Effective Terminology Support for Distributed Digital Collections

Martin Doerr
Institute of Computer Science, Foundation for Research and Technology Hellas
Heraklion-Crete, Greece

Paper presented on the Sixth DELOS Workshop, June 17-19, 1998

Abstract

Issues of embedding terminology service into large federations of multilingual digital collections are analyzed. A distributed system architecture is proposed, which preserves on one side a necessary degree of autonomy at the different sources, but allows on the other side to consistently correlate the chain of used terminology from the end-users to the collection maintainers. It should in particular allow for maintaining over the whole federation certain recall and precision properties of collections that use controlled vocabularies.

Introduction

One can roughly separate the problem of heterogeneity in distributed digital collections into a structural one - the differences in the schemata or document structures, and a terminological one - the differences in data values, which may refer to the same real items [Kram97]. There are two kinds of references, those to actual things, "***instances***", as "me, my house, my computer, my publications", and those to groups of things, either ***concepts***, as "researchers, buildings, PCs, essays, roads", or ***non-discrete sets***, as areas on the surface of earth. If one accesses a series of electronic collections, and wishes to retrieve data about certain things, there is the permanent problem of the identity of things referred to by terms. Each social group, be it a scientific discipline or a nation, uses other terms, and even individuals may differ in their use of terms. This problem is tackled by the use of so-called "authorities", which define and standardize terminology of a certain group and domain for consistent use in documentation and retrieval.

This works quite well on isolated databases, but is still insufficient for larger federations of databases. The hope to create a "world-wide" authority can be fairly regarded as an illusion. There are many reasons. First, authorities must be managed and evolved by initially autonomous groups at different paces; second, they often express different, not comparable views even on the same subject; last, the sheer size would be extraordinary, to mention only the most prominent reasons. In this situation, systems of interlinked thesauri or "domain knowledge bases" are proposed (e.g. [Wied94],[Soer96],[Doer97b],[Kram97]).

Terminology Support

Authorities basically try to solve two problems: The identification of a notion, and the definition of a concept (see also [Fosk97],[Soer96],[Doer97b]). For identification, linguistic expressions, so-called "terms", i.e. possible or preferred ***noun phrases*** or ***names*** are associated with a notion according to the practice of a social group. The notion in turn is described by attributes, as life-data of a person, free texts, geo-coordinates etc. The user may

select due to these descriptions fitting and unique terms, which can be used by the retrieval agent to match the notion behind with database records and occurrences in texts. If the database records use unique terms in the same sense (i.e. they use the authority for vocabulary control), the matching is immediate and precise. Else, the authority should list probable expressions for each notion, which can be used by a retrieval agent to calculate approximate matches (free text search). Much sophistication can be put into the context dependency of the probability of an expression and into the related matching algorithms. For "instances", this translation problem between notions and terms can be solved to the best possible, when we have gathered the terms of all groups for each notion. See e.g. the *United List of Artist Names* from the Getty Information Institute.

For concepts and non-discrete sets, however, each group tends to have its own definitions. Moreover, even the same items may be classified or referred to by coarser or narrower concepts. In addition to the pure translation problem, a correlation problem appears. If a data record uses a term for a real item, obviously all queries to broader concepts should include this item. Concepts of one group are therefore organized in subsumption hierarchies as thesauri. Concepts from different thesauri are correlated by equivalence and subsumption expressions in so-called "multilingual thesauri".

This knowledge allows retrieval agents to match data records about related items, but classified with different concepts (e.g.[Doer98]). The precision of this kind of matching is an open research issue. It may return larger or smaller answers than originally requested, e.g. subsumption properties invert in NOT clauses. Users may wish to have control on subsumption properties in the answers they get. If a literature subject is referred, the generality of the concept may or may not relate to the generality of the text. E.g. "Neural Diseases" may better match "Introduction to Neural Diseases" than "A New Approach to Therapy of the Creutzfeld-Jakob Syndrome", making things even more complex.

As above, free texts (scope notes), images etc. support further the identification of a concept by a user. Hence the translation of concepts consists of an identification and a correlation problem of all concepts of all groups, which is obviously an open ended task, as continuously new concepts appear. Authorities, in particular thesauri, can be regarded and dealt with as knowledge bases, which comprise domain knowledge in form of terminological logic, combined with a linguistic layer for concept identification.

Current Situation

From the point of view of implementation and system integration, the current situation can be described as follows:

- Either a separate, not integrated thesaurus tool is used, or there is an idiosyncratic implementation of a thesaurus management within the local collection management system or within a mediator component.
- Some libraries agree to use a foreign (typically English) thesaurus, as e.g. LCSH, ACM subject headings etc., thus giving poor support for the local language and any further specialization to local needs.
- Few systems support automatic query term expansion, in the same or to other languages.

- Evolution of the thesaurus on an external tool and consistent migration of new or changed terms **into** local collection management systems and into other external thesaurus tools is typically not foreseen.

Hence valuable information remains inaccessible, and retrieved information is incomplete and inconsistent with the request, at least by far more than necessary, even though thesaurus formats are standardized since a long time, and thesaurus merging and thesaurus federations are investigated by several groups.

The Architecture

This article describes a proposal for an architecture, which should be able to render integrated terminology services on large federations of digital collections in a scalable and manageable way with similar quality as currently on some local systems. It builds on the experiences and system developments from several co-operations of the author [Doer96], among which the AQUARELLE project (see e.g. http://aqua.inria.fr, [Doer97]) is the furthest-going in this direction, and conforms with more general integrated intelligent access systems ([Wied94] and many others). The equally important question of knowledge acquisition and the effective creation of thesaurus contents is deliberately not addressed in the following (see e.g. [Doer97b]).

We regard the whole as a heterogeneous database problem with vertical distribution and partial data replication. Collection management systems and thesaurus management systems overlap on the terms as shared identifiers. As thesauri undergo slow changes, and adaptations to changes may include manual actions by autonomous groups, partial data replication of thesaurus contents is efficient in a large network. For optimal results, the terms used for asset classification, in a search aid thesaurus and in the experts' terminology should be consistent. This led us to a three level architecture of loosely coupled components cooperating within an information access network: (1) vocabularies in local databases, (2) local thesaurus management systems of wider use and (3) central term servers for retrieval support. (See figure).

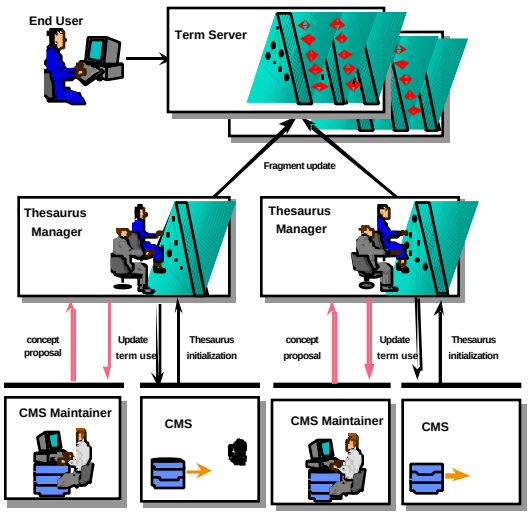

Figure: Terminology service at three levels

We foresee a separate thesaurus manager to which the vocabularies of several local databases can be loaded, and in the sequence organized as thesauri ("authorities") by some experts, following variations and extensions of the ISO2788 semantic structure. In addition, standard external vocabularies can be loaded to it. These authorities may be specific to one database, a user organization, or a whole language group. This thesaurus manager is locally consulted to enforce vocabulary control for classification and query formulation. This implies a knowledge (metadata) about which fields comply with which part of the authority. There are some advantages. A modern organization shares terminology within a group of heterogeneous applications – accounting, warehouse management, decision support, product development etc. Thesaurus software needs quite different structures and is in general maintained by other people than the individual application.

The local vocabularies and terms, that are already used for classification and are under the control of the thesaurus manager may need updating with the changes done at the thesaurus manager. This must be a semiautomatic process, which will be supported by a tool that compares the changes in the thesaurus manager and the use of terms in the local database, and makes proposals for the least changes to be made in the database. If e.g. a term is simply renamed, the change can be made automatically. If a concept is expanded into two, e.g. "tomography" into "CT" and "NMR tomography", a user may need to decide which one applies in each case.

Therefore the thesaurus manager must maintain a history of semantic changes from release to release. The same data can be used to translate or transform terms in a query formulated according to the new thesaurus release against a database consistent with an older release. The retrieval agent will know e.g., that the database understands only "tomography" instead of "CT". The returned answer is larger than expected (reduced precision), but not more imprecise than the database could answer anyhow. This is an essential element of this proposal. Without additional overhead for the thesaurus maintainers, we optimize the update of classification and allow for the simultaneous use of different thesaurus releases without loose of recall and precision.

Typically collection maintainers want to use more terms, than the managed terminology ever will contain (so-called local terms). If the local terms are related in the collection system with the next broader term of the authority, the collection system will give correct answers on the local and the broader term, and this relation can be updated as above. Even more, local terms could be automatically submitted to the thesaurus manager as proposals.

Term servers are used as search aids and need a limited management. Term servers are loaded with multiple thesauri from the local thesaurus management systems. Equivalence expressions will be introduced between the terms in the different thesauri, which on one side help users to select correct terms for databases using authorities he/she is not familiar with. On the other side retrieval agents should be able to make such "translation" automatically, in case many different databases are addressed simultaneously. Therefore term servers may give access to the necessary metadata about which data source uses at which field which part of a thesaurus. Term servers must be updated with new releases of local thesauri maintaining referential

integrity of the equivalence expressions. Again, the history of changes in the thesaurus managers is the key to that.

Equivalence expressions are not easily found, and their number increases with the possible combinations of thesauri. Scalability can be maintained however, if term servers are cascaded to support multiple translation steps (see e.g. [Dao96]). Of course, the precision will decrease over multiple steps. A suitable definition of the equivalence expression can allow to maintain the recall. If wanted, the precision may be maintained, but eventually empty answers are created. [Doer98]. Equivalence expressions may also be created by statistical and language engineering methods. The use of "interlingua concepts", as e.g. the European Education Thesaurus, reduce complexity as well. One may even think of cost-models for the cheapest translation.

In comparison to the size and complexity of multilingual thesaurus structures and management, the information needed to be transferred to classify a data item or to translate a query is small and has relatively simple structure. It is therefore quite efficient to access terminology resources through WANs. To standardize the interfaces between term servers and retrieval agents (a), between thesaurus management systems and local databases (b) and between term servers (c) may be more successful than the current effort to standardize the thesaurus structure itself in order to achieve interoperability. If such standards exist, thesaurus structure is only a matter of agreement between term server and thesaurus management system. In the sequence, more creativity of implementers can be tolerated and rapidly more intelligent services can be provided.

Within AQUARELLE we have enhanced our thesaurus management software SIS-TMS [Doer98b] to support cooperative development of multilingual thesauri. The system features release procedures with history of changes as described above and can also be configured as search aid thesaurus. By graphical visualization it allows for excellent understanding and control of complexly interlinked terminology structures. The solution consist of independent components with open interfaces. The system has found very good user response so far. Other groups have alternative and complementary systems necessary and useful in such an environment.

What to do now

We advocate for international cooperation to implement and experiment with a full architecture as described above. It means

1) To provide solutions for term translation within complex query expressions, e.g. in Z39.50 protocol requests.
2) To develop the methods to manage the consistent operation of such federations and to investigate questions of quality of service.
3) To develop the appropriate network managers, retrieval agents, and data exchange utilities.
4) To define the three basic open interfaces ((a),(b),(c) above). These interfaces become really valuable, when an open communication protocol can be established, which allows to combine freely thesaurus management systems and their term servers with retrieval agents and collection classification systems at an international level.

These activities must be harmonized with improved methods for knowledge acquisition and term correlation (as e.g. aimed at by the TELEMATICS project "Term-IT), and with developments on schema integration and mediation. Under these conditions we believe, that the separation of the terminology service from the retrieval agents and collection management systems into an overall federated architecture as proposed here, has the potential to make effective retrieval from a large number of multilingual data servers a reality. As well, the provision of a correlated terminology rather than reclassification of all data can make even highly specialized data widely accessible.

References

[Dao96] San Dao, B. Perry, "Information Mediation in Cyberspace: Scalable Methods for Declarative Information Networks", in: Journal of Intelligent Information Systems, 6, pp131-150, 1996.

[Doer96] M. Doerr, "Authority Services in Global Information Spaces." *Technical Report, ICS-FORTH/TR-163*, Institute of Computer Science-FORTH, 1996.

[Doer97] M.Doerr, I. Fundulaki, V. Christofidis, "*The specialist seeks expert views: managing digital folders in the AQUARELLE project*", in: Museums and the Web 97: Selected Papers", Archives & Museum Informatics, Pittsburg, 1997. ISBN 1-885626-13-4.

[Doer97b] M. Doerr, "Reference Information Acquisition and Coordination", in: "ASIS'97 - Digital Collections: Implications for Users, Funders, Developers and Maintainers", *Proceedings of the 60th Annual Meeting of the American Society for Information Sciences*, " November 1-6 '97, Washington, Vol.34. Information Today Inc.: Medford, New Jersey, 1997. ISBN 1-57387-048-X.

[Doer98] M. Doerr and I. Fundulaki. "A proposal on extended interthesaurus links semantics." *Technical Report* ICS-FORT/TR-215, March 1998.

[Doer98b] M. Doerr, I. Fundulaki "The Aquarelle Terminology Service", ERCIM News Number 33, April1998, p14-15

[Fosk97] D. J. Foskett. Thesaurus. In *Readings in Information Retrieval,* eds. K. Sparck Jones and P. Willet, publisher Morgan Kaufmann, 1997.

[Kram97] R. Kramer, R. Nikolai, C. Habeck. Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. In *International Journal on Digital Libraries* (1), pp. 122-131, 1997.

[Soer96] D. Soergel. "SemWeb: Proposal for an open, multifunctional, multilingual system for integrated access to knowledge about concepts and terminology." *Advances in Knowledge Organization,* 5, pp.165-173, 1996.

[Wied94] G. Wiederhold. Interoperation, mediation and ontologies. In *Proceedings of the International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogeneous Cooperative, Knowledge-Bases* (ICOT), Japan, December 1994, W3, pp.33-48.