

Kim H. Veltman

Six Steps in the Development of SUMS

Unpublished, Toronto, 1996

0. Introduction
 1. Preliminary prototype with basic examples in one subject: education
 2. Prototype with detailed example on one topic of education: mathematics
 3. Integration of detailed examples on various topics
 4. Coordinate with individual content providers
 5. Integration of Static and Dynamic Knowledge
 6. Universal Reference Data Base of Names, Subjects and Places
 7. Conclusions
-

0. Introduction

The vision is simple: to have the ability to search for information and knowledge in libraries, museums, archives and other repositories throughout the world systematically so that one can organize, archive, retrieve, re-arrange, re-structure materials at will. This will include both static knowledge (the corpus of traditional facts stored in libraries, universities and academies) and dynamic information (ever changing data in new fields such as satellite photography, weather, news, stock markets etc.). The System for Universal Media Searching (SUMS) offers a first prototype of such a tool. At present it effectively functions as if one had a number of bookmarks and organized these systematically. This preliminary version has a number of limitations. It uses only two subjects (Leonardo and perspective). It is largely hard-wired. It cannot do automatic searching. It applies only to static knowledge. This paper outlines a series of seven steps or stages which will take this software from a hard wired prototype to a fully operational, multi-valent product that is largely automatic in its search strategems.

Step 1. Preliminary prototype with basic examples in one subject: education

Basic materials on education are collected manually and organized in terms of questions: who, what, where, when, how, why. These connections are hard-wired but words in the server can be linked to an on-line dictionary such as Websters and find that term automatically. The presence of the Z39.50 protocol means that any person listed in SUMS under "who" can be searched in the author catalogues of a Z39.50 library, museum or archive.

This is still limited access because there is no inherent knowledge about what is found at the remote site. For example, if the user looks for Leonardo da Vinci in the University of Toronto Library with this version of SUMS, they will find all titles that happen to be in

the library. They will still need to organize that material manually. To achieve this there will be some elementary templates.

Step 2. Prototype with detailed example on one topic of education: mathematics.

To show the full potentials of making detailed cross-referenced links on the basis of who, what, where, when, how, why, will require a labour intensive proof of concept by an individual. To achieve this John MacDonald (MSSB) will relate course work in mathematics to the curriculum in the academic year 1996-1997. To help him in this process, various additional templates will be produced such that the process requires very little effort on the part of a new teacher.

Eventually many of these links can be automated with the help of mature natural language and neural network tools. Rather than wait a few years before these are in place, it is necessary to make a seeming detour by doing the labour intensive work of creating working prototypes manually or semi-manually. These prototypes can then be used as a standard when testing the ability of natural language techniques which make claim to be able to create all the desired links.

Step 3. Integration of detailed examples on various topics

Step three will extend this principle to a series of subjects to show that the same principle becomes more powerful when this cross-referencing is extended across the curriculum. In addition to going horizontally across different subjects there is also a need to go vertically among different categories of knowledge: from a general corpus, to a curriculum, a course, texts, and tests. Entailed in all this is an integration of access not just to individual facts of knowledge but links between the training, testing and evaluation thereof.

Step 4. Coordinate with individual content providers

Once one can show the full potentials of this re-contextualization of knowledge, it will be important to enlist the co-operation of major content producers to organize their materials in ways that are compliant with the SUMS framework. To achieve this the various fields in the SUMS system can be made available to an institution in list form. The institution in question will then add their equivalent fields in a parallel list such that materials from that institution can readily be mapped into the SUMS framework. For example the SUMS framework has the heading "Name". The institution in question may call this "Author", or "Creator" or "Artist" or even "Person". As soon as the equation Name in SUMS = Person in institution A has been made, then the SUMS server will know where to look for the materials. While this represents a considerable advance towards automation it means that searches remain limited to institutions who are members of the club.

Step 5. Integration of Static and Dynamic Knowledge

Traditionally those concerned with static knowledge (the corpus of accepted knowledge which does not change except that it becomes larger) have remained very separate from those concerned with dynamic knowledge. Librarians have tended not to interact with news reporters, stock market followers and the like. If my search is about something that happened in the last 24 hours, it will often be enough to limit myself to new sources such as Reuters and the like. In some cases, when I ask about a city in Bosnia, knowing something of the history of that city will be extremely useful to me. So an integration of dynamic and static streams of knowledge is necessary. Names, subjects, places in the two systems need to be integrated. In terms of media this entails integrating information from video and film with knowledge from books. This will also bring changes to the way we look at traditional knowledge. Statistics concerning publications can be transformed into charts such that one can see patterns in the form of spread sheets.

Step 6. Universal Reference Data Base of Names, Subjects and Places

To become universal requires an ability to access libraries and museums at random, which means that this basic knowledge of fields be extended everywhere. This will require a whole series of further initiatives.

First, one will need to introduce standards for each type of institution, such that libraries all conform to one set of protocols. The Z39.50 protocol is an important step in that direction. The standards among different institutions will then need co-ordination. Groups such as the Research Libraries Information Network (RLIN) have been moving in this direction.

Second, a system of aliases for names is required. If I am looking for a name such as Hondhorst I will not always find that Dutch painter under the standard version of their name. In Italy he is catalogued as Giovanni delle Notte. We therefore need to have a universal reference base of alias names such that if I am looking for Hondhorst the system also looks under Giovanni delle Notte whenever it is in an Italian database. That which applies to persons (who?) applies equally to subjects (what?), places (where?), times (as in different calendars; when?), and even techniques (how?). The way of restoring a painting in Italy may be different than in France or Germany. I may think I am looking for technique A but need also to look under technique B. So a centralized database of references and aliases is required even though the knowledge concerning those references may well be distributed all around the world.

Initially one would begin with various tools that exist already. In the field of art history for example there are standard reference works such as Thieme-Becker's *Allgemeine Künstler Lexikon* and the Getty AHIP's Union List of Artist Names (ULAN). Such basic works exist in all the major fields. In addition, at the national level there are national biographies, national bibliographies etc. Similar tools exist in all the major disciplines. In some branches of science, notably chemistry, physics and medicine, very significant steps towards an international framework have already been achieved.

Third, there will need to be another layer, probably distributed, to deal with the details of individuals, subjects and places. Let us take the example of Leonardo da Vinci. When a user makes a request from a local machine it goes to one of the mirror sites of the centralized reference base to acquire all the variant names, dates etc. It then goes to the standard database on Leonardo to get all the names of his paintings, manuscripts, instruments etc. Equipped with this context it can interpret the details of the question and know in what locations databases on Leonardo are found and/or likely to be found.

Fourth, the centralized database will entail a series of layers. Sometimes the question at hand will entail an idea that changes from culture to culture and/or a concept that changes historically. The term "perspective" in the sense of "linear perspective" is "perspectiva" in Latin. But the Latin term could also mean "vision". In such cases a simple mapping of one term to its translation in another language is not enough. We need to trace the etymologies of terms, culturally and historically. Such searches are much more intensive, even if much of the searching can be relegated to an agent. This is a very different layer than one which is asking for the location of Paris on a modern map. It is a level that is very much associated with the frontiers of scholarship, which is precisely that sector of society which can most help us in gaining a better understanding of the frontiers of knowledge.

In addition to various levels of knowledge, what may be needed therefore is a series of kinds of searching. Members of the general public will be discouraged from searching everything to the deepest level unless they wish to pay for it. On the other hand any scholar or even any member of the public (an amateur in the old sense of the term), as long as they have a demonstrable field of research where they are likely to make some contribution, will be free to search however much they choose, much in the same way that entrance to the British Library is effectively free, but only serious readers have free access.

Fifth, in order for the universal reference base to become fully sensitive to cultural and historical complexity will require the inclusion of major classification systems systems such that one can move readily from the mental cubbyholes in one culture to those of another.

Sixth, one will wish to complement verbal search techniques with visual methods such as those being developed in the Query by Image Content (QBIC) software of IBM. In elementary applications this entails a user drawing a shape which is then searched for among the images in a database. Such a strategy makes the search dependent in part on the user's ability and skill in drawing accurately. In future it will be desirable to combine verbal and visual query techniques. For example, a user is interested in the *Annunciation*. They choose this theme in a verbal classification scheme such as Van der Waal's *Iconclass*. The verbal classification has linked with it a series of basic types of *Annunciation* in the form of thumbnail images. The user decides which of these basic types interests them: e.g. *Annunciations* in a closed room or in a garden. The system then searches for that type.

Seventh, there will need to be new projects to translate materials into a common language. At present there are a series of impressive projects to scan in materials from the great collections of the world. IBM has begun to scan in the complete texts of the manuscripts of the Vatican Library. At present their purpose in so doing is to make the contents of this great library available to scholars able to read the manuscripts in their original state. These manuscripts are in a number of languages: while Latin and Greek may predominate there are also works in Arabic, Hebrew, Sanscrit, Russian and many obscure languages such as Babylonian or Assyrian which are accessible to very few scholars indeed. It is true that many of these works have at some time been translated into modern languages such as German, French or English. But many have not. Hence the quest for complete accessibility will require a complete translation of all works into at least one world language. This will take many years, especially in a world where there are forces at work to destroy almost systematically the heritage of libraries and museums in some cultures (e.g. Tibet, the former Yugoslavia and now Afghanistan).

Conclusions.

While the notion of an electronic butler as described by Nicholas Negroponte is a very attractive one, the possibility of achieving search mechanisms that truly reflect the complexities of cultural and historical change is not nearly as simple as it seems. Moving from static to dynamic links is feasible, but will require at least a generation or two to achieve in more than a superficial sense.¹

In the United States the notion has emerged that one can own content. This may be feasible as long as content is defined in terms of Hollywood films or television. In the grand scheme of things this material is relatively small. In the case of the libraries and museums of the world, the amounts of material are far too vast and no single company can hope to own it all. Companies such as Microsoft which began with that assumption have predictably been ridiculed by the major cultural forces of Europe. Needed therefore is a different approach.

Having abandoned the illusion of owning the content, the major players might fruitfully begin with a quest to own the references to the content. To achieve universal reference tools on a global scale will require many more resources than any single company or group of companies could ever hope to muster. It will require institutes dedicated full-time to developing the methodologies to make this possible. It will require a consortium of major players and international co-operation linking both the public and private sectors. Some of these initiatives dovetail with the goals of the G7 pilot projects which offer points of entry into the kind of global approach that must go far beyond the G7 if it is to become truly international in scope. The six steps outlined above offer some clear steps for making that future which still seems distant to many come closer to reality.

Perspective Unit, McLuhan Program
5 March 1996.

¹ Looking back to the example of printing would suggest that a much greater time scale is required. Moveable print was invented around 805 A.D. in Korea, moved slowly to China where it was used to censor knowledge. (That the Chinese now want to use the Internet to censor knowledge thus comes as no surprise to an historian). It was not until 600 years after the invention of printing that Gutenberg had the idea of using it to spread knowledge. And it was a good 150 years thereafter before the full fruits of a corpus of printed knowledge began to manifest themselves. The electronic revolution may be speeding everything enormously, yet it is sobering to remember that computers were invented over a century ago and we are still talking of them as if they belong to the future.