Authority Services in Global Information Spaces

A requirements analysis and feasibility study

Martin Doerr

Technical Report ICS-FORTH/TR-163

February 1996

Institute of Computer Science
Foundation for Research and Technology - Hellas
Heraklion, Crete, Greece
martin@ics.forth.gr

Abstract

The global access to a variety of heterogeneous information sources world-wide gives raise to new requirements for advanced access methods and the conceptual compatibility between those systems. We regard the usage of thesauri as a key feature to achieve the logical integration. Thesauri are not seen as passive dictionaries, but as active, evolving components in the network, as "authority services", which are continuously adapted to the needs of the users and the systems requesting authority service. The integration and adaption requires communication and coordination between information sources, their maintainers, authority services and thesaurus providers. We investigate the feasibility from a computer science point of view, proposing a general architecture, and establishing functional specifications for the authority service on the level of data management, data exchange interfaces and user interface.

1. Introduction

Traditional libraries could base and target their retrieval structures on an environment with certain, known limitations :

- A public of users more or less known in ethnic and cultural composition, and often in specialization and distribution of interests.
- A collection of material explicitly targeted at certain information goals, either defined by the library institution, or indirectly through the response to requests from the specific public, or both.

- A collection, even though large and slowly expanding, of known size.
- Users were typically in a long term contact with a few libraries by subscription. Hence it could be expected, that they invest a certain amount of time to get familiar with the individual organization and retrieval structure of the library.

Other collection maintainers, as museums, are usually targeted at a wider, often international, unspecialized public on one side, and at a specialized scientific community, by far more focussed in their information goals, on the other side. Nevertheless, most collection maintainers work traditionally under similar conditions; i.e. the access structures to their material, as systematic catalogues, keyword indices, paper archive organization etc., and the topical organization of the physical collection store, are tailored to the collection, and more or less grown in close interaction with a specific user community.

There is a tremendous amount of experience, knowledge, and research investment in these structures, of specific and general nature. Nevertheless, the "Information Highway", the global access from any point in the world to information sources from whatever location, provider, specialization and culture, changes essentially the preconditions as outlined above. In the following, I want to outline the characteristics of the new situation and propose an approach to meet it properly. Whereas the environment changes more in quantity and scale, rather than in quality, the solutions require new mechanisms. These solutions should ideally be elaborated in close cooperation of librarians, museologists, computer scientists and, may be, cognitive scientists, in order to exploit optimally the valuable expertise of each field. Here, I concentrate on the computer science aspects.

This paper reflects experiences from our own work on repositories [3],[5] and knowledge representation [2],[4],[6], several projects on cultural information systems and recent projects on library systems and distributed information systems. Valuable information provided a meeting [8] in San Diego, CA Oct ’95, and other informal contacts with the Getty Art History Information Program, Santa Monica, CA. Some design ideas presented here are similar to [1],[12]. Only very recently, an industrial interest in the intellectual access to large, heterogeneous collections can be observed.

2. Problems of the global information space

At first, the sheer amount of data itself can become counterproductive to our information goals. Typically a user of an electronic catalogue specifies search conditions, until he gets a selection of a dozen or less objects. He continues browsing extended information about the latter, to select finally the "relevant" or "best", that fits his goal. A sensible and precise enough request, which renders some dozen objects on a local site, may render thousands or more in a global search (let us assume that all compatibility problems were solved). On such numbers, the subsequent browsing becomes impossible. Without means to issue requests of higher precision, he may get a poorer service than on a single, local system. This is indeed realistic, because many current access systems are tailored in their precision to the size of the local collection.

Second, we are confronted with a rapidly increasing variety of:

i) access methods, formats and terminologies,
ii) specializations in field and subject,
iii) user interests, educational and cultural background.

Fortunately, access methods and formats are subject already to intensive standardization efforts, be it Z39.50, MARC, SGML, data models of the CIDOC, recommendations of the CIMI project, or "application neutral" procedures as full text retrieval or "indexing by content", and significant progress has been already achieved. Other sources of heterogeneity, more on the conceptual level, are not so obvious to deal with, and our main concern is on these.

Terminologies may be different due to the different natural language group. But even within the same natural language, specific groups may develop different terminology for the same field. Even well agreed on scientific terminology of a certain domain can become surprisingly ambiguous when used outside of the context of human conversation or a written document. Take e.g the meanings of "regression" in economy, biology, statistics, even "radius" has two meanings (or more ?) in mathematics. Global access deprives us more and more of any context that can be reasonably assumed by the receiver of a question as background of the questioner or vice a versa.

Specializations confront us with the problem, that a user may not have the knowledge to express a search request in the correct expert terms. He may e.g. not know the latin name of a "teddy bear cactus", whereas the expert may not be able to specify his request on the same species with sufficient precision in a "popular index". Specializations further hinder interdisciplinary or interdomain searches, and in general, all requests not conformant with the goals of a certain specialization. Many objects however have more than one relevant aspect. In smaller systems, users can develop a considerable skill to guess where to search, by trying to reconstruct the documentalist’s thoughts. Such strategies will suffer under the loss of context in global access. Last, specializations develop rapidly, confronting us every day with new concepts.

The aspect of increasing variety of user interests, educational and cultural background inverts the above considerations towards the service a single system must provide. On one hand, each user is confronted with systems made for different users, as discussed above, on the other hand, each system is confronted with different users.

Simultaneously, the amount of time the user can afford for learning specific features of each system is getting less and less. In other words, without further measures, the plane global accessibility of information sources will hardly fulfill our expectations, it can be even counterproductive. By sure the recall, i.e. information retrieved against the amount of relevant information available, will in many cases tend to zero while the number of attached information sources increases.

3. Requirements

Besides the obvious need for standardization of access methods and formats, terminology must be standardized. The experience in the field shows, that the usage of structured controlled vocabularies, especially semantically organized thesaurus structures, allows for high search precision. E.g. an experiment in Germany on accessing European Parliament On-line Query System (EPOQUE) demonstrated that the utilisation of a thesaurus in the first language of the searcher almost doubled the success in the retrieval of relevant documents. Previous experience of searching was comparatively less important [11].

Even automatical methods (full text retrieval or retrieval by contents), based on statistics, can be ultimately improved by incorporation of thesaurus structures, as e.g. the system MISTRAL from Bull, France. Moreover, structures as ISO2788 or USMARC ISO2709 seem indeed to allow for rapid orientation of even an unexperienced user in a large term space, if made with sufficient care [8]. They seem to be the first successful means to present concepts and their respective context in an intelligible way, a basic requirement in a global information space, as outlined above.

Personally I believe, that the thesaurus structures in the near future will evolve into a rich refinement of the current link types, and metastructures for dynamic term generation, rather than changing to other semantics than the current link types (see paragraph 4.3.3). I regard it hence worthwhile, to analyze the consequences of using current-like thesauri in large-scale, assuming that the concept will still be valid, when this becomes reality.

I shall not in the following distinguish specifically between "authorities" and "thesauri", as well as between "terms" and "concepts descriptors", for readability purposes. A preciser definition of the notion of authority used here could be: "multiply hierarchically or simpler organized authority files". What I am interested in here, is the handling of such systems in general, the complexity they introduce, and the actors and systems involved. Therefor the precise internal organization, and the precise definition of the elements, have no impact on the following discussion, and I shall not make such distintions.

The construction of thesauri requires a tremendous investment of human work. To become really useful, a "critical mass" of some hundred thousand terms must be gathered. The success and broad acceptance of such investment, as e.g. the LCSH (Library of Congress Subject Headings) in libraries world-wide, fully justifies the effort. The solution is however not, to provide once and forever a "good" thesaurus for each domain. First, even very specialized thesauri overlap in their more general notions, or cannot be used without terminology of neighbour disciplines (e.g. chemistry and physics, biology and chemistry, computer science and mathematics etc.). Second, slowly but continuously new concepts are created. In order to provide optimal service, and to preserve the investment, the tendency is to built dynamically on few, large, integrated knowledge structures. Moreover, it seems to be easier to develop a thesaurus in some natural language from one example in a foreign language than from scratch.

It means, that a thesaurus in the future will include multilingual terminology and an increasing number of domain specific subdivision, ever growing by new concepts. Besides for reasons of economy, such large, homogeneous and consistent knowledge structures seem to be apt to counteract the initially described increasing disorientation and loss of context in a global information space. What are the problems of such a development?

Technically, the sheer size poses various challenges of storage and maintenance under global distribution, of both, users and providers of various expertises. Already now, most library systems will not incorporate the full LCSH, which is "only general purpose". Incorporation of specific terminology of the major sciences will easily reach millions of terms. Even the promotion of updates into the fields using standard terminology in some object record is not solved in current systems.

Large uniform structures are always controversial. On one side, we need the compromise of "common" understanding, and on the other side, we need the diversity and individuality for evolution. To my opinion, we must regard thesauri really as a help for retrieval only, a compromise of common understanding of concepts rather than an agreement on world-wide uniform definitions of concepts, and thesaurus providers should take that into account (e.g. by defining concepts eventually "wider"). But beyond that, the structures employed should be flexible enough to capture alternative concepts, their correlations, and eventually future semantic changes of the structure itself.

Summarizing, the problem is to maintain a preferably global knowledge structure (at least thesaurus structure), which is consistent internally and with its client collection management systems; which undergoes rapid growth of contents, and slower logical evolution. In other words, to provide an evolving global authority service. In the following we propose a series of services and components to satisfy the previous general requirements. These services may be gradually implemented towards the full "ideal" system.

I shall use in the following the term "collection management system" (CMS) in the widest sense, comprising traditional library software. I shall call "authority service" (AS) the service, to provide the CMS and CMS user in all appropriate functions with terminology of wider agreement, and informations about terminology.

3.1 Considerations on Services

A set of software tools must be developed, which deal with the various aspects of controlled vocabularies from the creation and the provider, until the end-user and his systems. As outlined above, the "authority" is not so much an authoritative declaration, but a democratic compromise of loosely organized global communities, with heterogeneous environments. Hence these tools must be open systems, based on a common understanding of their function, common data exchange and data processing standards. A global cooperation as required here can only be achieved, if enough software providers can be invited to care for local support, local adaption, and the parts of the global distributed services.

Thesaurus providers need development tools. Characteristic tasks are the merging of specialized vocabularies in to wider ones, control and reasoning on the logical principals of hierarchy organization, and the management of individual proposals for new terms, mainly from the user community. Only on thesaurus merging some research has been done. Most of the current thesaurus software is not more than an electronic notebook and dictionary. The most time-consuming task of thesaurus providers is the assessment of the acceptability and precision of a proposed term. It could be done ideally in a global communication through the network (CSCW). Few thought seems to have been invested on that. Even more, it seems that thesaurus providers should extend their role to coordinate with their expertise distributed consortia of highly specialized domain experts as codevelopers. The main concern of this paper are however the services to be installed at the user side: Thesaurus providers publish their products like (electronic) books. Thesaurus users want "authority control". They embed the terms in their data bases, and refer to them in object descriptions. That would work, if the thesaurus existed a priori. As outlined above, thesauri will evolve and expand rapidly. Without further help, it is absolutely impossible to compare some millions of objects against a new set of some hundred thousand terms, in order to update consistently the term references. The fields foreseen so far in the ISO2709. to denote changes between thesaurus versions are incomplete. Thesaurus provider must give more detailed informations. I shall elaborate this in more detail below.

Let us assume, that any migration from one thesaurus version to the next one is done automatically as far as possible, based on informations from the provider and the actual usage in the collection data base. Still a series of decisions must be done by humans, as e.g. if a term reference is still valid after a change of the terms scope (as expressed by the scope note). Statistical means may quite well propose solutions, but cannot replace the human decisions. These need time: until a person becomes available to make a decision, and then until he has decided.

On global scale that means in general, that never all collection management systems using a certain thesaurus will be up-to-date with the same version. In other words, not only multilinguality, but also the coexistence of versions forces us to maintain equivalence relations between thesauri, in order to translate queries according to each local vocabulary status. Necessarily some loss of precision will happen on the translation, but loss of recall should not, and that is the ultimate criterion for a correct implementation.

Summarizing, the thesaurus users need support to update their term references, to translate terms in global queries to vocabularies locally in use, and last not least mechanisms to access efficiently vocabularies, which exceed any sensible local storage capacity in size. As follows from the previous chapter, the users want to find and understand rapidly concepts, which implies the need for suitable (graphical !) user interfaces, which can be easily adapted to improved knowledge representation models.

4. Design proposal

Considering these requirement, we conclude, that the authority service becomes a data base problem in its own right. The basic structures are alien to the others appearing in a collection management system. Deep networks of indirectly related logical descriptors (e.g. narrower of narrower terms), against large tables of uniform records with statistic data (type, provenance, storage place, conservation status etc.). Information that grows slowly and changes seldomly, against frequent status changes (transfer, renting, trading, conservation etc.). Access following links against selection by combination of criteria.

The data exchange between the collection management system and the authority service is relatively simple, and mostly not performance-critical. Some terms are assigned to object record fields. Some terms are selected as query parameters (or, in both cases, some logical constructs in more advanced systems). Only for migration between thesaurus versions massive information must be exchanged. But even then, it is based on flat object-term relations, and will not be done frequently. The data exchange between thesaurus providers and the authority service, and eventually between a query generator and local vocabularies, can be by far more complex, especially when we are forced to use distributed systems due to the amount of data.

If data distribution is required, both systems need completely different patterns. Distribution of the frequency of specialized requests, against the physical distribution of objects, curators, librarians.

In this situation, a heterogeneous distributed architecture seems to be ideal. The collection management system is according to the state-of-the-art implemented on a relational

(RDBMS) or object-oriented (ooDBMS) data base. The authority service is better implemented on a semantic network, at least as long as thesaurus-like structures in the widest sense are used. It could be also more advanced KR-tools, if the performance requirements were met. The communication protocols between the collection management system and the authority service can be standardized, widely independent from any specific application, domain, and detail of knowledge representation (e.g. thesauri). The communication protocol between thesaurus providers, authority service and eventually global-to-local query generators can be standardized for respective KR model groups (e.g. derivatives of ISO2788 under a certain metamodel). I shall give details on that below.

The benefits of this approach are clear. Instead of implementing on any idiosynchratic CMS more or less authority control functions and logical structures, running after logical changes and increasing data amounts, the CMS can "plug in" the authority service. The latter can be harmonized with thesaurus providers and information network maintainers, and tuned to the network and local performance capabilities. Even more, CMS maintainers may change an authority service without affecting the rest of their S/W investment.

4.1 Communication with the CMS

4.1.1 Searching terms for classification

Curators and librarians assign terms to fields of object records. Terms may be selected from the user interface of the authority service, and sent to the CMS user interface as parameter. The user understands terms by suitable display of its environment in the thesaurus, but also by usage examples. For that sake, he would like to see quickly, if the term has been used already in his collection, how often, where and together with what other terms. For objects of a type, that is seldom in his collection, he may even want to look for the usage of terms in other collections which use a similar authority. In such examples, he finds further terms, he would like to understand better. It means, that he wants to send terms also back to the AS for clarification purposes.

Technically, these functions can be reduced to exchange terms between user interfaces, together with some term-type code and options. The receiving system can either be already in the state, to take the value sent as a specific parameter, as "zoom" to a controlled value list in usual RDBMS user interfaces, or it shows the value in a buffer/window, where it can be taken by any local command as parameter. I.e. the initiative, to issue a command remains in each system. Typical term-types are facets as subject, person, placename etc.

Such a communication is more application neutral and flexible, than issuing a series of specific CMS- functions from the AS and back. Respective type checking on both sides can provide the same security from user errors. We regard it as useful, that the CMS registers and assigns a local identifier to each used term. That improves performance, eventually the local system autonomy, and the communication with the AS. The AS may further have knowledge about the usage of terms in an attached CMS, storing together with a term either the local CMS-identifier or local usage frequency in the CMS or both.

4.1.2 Searching terms for retrieval

Users searching for information want to select suitable terms for queries. As above, they should be transferred from one user interface to the other, in both directions, for the same reasons. Information about term usage is still more important. Already on browsing through a term hierarchy, users would like to see if:

- a term is used at all in a collection,
- if a term is distinctive, i.e. not frequently used
- if a collection has rich material of the type he looks for.
- if a collection has no object of the type he looks for.

Especially the last point is usually not supported by current systems. They tend to display the used terms only. Consequently, the user can hardly decide, if what he looks for does not exist there, or is classified under other terms. He should see at least the unused siblings of each used term. In current systems it is very time consuming to find out, that you search in vain.

Another useful function of an AS is the term expansion, into sets of narrower terms, or "similar" terms - nearby in hierarchy or associated by some "related term"-link. Obviously, it is better not to list the broader terms in the object records of the CMS, but leave the expansion of broader to narrower terms to the AS, for reasons of consistency and maintenance. Especially in libraries, or image indexes, however, the usage of broader terms may indicate the level of genericity of the object, e.g. an introduction to the history of Greece. Respectively, the user would like to be able to specify how specific his interest is, which does not necessarily correspond to the specificity of terms used. The whole problem seems not to be completely understood, and needs further research. Fortunately, the communication protocol between the CMS and the AS is not much effected. In general, sets of terms have to be exchanged. May be in the future, level indicators will accompany the terms. It is an open issue, how absolute levels as in biological genealogies could be defined for term hierarchies. Most complexity of the problem however could be handled by functions of the AS and suitable term usage rules in the CMS.

4.1.3 Maintaining term consistency

Two phases must be distinguished. An initial phase, where a from the beginning uncontrolled vocabulary of a CMS becomes controlled, and the maintenance phase, i.e. the regular updates with new thesaurus editions. During the latter, suitable information on interthesaurus relations can reduce considerably the effort, as proposed here. Term comparison can be of an open-ended complexity, and can only partially be automated, as mentioned above- at least with current technology. Typically, each field, or field group sharing one facet (Persons - Places - Subjects etc) in the CMS is treated independently. The update process of the CMS always consists of three parts: Read out term usage from the CMS, compare the terms with the authority, and update the CMS.

Reading out term usage information can be on terms only, e.g. term-identifier, or, more advanced, term combinations per object identifier. More information may be helpful, but the more, the application independence of the communication protocol and its suitability as standard will be lost. The respective retrieval commands and output formats should be standardized.

The term comparison function could be part of the AS, or a third independent module. In the latter case, a similar access protocol as above can be defined for certain KR model groups. The benefit would be the ease of migration to better comparison algorithms, or the possibility to take advantage of a more specific module for the local CMS. The result of the comparison will be:

i) A list of accepted terms, unchanged or to be changed with another term.
ii) A list of not accepted terms. A proposal for a better term may exist or not.
iii) If term combinations per object identifier are available, a list of definite or proposed new term combinations per object, and a list of undecidable cases.

If the standard protocol should not foresee term usage per object, the latter functionality must be implemented locally in the CMS. If electronic texts or captions exist for the objects, statistical tools can be valuable to propose terms. It is in the responsibility of the local CMS administration, to accept or not proposed terms. Undecidable cases are ultimately subject to human decisions. For those, the system may propose alternatives or not. The system may present as smallest undecidable units groups of objects and terms, which need reorganization as a whole. All those interactive procedures should be implemented as work flow systems, to guarantee a processing complete and as fast as possible, in order to keep the number of world-wide used versions of any authority small.

After eventual human decisions, the update process results on the CMS side in the replacement of some terms in a table of used terms, and some changes of term references in fields describing directly or indirectly objects. As outlined, for some of these functionalities it is not obvious, on which side of the proposed three modules, the CMS, the term comparator, and the AS, they should be implemented best. This point needs further clarification, theoretically and experimentally.

4.1.4 Maintaining local terms

As it seems very clear however from comments of users [8] and from this analysis, at least in specialized collections, it cannot be expected, that all terms can be replaced appropriately by thesaurus terms. On the other hand, just leaving those terms in the object records would corrupt most benefits of retrieval by standard terms, especially the recall on those items by higher terms. A sensible solution seems to be, to introduce local terms, which are preferred by the local CMS administration, to the AS as such, under suitable broader standard terms.

Such a decision has some interesting aspects. On one side it is elegant, all terms are controlled by the AS. The distinction between local and standard terms is internal to the AS. It does not even introduce an additional complexity. It can be handled with the same mechanisms mediating between versions etc. , that will be described below. Even more, the AS may have built-in mechanisms to communicate those terms in large scale to thesaurus providers. That makes pretty much sense, since thesaurus providers will have to do directly with the authority services, and nearly "complete" statistics on the worldwide usage of some term may be obtained. (Of course within the group using such tools).

On the other side, the AS becomes even closer connected to the local CMS, somehow in contradiction to its global mediation role. It means, that a respective "authority server" will have several "attached" collection management systems, keeping their: local terms and their relations, local usage of standard terms, and may be also local identifiers of standard terms. Nevertheless, that does not violate the homogeneity of the internal structure of such an AS, and the informations about the local systems may be rather helpful in future systems for global query routing. The basic decision is between a distributed system, that actually acts as a whole towards the user, and a consulting system, which does not deal with consistency. Distributed systems necessarily share informations between communicating modules, "their common language". "Consulting" systems are mostly current practice - paper or CD editions read by the user optically only.

4.2 Interthesaurus relations

I call here relations, which connect a term from one thesaurus to a term in another one "interthesaurus relations". They can serve the migration from one version to another, and the "translation" of terms from one to another thesaurus simultaneously in use. The latter includes the determination of indexing capability of a target system, degree of narrower terms analysis etc. They may further be links to specialized subhierarchies, for dynamic loading.

4.2.1 Migration between versions

The main assumption is, that any user community can for any collection object agree on, if it is an occurrence of a certain thesaurus term or not. In other words, that the scope of a term is clear for a commonly known object set. That does not mean anything for the natural use of the term, and personal preferences of a user. It is a compromise for communication purpose. It does not mean, that there is a term for every object. Human concepts are fuzzy. In boundary cases, a preference to recall over precision should be given, i.e. a slight widening of the scope. It means however, that we can compare terms between different thesauri, be it versions from the same provider, or of different origin, or in another natural language, with respect to their scope. If the above assumption is acceptable, the rest is a mathematical exercise outlined below.

Consequently, migration procedures as defined above need as input all changes of scope, and not just rename and split into two (ISO2709) We need links, that denote the distribution of occurrences from one term to a set of other terms, not just one, i.e. a 1-n relation. Imagine, e.g. the change of a guide term criterium like <bridges by rivers> to <bridges by construction periods>. For the bridges of each river, a link to all periods of bridges must be drawn, at least after the first known bridge there (for the meaning of "guide terms" see [10]). These links must be given by the thesaurus provider. The information is anyhow available in the development process, it should just not be forgotten.

We distinguish complete transfer, which leaves the previous term obsolete, from partial transfer. If the semantics of the broader term relation are understood suitably, changes of scope of broader terms can be concluded automatically from narrower terms changes. Besides that, only complete transfer from one term into one other can be handled automatically. It corresponds to a rename (target term is new), or a merge (target term existed). The term comparison process will take the terms, which "loose" occurrences, find the respective objects, and propose the new terms, i.e. the total of terms, that took over occurrences from those previously used for the object. A curator or librarian must decide, which one is valid. More advanced term comparison procedures may take advantage of term combinations assigned to an object to narrow down the candidates for new terms. E.g. "vase", "burial utensil" may be replaced automatically, when "burial vase" is introduced as new term.

It should be mentioned, that only the introduction of a new facet may lead to absolutely new terms. For all other cases, the above procedure is complete - in particular a new term/concept takes its occurrences from some broader term at least. Typically however, siblings are affected. Interesting enough, transfer of occurrences between terms may happen by nothing else than a change of a scope note, without any change of the terminology at all. Imagine, the term "chairs" to be introduced as first and only narrower term under "furnitures". It will take its occurrences from "furnitures", i.e. exclusively objects so far classified as "furniture" are candidates to be reclassified as "chairs". Imagine further, the scope note of "chair" referred the chairs having backs. Next, the scope note is changed to include "stools". As a result, another set of occurrences is transferred from "furniture" to "chair".

The same procedure applies not only to object descriptions, but also to the local non-standard terms, and their relations to standard terms. As special case the thesaurus provider should notify, if such a term became accepted, eventually under another name and position in hierarchy.

4.2.2 Term translation and AS communications

Obviously, the above mentioned relations are equivalent to those of the ISO5964 (Guidelines for the establishment and development of multilingual thesauri) equivalence relations. The difference is only the intention - the ones defined here are thought to be temporary, until the migration is done, whereas the ISO5964 ones permanent, to adapt the query parameters and not the object records. Consequently, the same translation procedures for query terms can be applied between version (and local terms) as between languages, as long as they are simultaneously in use. If the respective relations are complete, and the thesaurus of appropriate quality, recall is always preserved under term translation. In certain cases, precision may be preserved, under loss of recall, as a user option. If no other source of information than interterm relations is used for the translation, only in the trivial case of renaming recall and precision is preserved.

We can think of networks of pairwise related thesauri, that allow query propagation, loosing slowly precision under each transition. One AS must incorporate as "contact point" at least the referenced term identifiers of the neighbouring thesauri. Each AS "knows" by suitable declarations with which thesauri its attached CMS are consistent, eventually parts of the CMS. "Contact points" can exist also to specialized vocabularies, vertically, which may be loaded on demand. Probably a global identification scheme for thesauri, e.g. a prefix to identify to which thesaurus a term belongs, may be necessary.

For such a scenario it seems to me realistic, that global systems with consistent term usage can be built. On one side, it allows a graceful integration process, and is tolerant to time-consuming adaption processes. On the other side, it will become more and more stable in the upper (broader terms) part, never requiring completeness.

As becomes clear however from this discussion, the communication between authority services depend by far more on the KR model than that between the CMS and the AS. The design of a standard protocol should take expectations on the evolution of the KR model in the next future into account. There are several research issues about the creation of multilingual thesauri and hierarchical compatibility. I can imagine, that requirements as presented here, can also help to understand better the nature of term hierarchies for retrieval.

4.3 Internal issues

The user interface of the authority service for term look-up is rather demanding. It must serve a very fast orientation in the term space and understanding of a displayed term, especially for non-expert users with arbitrary educational background. Orientation means, that the user can easily find out, in which branch of a large hierarchy he has to look for a specific term, but also the security, that he did not miss another branch also relevant to his request. The understanding of terms is supported by their environment - broader terms, related terms, neighbouring concepts - by definition - text, image - and by example - object description, image, sound.

Few of the current thesaurus software on the market seem to be based on a thorough analysis of the processes users like for finding out terms. Graphical representations of term relations are often advantageous. Our experience indicates, that term relations can be better understood in larger views, with siblings and neighbouring levels. Many terms relate to concepts, that are not adequately described in words, especially in fine arts, technology and natural sciences. Graphics and multimedia techniques can be very helpful.

Not to overload this paper with too much detail on this aspect, I shall enumerate the most important requirements, we could establish so far from various information sources, including direct interviews of users, at the Museum Benaki Athens [7], at Siemens-Nixdorf, Berlin, in the framework of the project ITHACA, and others.

4.3.1 Presentation

- A "global view" on the whole term space, with indication of the current position. We could not yet objectify this requirement, even though everybody wants it. Networks are not "flat" in general, nor is there a clear notion of "major cities, land and sea". May be the frequency of occurrence and genericity ? An interesting research issue.

- Alternative textual and graphical presentation of any view, i.e. a connected part of the network. Real graphs from vocabularies are often unbalanced, sometimes flat, sometimes deep. The readability depends heavily on the suitable presentation mode for a certain structure. Users differ in their preference and perception capabilities too. Mixed modes may be even better. They must be automatic and fast. Users recognize known graphs in tenths of seconds and issue the next command.

- "Moving views", like a magnifying glass moving over a large map. It means the possibility to extend a view by the next level of a hierarchy or connection type, without loosing the previous. Users loose orientation, when the display order of the same information changes. On browsing, they would like to see, how one view "emerges" from the previous. Rearrangement as gradual (!) motion on the screen may preserve the orientation.

- Display of environments of a view. A view, showing in a certain depth some relation, as narrower terms, related terms, or whatever, should be "surrounded" by indications of information next to it. These may be the existence and number of further narrower and related terms, all broader terms or other descriptional elements of terms in the view. Such information clarifies the identity of a term, the placement of the view in the larger context, and indicates directions for further search. It must be user selectable, not to clutter the display. [Ben Shneiderman]

- Multimedia data. Storage and loading of multimedia data is computer resource intensive. They should be carefully selected according to their usefulness for the user. Multimedia data have more importance for the thesaurus developer and the curator, i.e. for the proper classification. They are very important for specialized, visual, geometric or accustic concepts, may be also for dynamic concepts. As modern user interfaces demonstrate strikingly, even small icons can be more informative than equivalent words [9]. An interesting problem of multimedia data is, that they usually show an example, not the variance of a concept. Sketches of biological species often mark the invariant features for that reason, or comment them with text again. Note that many biologists prefer sketches over photos! It means, that the acquisition or creation of suitable multimedia data is not trivial, and needs some investment.

4.3.2 Functions

The user interface functions must support direct term selection, navigation, and term environment exploration. I shall not mention requirements for the handling of the various communication functions and update procedures, as defined above. They require either direct mappings to windows, buttons, etc. by state-of-the-art means, or need again term space exploration.

Direct selection means to find preferred terms by typing a word, or by matching of a character sequence, eventually of a phonetic equivalent. Many good algorithms for partial matching do not scale well, i.e. lack performance for very large sets, whereas precise matching is fast on arbitrary sets. Designs may be better, which do partial matching in suitable parts of a hierarchy. Users will always know some general facts about a term they are looking for. If these conditions can be evaluated first in order to restrict the search space, and indexes built to support partial match are able to make use of it, the overall performance and functionality can be improved.

Navigation should be possible with the same ease forward and backwards along all relations. We could not identify so far preferences in direction in searching a hierarchy. Users make smaller and larger cycles to explore branches, abandon them, go deeper. That back and forth usually dominates over the initial general-to-specific direction. Users would like to return back to the origins of those cycles, which are nested in general. Most current systems do not show or exploit, how "bookmarks" of "back" moves (e.g. NETSCAPE) are positioned in a hierarchy of nodes. It should not be difficult to define a "back" by a direction "up" in hierarchy, instead of blindly following all loops the user previously made, - in inverse direction.

Our experience indicates, that users would like to hold more than one screen or view on the display, to recall in which branch they are, to compare, to iterate over another property (e.g. display in parallel some scope notes of a subhierarchy). Some "intelligence" is needed, to assist a user to keep what he may need, and not to "drown" him in a multitude of windows. Several questions in this context can be still regarded as open issues.

To our opinion, navigation and the exploration of the environment of a term needs more flexibility to select properties, attributes or link types. It should be possible, to expand a view dynamically by some attributes. That seems to be more a decision of the moment in search, than a user preference, as many software providers assume (by "user profiles" or "predefined views"). We expect, that in the future a multitude of hierarchically organized link types will be used - see only the ambiguities between "broader term" and "related term". User interfaces should be prepared for that.

4.3.3 Logical data structure

Any term structure organized by some meaning must be seen as a more or less developed knowledge representation. The step from a ISO2788 thesaurus to an object-oriented semantic model is not too big. There exists an overwhelming literature on knowledge representation from many years of research to develop the artificial intelligence, and it makes no sense to refer even to the main research directions here. More, an estimation of what are the most urgent needs to extend the current models may help us, to define a successful system for the near future. Obviously many sensible solutions can be taken from respective knowledge representation systems.

In contrast to e.g. expert systems, an authority service neither needs to guess nor to predict (at least for the time being). It has to deal with very large data amounts, more like natural language understanding systems. So the structures must stay as simple as possible. Someone has to pay also for the data acquisition. Nevertheless, I propose to motivate thesaurus or authority structures more with methods of knowledge representation and mathematical logic. Most literature on thesaurus merging take the structure as granted (especially the decision to monohierarchies) and leave many riddles. On the other side, authorities seem to become together with natural language understanding systems the largest knowledge bases at all, tested and retested by many many users. So Artificial Intelligence can learn at lot from them.

For the link system, we need the possibility to specialize the current types. Broader terms should not be mixed with part relations e.g., related terms may be a producer-process-product structure etc. Many broader terms must be allowed, things have more than one aspect. Broader terms should not be mingled with user guidance. Users may like to follow narrower terms, but they also would like to associate a context, e.g. aeroplanes - design, aeroplanes - social impact. Some statistically determined "main" associations could be very helpful, and any research on human associations.

The "producer-process-product" example bears another important aspect. Facetted schemes are thought to provide means to create dynamically concepts by term combination. Obviously natural language is by far more flexible in that, which does not mean, that it reaches always the necessary precision for an authority. Moreover, combinations and single terms may appear in analogous cases in natural language, as "carpenter" and "shoemaker", making rules difficult. Anyhow, some more semantics are useful, e.g. a "portrait-shaped-vase" is not a "vase with portrait". A very careful and practise-oriented analysis of such problems can be found in [13]. To our opinion, meta-structures can solve efficiently many such problems, without introducing more complexity [2],[4].

5. Conclusions

An authority service should and can be implemented as a database of its own, federated with the collection management systems using its services and other authority services. It should not just be an Authority Control, a kind of constraint enforcement, but an object in the sense of object-oriented programming, providing a knowledge service. Such software should be harmonized with the authority providers as an integrated service, and the integration with the CMS should be based on a application and platform neutral protocol.

It seems realistic, that a kind of global term consistency can be achieved, providing the quality of service or better on a global level, which a trained user has currently on a local system. Global consistency cannot be enforced. The idea is rather, to create flexible federated systems, that provide well defined degrees of consistency, and tend to evolve to more consistency. There are however several research issues to be solved.

There seems to be not much awareness about the complete amount of the problem, neither from the users, nor from software providers. We propose an attempt of the respective interest groups for the necessary standards. Thesaurus structures seem to be a sensible base to define standards for the near future. The interesting point is, that the vision presented here relies heavily on wide compatibility to come alive. On the other side, it may be much too early in this phase to agree on standards, which will delay the solutions. With this paper, and some related projects of our group, we want to start the discussion.

6. References

[1] P.A.Bernstein, U.Dayal, "An Overview of Repository Technology", Proceedings of the 20th VLDB Conference Santiago Chile,1994.

[2] M. Christoforaki, P. Costantopoulos and M. Doerr, "Modelling occurences in cultural documentation", III Convegno Internazionale di Archeologia e Informatica, Roma 22-25 November 1995.

[3] P. Constantopoulos, M. Doerr and Y. Vassiliou, "Repositories for Software Reuse: The Software Information Base", Proc. IFIP WG 8.1 Conference on Information System Development Process, Como, Italy, Sept. 1993.

[4] P. Constantopoulos and M. Doerr, "An Approach to Indexing Annotated Images", Multimedia Computing and Museums, Selected Papers from the Third International Conference on Hypermedia and Interactivity in Museums (ICHIM ’95 . MCN ’95) San Diego, CA., USA, Oct. 1995, edited by David Bearman, Pittsburgh 1995, pp 278 - 298.

[5] P. Constantopoulos and M. Doerr, "Component Classification in the Software Information Base", in O. Nierstrasz and D. Tsichritzis, eds., Object-Oriented Software Composition, Prentice-Hall, 1995.

[6] I. Dionysiadou, Martin Doerr, "Mapping of material culture to a semantic network", 1994 JOINT ANNUAL MEETING, International Council of Museums Documentation Committee and Computer Network, Wasington USA, August 1994.

[7] Anita Kvamme,"Task Analysis in the Context of Human-Computer Interaction", Diploma thesis, Norwegian Institute of Technology, Facultty of Electrical Engineering and Computer Science, Trondheim, Dec. 1994

[8] E.Lanzi, "Points of View Forum: AHIP Vocabularies, Collection Management Systems, and Users", ICHIM/MCN Conference, San Diego, CA, Oct 11, '95. Meeting Report, AHIP, Nov 21, 1995.

[9] M.Lesk, "Experiments on Access to Digital Libraries: How can Images and Text be Used Togetther?", Proceedings of the 20th VLDB Conference Santiago Chile, 1994.

[10] T.Peterson, "Introduction to the Art and Architecture Thesaurus", second edition, Oxford University Press, New York, 1994, page 37.

[11] A.S. Pollitt, G.P.Ellis, I. HOSCH, "Improving search quality using thesauri for query specification, and the presentation of search results." Advances in Knowledge Organisation 4, 1994, pp382-389.

[12] J.Schnase, J.Leggett, D.Hicks, R.Szabo, "Semantic Data Modeling of Hypermedia Associations", ACM Transactions on Information Systems, Vol.11,No1, 1993, pp27-50.

[13] D.Soergel, "The Art and Architecture Thesaurus (AAT): A Critical Appraisal", Visual Resources, Vol. X, pp369-400, Malaysia, 1995