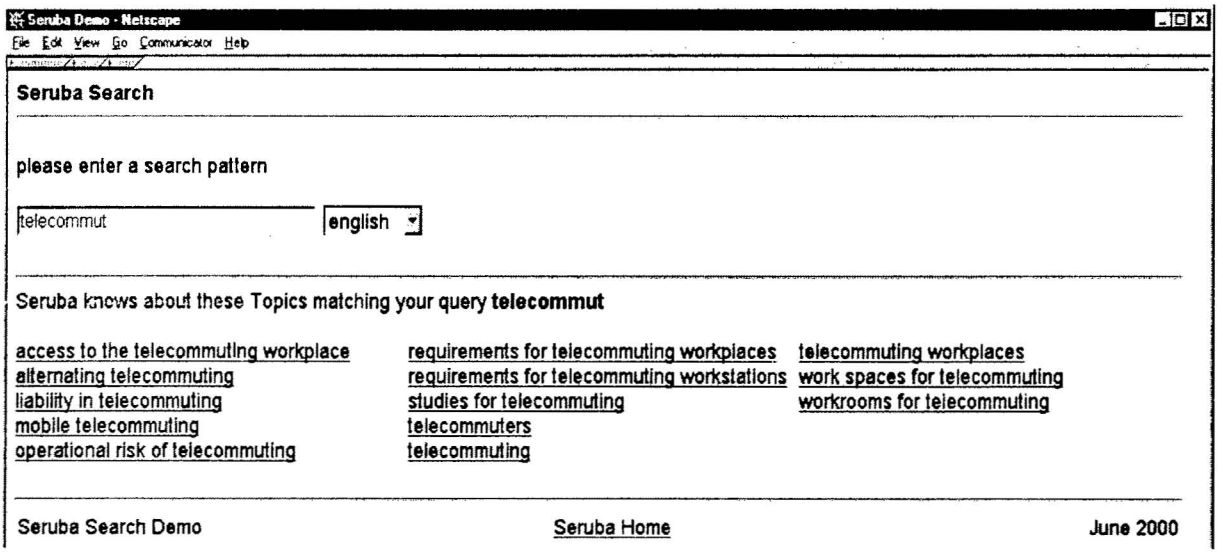
Winfried Schmitz-Esser
University of Applied Sciences, Hamburg, Germany
schmitz_esser@csi.com

Abstract

The paper describes a multi-lingual, ontology-based system for user support and learning in very large, non-domain specific network environments. The languages implemented are English, Spanish, French and German. SERUBA, named after its SEmantic and RUle-BAsed approach, will hit the Internet market early in 2001.

1 A search in SERUBA

Assume a user is interested in *telecommuting*. He is at the point of keying in this word, but stops, since he is aware that he would miss *telecommuter* , *telecommute*, and some others. So he restricts his search string to *telecommut*.

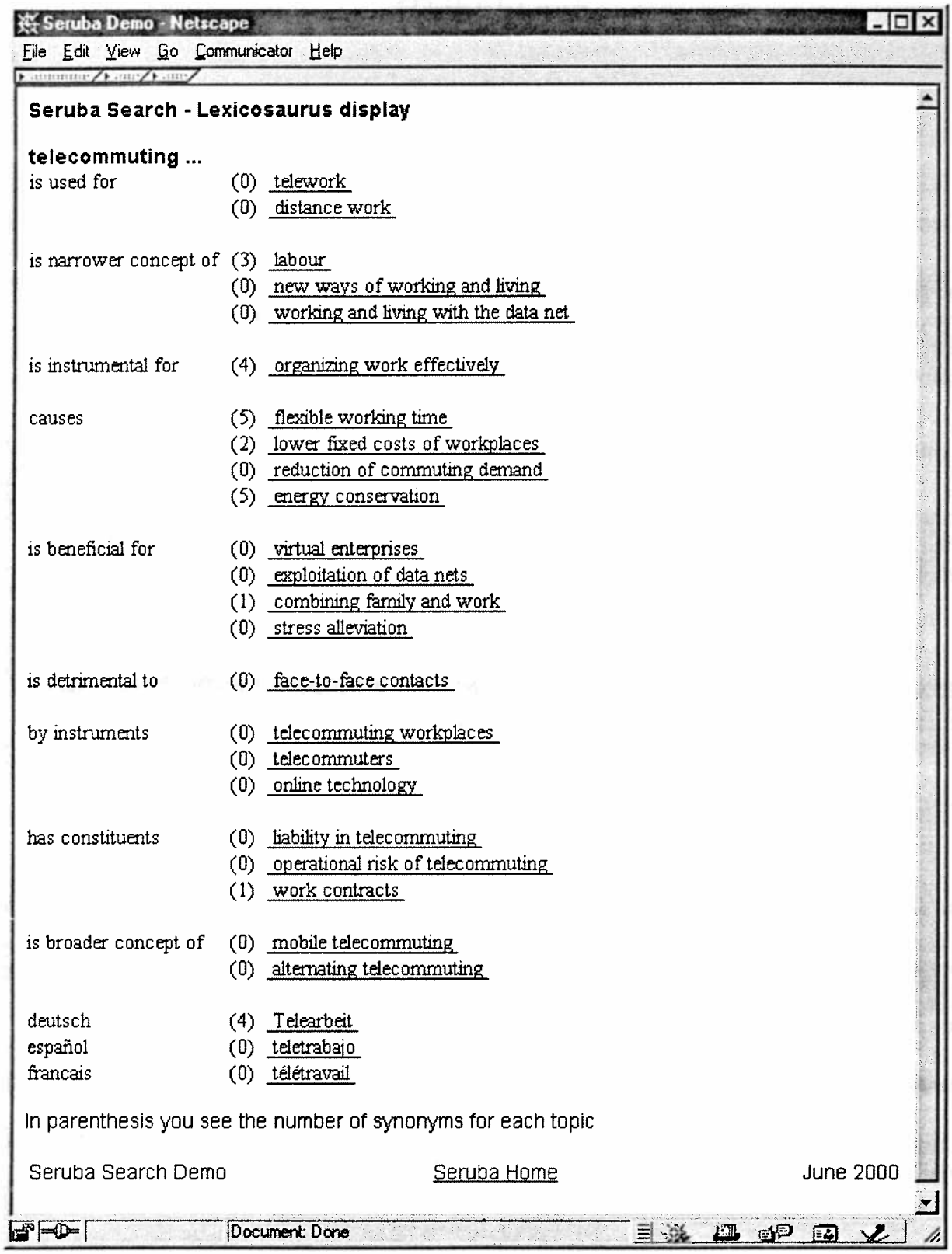


Among the topics shown, the user doesn't feel sure yet, falls back to his initial idea and explores *telecommuting*.

ISSN: 2324-9773

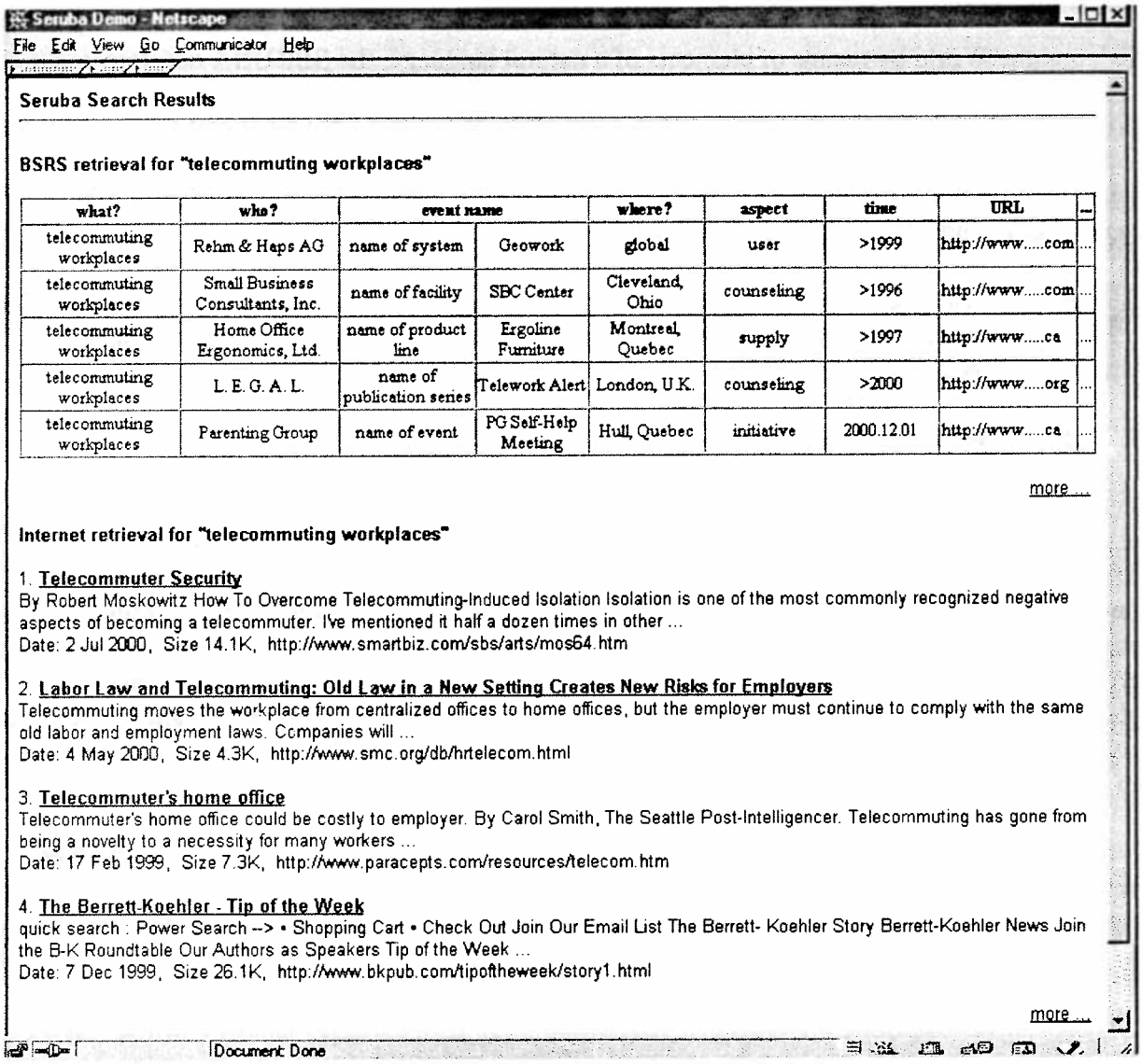
Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

Now the user sees a display of concepts and whole themes related to *telecommuting*, arranged by SERUBA's up to ten semantic relationships. He chooses *telecommuting workplaces*.



ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772



Search results are divided into

- entries from the BSRS – Basic Semantic Reference Structure, and
- matches from the Internet.

Further columns in the BSRS, with pictorial information (e.g., drawings, film spots), functional information (e.g., operating hours), and referential information (e.g,. DDC numbers) may be viewed. In part, entries in the BSRS come as paid listings, offered under SERUBA's four-language publishing scheme of YELLOWTRONIC PAGES.

The remainder of this paper deals with the inner workings of SERUBA and shows how a result of this complexity and universality can be achieved.

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

2 The SERUBA Basic Semantic Reference Structure (BSRS) and ontology

Modeling the World by means of elements of a natural language, the SERUBA ontology makes a fundamental distinction between *universals* and *instances* and accommodates them both. While universals reflect concepts, themes and ideas, i.e., all that is, or may be, an object of common thought and expression, instances in this view designate phenomena unique in space and time - a human being, a pet, a company, a product, a trade mark, an event, etc. - in short,: all that is regarded as an individual and therefore has a name.

SERUBA is a search and knowledge acquisition technology based on an ontology that allows for this fundamental distinction. This section, the main section of the paper, presents its main features and functionalities.

2.1 The Basic Semantic Reference Structure (BSRS)

According to this fundamental distinction, the SERUBA ontology uses designators representing the two types. A template, engineered according to the requirements of a Basic Semantic Reference Structure (BSRS), is filled with the entries, each being assigned a distinct place in a line of entries, according to its ontological quality (e.g., a person, a corporation, a product, a concept, etc.). Each line is a row in a relational data base table. The tuple represents the elements needed as retrieval cues for expected queries, to wit, *What?*, *Who?*, *Event?*, *Where?*, and *When?* questions. Each tuple is controlled by an " aspect", which is the head of the tuple and governs its respective sense. By means of this syntax, a clear-cut, univocal read-out of the desired meaning can be assured. Each aspect has a way of read-out of its own.

More complex statements can be formed by more than one tuple, whereby the tuples involved are held together by so-called " onto-links". These onto-links express distinct concept relationships, so as to enable the modeling of statements such as: *deforestation* in *Amazon region* causes/may cause *desertification* in *Africa East of Sahara*.

Moreover, when applying the BSRS in indexing procedures, complex content encountered in a text, a picture, or a film, etc. can be assigned different single tuples, or sets of tuples, by means of " source links". The same may apply to represent classes as found in library classifications and other archiving schemes, including more complex expressions, like Dewey or UDC classification numbers.

Fig. 1 gives a few sample BSRS entries.

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

What?	Who?		Event?	Where?	When?	How?
Universal, concept, theme	**Person**	**Corporate body**	**Name of event**	**Space**	**Time**	**Aspect**
General manager	Mike Osborne	Asia Trading Co., Vancouver		Canada	>1998-11-1	Definition
General manager	Phil Hawking	Asia Trading Co., Vancouver		Canada	>2000-4-1	Definition
Planting of St. John's trees		Ministry of Agri-culture, Lima	El Algarrobo project	Peru	>1984	Propagation
Conservation of soil humidity		Cooperativa Suelo y Água, Grau	El Algarrobo project	Grau (Peru)	>1984	Impact desired
Coffee substitutes		Cooperativa Suelo y Água, Grau	El Algarrobo project	Grau	>1995	Production
Coffee substitutes		Coffee Trading Co., Lima	Carob ™	Andes region	> 1996	Offer

Fig. 1. The Basic Semantic Reference Structure (BSRS); referential columns, onto-links, source links omitted

2.2 The classification of universals

The column with the entries for the *What?* question gives universals, concepts, or themes designated, in a linguistically univocal way, through words of a natural language; these concepts or themes are rigorously conceived as classes.

2.2.1 Vocabulary control: Descriptors and Additional Access Expressions (AAEs)

Semantic control of the universals is crucial. This is performed by a new type of multi-lingual knowledge base, combining the characteristics of a classic information retrieval thesaurus, a semantic network, and a multi-relational lexicon, or encyclopedia, which at SERUBA is called a LEXICOSAURUS®.

Its unique conception follows ideas developed by a group of scientists who early in the nineties teamed up for an advanced, standardized thesaurus structure, the German Committee for Classification and Thesaurus Research (KTF) of the Deutsche Gesellschaft für Dokumentation (DGD). In the mid-nineties, the resulting thesaurus proposal was submitted to a field test during the preparations for the EXPO 2000 in Hannover. In the past few years, these ideas have been exchanged and further discussed with scientists from various other groups and schools of thought on an international level.

The technology applied is onomasiologic, which means that it is based on clear-cut definitions of concepts and themes. In the model, these are represented as objects, each being assigned a corresponding MetaLanguage Identification Number (MLIN), which is a clear-cut and unequivocal designation of the object. As is shown in Fig. 2, each object is represented by a MLIN linked to an Equivalence Chain of Descriptors (ECD) which stands for the object. The ECD has a natural language expression, called the descriptor, in each language considered. (In

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

SERUBA, four languages are implemented, - English, Spanish, French, and German. With these four languages we will be capable of reaching almost 80 per cent of all users on the WEB.) The main condition of functionality is that in each of the languages, the descriptor designates the object represented by the MLIN in a clear and unambiguous way.

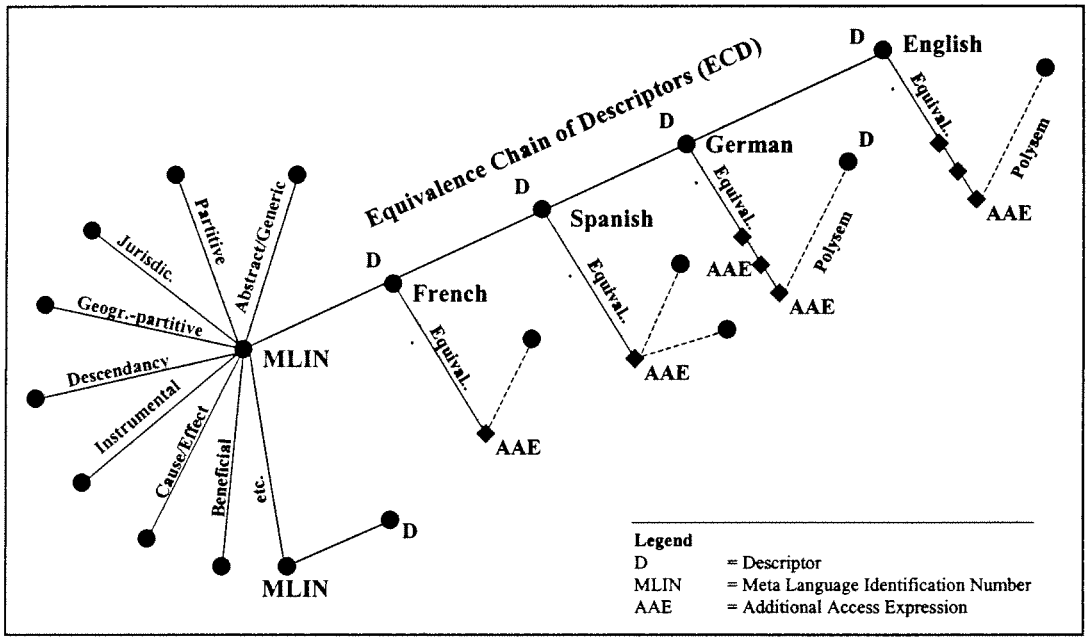


Fig. 2. The Semantic Network

Starting out from the concept definition linked to the MetaLanguage Identification Number (MLIN), Additional Access Expressions (AAEs) are found in each language, using a special procedure that finds the most current alternative expressions used in common communication, with proof of occurrences in the media being documented. Each of these alternative expressions opens up one more way of access to the object.

Many AAEs have only one meaning; such unambiguous synonyms can be automatically linked to their respective descriptors. Other AAEs have multiple meanings (homonyms or polysemes); these must be returned to the user asking him for disambiguation. So, if *Amazon* were the question, the user would be presented with various different meanings: What do you mean?, - *Amazon river*?, *Amazon basin*? *Amazon region*? *Federal State of Amazonia*?, Greek mythological figure, etc.

AAEs are a matter of each separate language, and as such are typified as Class I relations (See Fig. 3.)

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

Equivalence	US Use preferred Synonym *Reforestation* *US afforestation*	UF Preferred Synonym Used for *afforestation* *UF reforestation*
Polyseme	UD Use Descriptor *Bank* *UD credit institute* *UD riverside*	UF **Descriptor Used For** (among others) *credit institute* *riverside* *UF bank* *UF bank*

Fig. 3. Class I relations, valid for each language separately.

Class I relations perform two services:

1. They provide multiple ways of access, ideally taking the user from whichever expression he may come from and leading him to a well-defined concept or theme.

2. They provide the ability to automatically expand the search for a given concept or theme by adding all the other elements of the respective Class I equivalence chain.

In addition this expansion, which is based on knowledge of specific terminological relationships that requires a considerable investment of intellectual work,. SERUBA will feature lingware for each of the four languages to automatically cope with certain rule-dependent variations on the " surface" of the language, such as: English/American spelling, singular/plural forms, nominalizations, and enumerations.

In sum, the goal of clearness and unequivocal ness requires that the object represented on the Equivalence Chain of Descriptors (ECD) be expressed by a universal, not a name. Universals are concepts or themes or topics that are common knowledge in civilized communities. All modern languages allow for the formation of a clear and unequivocal expression based on existing universals even for an unknown concept (i.e., a concept without a lexicalized designation in the language). This makes it possible to set up multilingual ontologies.

2.2.2 Relations between concepts

Universals are found to relate to each other in certain ways (Fig. 4). A *welding torch* is instrumental for *autogenous welding*, *deep rooted trees* favor the *conservation of soil humidity*, a *cat* is, or may be, a *pet*. The stipulation of such relationships between universals is part of the ontological description.

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

The SERUBA LEXICOSAURUS® distinguishes ten different relationships proper, or class II relations:

Abstract/generic	Cause/effect
Partitive (physical and theoretical)	Beneficial
Partitive (habits, law and jurisdiction)	Detrimental
Partitive (geographical, topographical)	Process applied.
Instrumental	Derivative

Four of these are illustrated in Figure 4. Other, more specific relationships, can be added later as needed. . The application of each of them is subject to a set of rules, including a definition as well as some clarification regarding the boundaries between them.

Type of relationship	Up	Down
Abstract/generic	St. John's trees *General* Deep-rooted trees	Deep-rooted trees *Specific* St. John's trees
Instrumental	Welding torch *Instrumental for* Autogenous welding	Autogenous welding *by instrument of* welding torch
Beneficial	Deep-rooted trees *favor* conservation of soil humidity	Conservation of soil humidity *profits from* deep-rooted trees
Derivative	Carob beans *Come/may come from* St. John's trees	St. John's trees *deliver/may deliver* carob beans
Derivative	Coffee substitutes *Come/may come from* Carob beans	carob beans *deliver/may deliver* coffee substitutes

Fig. 4. Class II Relations, - Relationships between universals (4 examples)

The yield of such fine-tuned distinction is rich and manifold. In searches, the user can be guided in a meaningful way from a fuzzy idea, from a word, or from a fraction of a word, to a theme, and from there through a display of themes to the most specific topics he might be interested in. By doing so, he obtains a clearer picture of how themes interrelate. This gives pleasure and surprise. It is an ideal tool for easy learning. With higher bandwidth to come in the near future, one can imagine an information seeker traveling through an audiovisual cosmos of themes, from star to star, from galaxy to galaxy, exploring with pleasure worlds of knowledge hitherto unknown to him or consolidating knowledge acquired previously but half forgotten. Many of these qualities can be witnessed even under the obvious limitations of present-day, page-oriented HTML platform environments we have to live with for now due to market conditions.

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

2.3 Treatment of instances

The ontology must be hospitable enough to house the instances as well. Universals basically depend on definitions stipulated in terms of the different natural languages, whereas instances, in general, have designations given for application in one language only. Normally, this is the language of the community where the instance lives or occurs. However, many exceptions are encountered in practice, and their number is on the rise as a result of a steadily growing international exchange.

What in German is *NATO*, is known in French as *OTAN*; in German you may find the full equivalent as *Atlantisches Verteidigungsbündnis*. Acronyms and other designations of such instances often come in mixed form, as, in German, *UNO*, *UN*, *Vereinte Nationen*, *VN*. Where you find full-fledged single language versions, language-specific treatment is inevitable, as in cases such as *Médecins sans Frontières* in French, the German equivalent being *Ärzte ohne Grenzen*, the English one *Doctors without Frontiers*.

So the developer of an ontology must earmark names, acronyms and equivalent expressions for language-dependent use in as many languages as are served by the ontology. Otherwise, the system would be unable to prompt the user for disambiguation.

Such a case could be given in German, if the user asked: *Uhu*, which is a brand name for a series of glues, and thus the designation of an instance. In this case, a need for disambiguation occurs only in searches carried out in German, where *Uhu* may also mean a bird, which in English is the *great horned owl*. This is not a problem for English. Species are not individuals and therefore must be seen as universals, in accordance with the logic of this approach.

Moreover, among the different types of instances examined for representation in the ontology, a need was felt to differentiate between instances which normally appear as active players or " actors" on the one hand, like persons, groups, corporations etc., and the usually passive instances, like products, services, brands, events, on the other. Both types can be real, like *Yves Montand* or *Robert Malthus*, or virtual, like *Donald Duck* or *Batman*, and relations may exist between the active and the passive instances, as well as between instances and universals. You also have to allow for relations which may exist between two or more active instances or two or more passive instances. *Richard Wagner* may have written on a fictive *Beethoven* (which he actually did), and *INA*, the *International Air Show* in Berlin may affect the *EXPO2000* in Hannover because, unfortunately, both events have been scheduled at the same time.

2.3.1 Space and time

Defining spatial criteria has its own problems. SERUBA's ontology allows for model (a) locations and areas on Earth and on planets, as well as (b) connections between them. Although geographical designations usually feature characteristics highly dependent on one particular language, they are dealt with as universals in the SERUBA ontology for practical reasons.

Time can be expressed as (a) a certain point in time, and as (b) extensions in time. Non-Christian equivalents can be given to users if the need arises.

ISSN: 2324-9773

Schmitz-Esser, W. (2000). SERUBA – A new search and learning technology for the internet and intranets. .11th ASIS SIG/CR Classification Research Workshop, 91-102. doi:10.7152/acro.v11i1.12772

3 How SERUBA works in searches

A search on the Net via SERUBA works as follows: The user enters a character string, e.g., *civil*, whereupon the system displays a KWIC index of descriptors and additional access expressions, including *civil engineering*, *civil law*, *civilization*, *uncivilized,* ..., etc. A click on one of these opens a page of the Lexicosaurus, where terms can be seen that have the different synonymous, hierarchical, partitive, and other relationships to the selected entry. Next, one or more descriptors are clicked into a " topic bag", which constitutes a Boolean AND. Synonyms are included automatically with a Boolean OR. Free character strings which are not in the Lexicosaurus may be added to the query, which is serviced by entries from the BSRS (" electronic yellow pages"), and, in parallel, submitted to one or more conventional search engines. Users profit from the vocabulary resources of the Lexicosaurus and language specific expansion lingware by retrieving documents using different but synonymous wording (recall device) and by suggestions from the displays for formulating their search in an unambiguous phrase or word string (precision device).

Thus SERUBA is not based on the often erroneous assumption that a user, when beginning a search, knows exactly what he wants, and, even more, that he is capable of figuring out which terms for the sought-for subject may have been used by the different authors of the texts he is browsing.

4 Applications

The ontology described can be used in many different ways. Direct application as a search tool on the Internet is just one of them. Another application is the following:

At SERUBA, a set of a dozen or so different aspects have been developed so far, offering a surprising range of possible standard types of expressions/declarations. Six of these which can be applied most easily will be used for a commercial product on the Internet which follows the idea of paid listings. This will be our four-language, electronic yellow pages service.

Moreover, the structures stipulated in the searches can be automatically exploited for searches and other undertakings in linguistic and knowledge engineering in many ways. Just think of the possibilities of expression offered by a combination of the BSRS and the LEXICOSAURUS. Easy, fun-filled, surprising access could be provided not only to knowledge of the ontological kind, but also to knowledge sources. URLs of the Internet need not be the only ones. I can imagine such new ways of access also being established for existing taxonomies, branch category schemes, and library classifications, along with their respective document delivery services.

Finally, its quality as an almost universal indexing tool qualifies SERUBA as an advanced technology for Media Asset management, and its quality as an ontology of contemporary knowledge or as a universal electronic encyclopedia.

ISSN: 2324-9773

References

Benking, H. 1996. *Concept and context mapping – towards common frames of reference.* In Terminology and Knowledge Engineering, Frankfurt am Main : Indeks, p. 35-47

Damer, B. 1998. Avatars! Peachpit Press, Berkeley

Fischer, D. H. 1998. *From Thesauri towards Ontologies?.* In W. Mustafa El-Hadi (ed.), Structures and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference, Würzburg : Ergon (Advances in Knowledge Organization ; Vol. 6), pp. 18-30

Gibbon, D. 1999. *Computational lexicography*, Doordrecht : Kluver

Mustafa el-Hadi, W. 1998. *Automatic Term Recognition & Extraction Tools: Examining the New Interfaces and their Effective Communication Role in LSP Discourse.* In W. Mustafa El-Hadi (ed.), Structures and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference, Würzburg : Ergon (Advances in Knowledge Organization ; Vol. 6), pp. 205-212

Rahmstorf, G. 1998. *Concept Structures for Large Vocabularies.* In W. Mustafa El-Hadi (ed.), Structures and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference, Würzburg : Ergon (Advances in Knowledge Organization ; Vol. 6), pp. 198-204

Rahmstorf, G., Schmitz-Esser, W., Schubert, K., Zimmermann, H. 1995. *Skizze zur Standardisierung sprachbezogener Begriffssysteme (Thesauri).* Paper presented at the ISKO-KTF Thesaurus-Workshop, Trier, October 17, 1995

Rahmstorf, G. 1994. *Semantisches Information Retrieval.* In W. Neubauer (ed.), Deutscher Dokumentartag 1994, Proceedings, Frankfurt am Main: Deutsche Gesellschaft für Dokumentation, pp. 237-260

Rahmstorf, G. 1983. *Die semantischen Relationen in nominalen Ausdrücken des Deutschen.* Dissertation, Mainz

Rostek, L., Möhr, W., Fischer, D., 1998. *Weaving a Web: The structure and creation of an object network representing an electronic reference work.* In: Fankhauser, P., Ockenfeld, M. (eds). Integrated Publication and Information Systems. GMD Sankt Augustin, pp 189-199

Schmitz-Esser, W. *2000. EXPO INFO 2000.* Springer, Berlin, Heidelberg

Schmitz-Esser, W., 1999: *Thesaurus and Beyond: An Advanced Formula for Linguistic Engineering and Information Retrieval.* In: Knowledge Organization 26 (1999), No. 1, p. 10-22.

Schmitz-Esser, W. 1998. *Defining the Conceptual Space for a World Exhibition – First Experiences.* In W. Mustafa El-Hadi (ed.), Structures and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference, Würzburg : Ergon (Advances in Knowledge Organization ; Vol. 6), pp. 146-152

ISSN: 2324-9773

Schmitz-Esser, W. 1994. *Thesaurus - frischer Anlauf: lexikographisch, mehrsprachig, maschinengängig, universal, für Informationslinguistik und Information Retrieval. Vorstellung eines im KTF erarbeiteten, neuen Thesauruskonzeptes*. In W. Neubauer (ed.), Deutscher Dokumentartag 1983, Proceedings, Frankfurt am Main : Deutsche Gesellschaft für Dokumentation, pp. 261-274

Schubert, K. 1995. *Parameters for the design of an intermediate language for multilingual thesauri*. In Knowledge Organization Vol. 22, no. 3/4, Frankfurt am Main : Indeks, pp. 136-140.

Zimmermann, H. 1993. *Aspektierung von Thesaurus-Relationen: Öffnung in universale Anwendbarkeit?* In W. Neubauer (ed.), Deutscher Dokumentartag 1993, Proceedings, Frankfurt am Main : Deutsche Gesellschaft für Dokumentation, pp. 275-290

ISSN: 2324-9773