International Research Information System: Support to Science Management

Barend Mons[2,4], Renée van Kessel[1], Albert Mons[3], Ruud Strijp[1], Bob Schijvenaars[2,4], Erik van Mulligen[2,4]
[1]Netherlands Organization for Scientific Research, The Netherlands
[2]Collexis B.V. Geldermalsen, The Netherlands, [3]Collexis Solutions, USA,
[4]Erasmus University Medical Center, The Netherlands

Summary

In response to the ever increasing complexity of international scientific networking, the Dutch Government through NWO, The Netherlands Organization for Scientific Research, has taken the initiative to develop a global information system for Research Councils with the working title IRIS (*International Research Information System).* Most Research Councils consider finding referees a frustrating and time-consuming process. Too often major resources are spent on maintaining websites, indexing research proposals, trying to find the perfect referee, and evaluating researchers and research institutes. Science Managers should ideally have access to information about referees from any country or discipline in the most effective way possible. This paper describes the pilot phase of the IRIS project, aimed at the construction and set up of an International System to share reviewers.

1 Introduction

1.1 Preamble

This paper will focus on the basic principles of a very specific form of knowledge management, being the optimal use of validated explicit knowledge worldwide to support the upward knowledge spiral through high quality scientific research. It will first give a minimal background description of the basic philosophy behind the system and the underlying technology, followed by a description of the plans to set up an International Research Information System (working title IRIS) that is designed to support the management of the increasingly multidisciplinary and international research arena.

Many previous attempts to set up networked knowledge resources have failed, and yet another one is therefore likely to meet with a fair deal of skepticism. However, the newest generation of Information Mediation (IM) technology and the advent of a growing validated body of information available via the Internet have created an unprecedented challenge, as well as a unique historical opportunity to make it *work* this time.

1.2 How is knowledge generated?

Before we embark on the description of a system to organize optimal exploitation of existing knowledge word-wide we will briefly reflect on some basic aspects of knowledge generation and try to draw some conclusions and lessons from that reflection.

Many failures of "Knowledge Management" initiatives have been blamed on either a fundamental lack of distinction between the various levels of knowledge components, namely *data*, *information*, *knowledge* and finally *expertise* and *competence* (Sveiby 1997).

Even the basic building blocks of information (data) are not free from human interpretation (Heisenberg 1976) as they can only be observed, measured and stored by organisms with at least a basic form of knowledge. This implies that knowledge generation is in essence a *cyclical process* (Tuomi 1999).

Knowledge as it exists in the heads of people can only be effectively communicated to other human beings after being made explicit in written or spoken language or in visual form. This is a crucially important basic assumption for knowledge management. After being made explicit in communicable form, *knowledge* has in fact been reduced to *information* with the loss of one level of complexity (human creative and associative power). It is therefore crucial to deduce that Information Mediation is not the same as Knowledge Mediation; if knowledge is indeed confined to people, effective Knowledge Management is therefore more complex than simple Information Management. An international system aiming at optimal exploitation of knowledge should therefore take into account the enabling environment for *direct human interaction*.

It should also fully encapsulate the notion that information as captured in natural language, introduces multiple variations in expression of the same concept. Different national languages can be used, but in addition jargon can introduce multiple synonyms for the same concept within one and the same national language. In addition, language introduces the homonym problem (multiple meanings of the same expression).

The human drive to share knowledge has also been assumed in a rather naïve way in some systems. Before success of a networked knowledge system is even a viable concept, one has to realize that the Immediate Return on Investment (mainly time) for all distributed content owners should be obvious. Both scales on the balance should therefore be addressed:

On the "entry side", the investment needed to make data and information "exchangeable" via the network should be absolutely minimized and duplication (filling in the same data repeatedly for different initiatives) should be banned. Such a reduction in time investment has both a technical (no forms) aspect and a networking (political) aspect: It requires an upfront collaboration between the major players in the field to be networked to avoid duplicative submissions wherever possible.

On the "output side" the focus should be on *immediate benefits* for the users who decide to share information. Scientists have been intensively trained to disagree *by default*, but there is probably at least one aspect on which they all agree: they hate the time investment is searching for relevant information, partners, meetings, (re-) writing applications in horribly complicated forms and filling in their registration details for meetings, applications and surveys over and over again.

Scientific policy makers and science managers on the other side of the table have similar time-consuming problems in finding the right referees, distributing calls for proposals to the right people and institutions, keeping their public information updated and to optimally inform scientists as well as the society about ongoing research.

If the output of a system where people register by sharing their information, would be the *immediate* return of relevant and validated information on people, literature, projects, meetings and other relevant issues based on the information provided by a registered user, including a subsequent alert function for "more like this" the incentive to join the network by sharing information would be optimal.

With the enormous explosion of data and information going on at present and the resulting drive towards international and multidisciplinary scientific networks all these aspects become orders of magnitude more complex than they have been for the past few decades. One major constraint is introduced by the advent of the World Wide Web as an essentially uncontrolled infor-

mation carrier with unprecedented, low cost publication possibilities. As a result, the user is confronted with massive amounts of information, of which much may be irrelevant and even wrong and thus counterproductive. *Validation* of information is therefore crucial and scientific publishers see their role in this area for the future as a crucial part of their core business.

Last but not least, players that have an immediate interest in keeping parts of the system validated and updated should be placed at the core of the network as they will have a radiating quality effect in their specific sector. In the case of the IRIS project national research Councils and comparable Research Funding institutions will be forming a core backbone.

In summary, the critical elements of an International Knowledge-driven Communication System should include:

- A clear basic understanding of the difference between data, information and knowledge and their specific interaction (knowledge being essentially confined to people)
- A technology to deal with natural language variation and jargon issues
- A clear and immediate incentive for content providers and evaluators/editors to join and remain active in the system
- A minimal need for "forms to be filled"
- A clear distinction between validated and doubtful information
- A strong network of prominent partners in the area from the very start.

2 The Technology

2.1 The Matching on concepts

The selected technology for IRIS was originally developed to match across jargon and languages in large, distributed text corpora. Collexis® is based on proprietary technology, originally developed in the public sector (Van Mulligen *et. al* 2000, http://www.collexis.com)

The basic theory behind the core technology is that, although humans communicate in explicit language, including many variations and ambiguities, the final aim of communication is to share "concepts". Concepts are in this context, the "real life entities" that constitute the reference framework of human knowledge. An effective Information Mediation technology should therefore search and match at the *concept* level rather than at the *word* or *term* level to enable cross-jargon and cross-language communication. Collexis® has realized just that. It is a web-based technology, essentially driven by pre-existing, validated knowledge as made explicit in thesauri and ontologies. A process of "pre-training" enables the system to accumulate all existing human knowledge in a given field based on the validated knowledge resources in that field before starting to analyze large amounts of information. After being "made smart" by incorporating validated human knowledge into its text analysis component, the abstraction engine is equipped with the most unique feature of Collexis®, namely the *immediate* "normalizing" of all textual variations of words or terms referring to the same concept to a unique concept ID, without the need of extensive cycles of machine learning. The abstraction component thus creates a *Conceptual Finger Print* (CFP), a numerical representation of the real content of full text through a list of approximately 50 concepts listed in order of their relative importance.

Queries can be either pre-computed CFP's of existing text (the "*more like this*" function) or natural language queries typed by users, which are abstracted in the same way as the content in which the search is done and again, all natural language variations of a concept will lead to the normalized unique ID of the concepts searched for. In addition the underlying thesauri allow conceptual searches with definition support as well as dynamic categorisation leading to "dynamic portals"

Accumulated CFP's from multiple publications or projects form dynamic *interest, activity* or *publication* profiles of scientists and experts (Interactive Scientist Information Card [*ISIC*]).

Hundreds of millions of CFP's can be stored on one server and can be compared to each other by vector matching in a matter of milliseconds. This matching process is language and jargon independent and can be used in conjunction with any existing local platform or database used by networked partner institutes. There is no need to change existing Web Based Information Systems, Document Management Systems, or local databases. Web access to the pages containing the actual information is the final step. Entire data collections can be fingerprinted at a speed of 250.000 pages per day per PC independent of the local data system used to present the original information on the user's screen. This new web based approach thus allows unprecedented search and networking properties across distributed and non-standardized data sets.

Collexis® also contains drivers to automatically detect words and terms in texts that are neither identified as irrelevant (classical and user-defined stop words) nor incorporated in the word or concept lists of the thesauri used to train the Collexis® application. Such words and terms (word combinations) can be presented to the application administrators or authorized users as lists of suggested concepts. The system has recently been expanded with homonym detection and disambiguation based on context (semantic laterality) and is currently adapted for the life sciences by the collaborating teams of Collexis and the University of Rotterdam. An ambiguous term like BSE, which can either mean *Bovine Spongiform Encephalopathy* or *Breast Self examination* in a medical context, will normally lead to two possible, distinct concepts. However even in a very short query (bse scrapie, or bse behavior) the system will already make the distinction, based on the contextual concepts used by the corrective technology (figure 1).

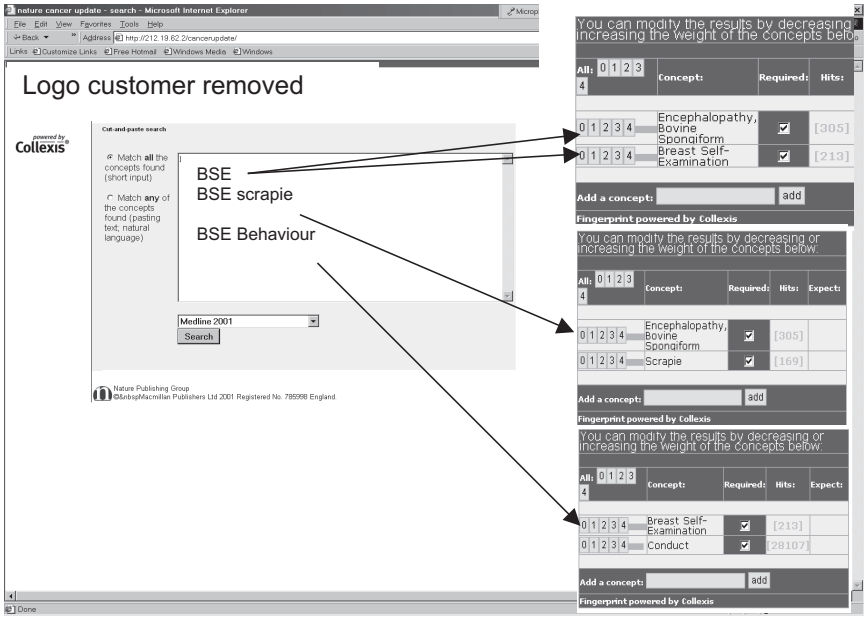


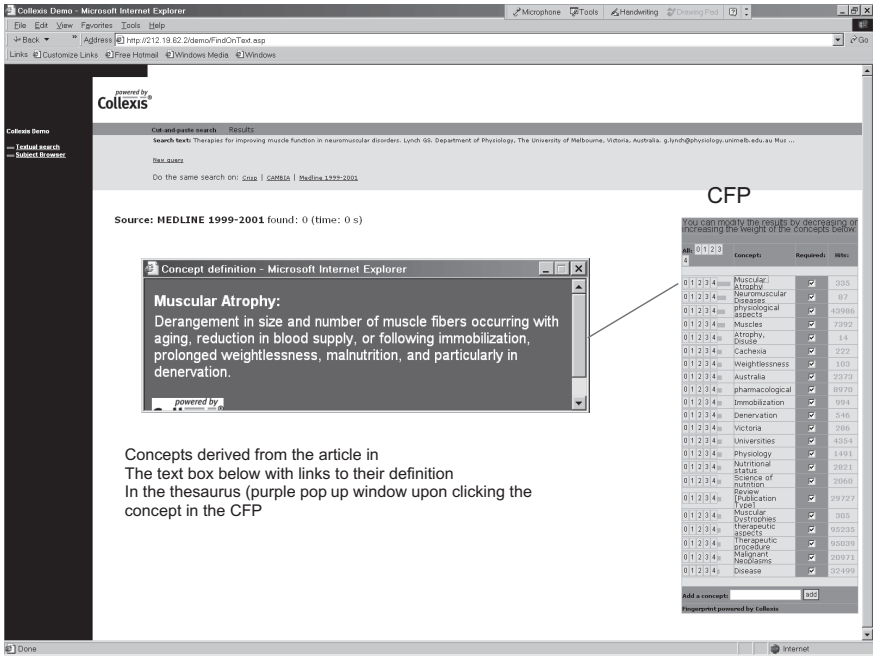
Fig. 1: BSE, either alone or with one context word attached leading to suppression

An additional value in Collexis® technology is that background information on concepts found in text can be used to inform the user directly about existing background knowledge about the concept. Figure 2. depicts what happens in the current public life sciences demo of Collexis when the user clicks on the textual expression of a concept in the CFP (in this case Muscular At-

rophy): A window opens which depicts the description of the concept as provided by the Medical Subject Headings (MeSH)

This example is also illustrative for the difference between Collexis and manual concept or keyword assignment: Many words and terms in the abstract depicted below are almost literally mapped to the corresponding concepts (blue) and some are associated and "interpreted" by Collexis (orange) in the CFP shown in figure 2. The coverage of correct concepts in the example is very high, which is the general trend. The precision and recall aspects of Collexis can be influenced by tuning the system to the need of individual user communities and applications.

Therapies for improving muscle function in neuromuscular disorders.



Lynch GS

Department of Physiology, The University of Melbourne, Victoria, Australia.
g.lynch@physiology.unimelb.edu.au

Muscle atrophy or wasting is a loss of muscle tissue resulting from disease or lack of use. This review examines recent pharmacologic or nutrition interventions for ameliorating wasting and improving muscle function in neuromuscular disorders. The information has application for treating the muscular dystrophies, cancer cachexia, weightlessness, immobilization, denervation, and disuse atrophy

Fig. 2: Definition support directly from the Conceptual Fingerprint (CFP).

The CFP based connection of "similar" information in different and distributed databases can be achieved with minimal effort of the content providers. Simply providing access to their data will allow the system to create CFP's of all desired content fragments and when these fragments (CV's projects, publications etc.) are linked to contact data of people and organizations, the CFP's of people and organisations will automatically accumulate into activity profiles and

knowledge profiles. The matching process is now carried out completely at the (language independent) concept level and can be started from any text fragment of interest. Most systems exploit a simple full text Boolean search engine in addition to the Collexis® matching technology, in order to be able to find highly specific and infrequent words that are not covered by the thesauri used in Collexis®.

3 The International Research Information System (IRIS)

3.1 Background

The Internet, although contributing significantly to the almost unmanageable information explosion itself, also offers unprecedented opportunities to reduce the burden of science management. For the first time in the relatively short history of Information Mediation, the newest technologies provide for tools to handle information in a revolutionary different way. Matching proposals across languages with the best referees for example, and the efficiency of Information Exchange can be improved significantly through the method of matching information based on knowledge-driven indexing described in chapter 2.

Effective International Research management will contribute significantly to the increasing societal demand for accountability, transparency, effectiveness, accessibility and productivity of public and private resources spent on Scientific Research. IRIS is intended to enable all research councils to utilize the IRIS technology, which will provide for inter-council exchange of information and international selection of referees and much more. The Netherlands National Research Council (NWO), the Ministry of Education, Culture and Science (OC&W) and the National Institute for Scientific Information (NIWI) have taken the initiative to develop a prototype of a comprehensive international research management system tailored to the needs of Research Councils and other institutions dealing with research management.

3.2 Scope and aims

IRIS intends to create an International network of National Research Councils (NRC's), sharing minimal information on both national research and scientists on a global scale, while allowing each participating organization to keep using its existing Information Management system. IRIS is meant to function as a decentralized system with national responsibility for the quality of all data. The basic principle is a collective back up of local data through an ASP based approach. Each participant will manage its own domain in the system, which allows for an organization's own "look-and-feel".

The project will result in a working prototype of a Web based information system that allows for:

(a) Web-based, global and interactive selection of referees.
(b) On-line submission of proposals in a variety of formats.
(c) On-line reporting and updating.
(d) Personalized support for scientists and science managers.
(e) Real time evaluation of scientific performance.

The ultimate goal is to make the human and institutional aspects of scientific networking and management much more meaningful, efficient and thus less time consuming. In addition, a decrease in monetary overhead can be achieved. If ideally all (or the vast majority) of active scientists would be represented in the system, the search for (international) referees on any given proposal or paper would be reduced to seconds rather than hours and collective contacting of the referee panel could be performed directly from within the system.

3.3 Status of the IRIS project in May 2002

NWO is in an advanced stage of discussion with a number of research councils in Europe, The European Commission, The United States, Asia and Latin America to join the prototype phase. Several councils and major scientific publishers have expressed their keen interest in joining the project, which is crucial for its success.

NWO has found financial and moral support for its initiative to launch the IRIS project. In The Netherlands the participating organizations are convinced of the need for IRIS and have decided to fund and implement it. At the international level NWO has already found major interest in the development of an internationally applicable system. To NWO's knowledge no similar initiative has been launched with the same scope IRIS.

Since the first stage of development of the necessary applications has been completed, NWO intends to test all features implemented in the Netherlands with several international partners, including but not limited to a selection of National Research Institutes.

IRIS will become operational for all fields of expertise. The Medical and Heath Sciences field plays a major role in the pilot phase based on the exceptional quality of the thesauri available for these fields. Other fields will be subsequently included in the project as more thesauri become available.

NWO intends to honor the general understanding that validation of knowledge and management of content and personal data is foremost the responsibility of NRC's. Therefore the choice has been made to implement a fully de-centralized system with optimal opportunity for NRC's to control their own data and give them autonomous authority of quality control through National Focal Points. Based upon the technology used it will still be possible to provide global access to all data entered through the CFP's stored on an ASP server.

In order to ascertain the feasibility of the proposed approach at an international level NWO intends to have 5 major NRC's committed to participation in the pilot phase, prior to launching the formal try-out. At a formal meeting in September 2002, the project will be presented and a broad representation of NRC's will be invited to participate in this meeting. Up to 10 additional Councils can participate in IRIS during the Pilot Phase at *no cost* other than local overhead and training of personnel.

The final goal of IRIS is to develop a functional global network of NRC's and related content providers for the benefit of international scientific networking and management. As many Institutes as possible should have joined the initiative by then. After successful completion of the pilot phase, it is intended to spin off IRIS as an independent organization. The legal structure will be discussed in detail with all IRIS partners in the evaluation phase (2003).

The term for completion of the pilot phase of IRIS is two years. The project was launched in October 2001 and should be completed in October 2003.

Participation in IRIS has several immediate benefits for IRIS partners. It provides the ability to find referees online based upon available knowledge profiles of all scientists available. In addition each Partner will have the ability to search all fingerprinted scientific data and match them with queries based upon Search Fingerprints created from full text sources.

Another obvious advantage is the ability to compare grant-applications with other already funded projects on a global scale. Lastly, all scientists working with or for the Partner will have the ability to utilize the features of IRIS for their own projects.

3.4 Technical Scope of subsequently developed tools and steps in IRIS

Note: The steps and tools are described following the chronological order of the science management process: Calls for proposals-submission-review-reporting and evaluation. The actual technical development of the tools does not necessarily follow the same sequence. It has been de-

cided to take *Step 2*, the "referee finder" as a major first step as it contains the greatest challenges, both technically and in terms of networking.

Moreover, several potential partners may have excellent systems in place for on-line submission and/or reporting. In such cases the interactive connection of the existing systems to the central referee selection system would be a first step and that makes the referee module the only default centralized tool in IRIS.

Several partners may decide to implement only the referee tool (2) and this would not in any way jeopardize the successful implementation of the IRIS concept.

Step 1. On-line submission of proposals in a variety of formats

IRIS will design and implement an extremely user-friendly on-line submission tool for scientists. It can be used by organizations that do not have such a tool available. For those councils that already have on-line submission of proposals operational, only additional functionality may be imported, which will render existing submission procedures intact

Step 2. Web based, global and interactive selection of referees

Most research agencies or scientific content management organizations have a major burden during the selection of referees for submitted proposals. Not only the selection of the correct reviewer in terms of expertise and knowledge, but also the practicalities of finding these people, composing mailings and organizing the review process are a serious management issue. The IRIS Referee Finder Tool will include a feature that allows semi-automated collection of publication profiles of referees based on literature available in the public domain.

The Referee Finder will allow for an efficient decentralized tool with global access to all data available in IRIS through the shared back up of all National data. As the first step in the process, the agency approves a submitted proposal for the review process and the following steps are "interactively automated". Selected text fields in the on-line application will be fingerprinted and will generate a "proposal profile", which can be reviewed for inconsistencies by the project manager. The council manager will be able to see publications, projects of other research councils including knowledge validated anywhere in the world. The fingerprint of the proposal will be used in the matching engine to find the most suited reviewers. With this application, the organization can find the best referees for a particular project proposal in the most efficient and expedient manner possible.

Step 3. On-line reporting and updating

Research councils usually have a contractual relationship with the scientists whose research was funded, which includes regular reporting. This is at present a cumbersome process, requiring regular reminders, approval of the reports and updates of existing databases at the Councils' offices. IRIS' technology allows far-reaching automation of many elements of these processes.

Step 4. Personalized support for scientists and science managers

All the scientific output of a researcher represented in the IRIS knowledge or activity profiles (ISIC's) is constantly updated and used for search by institutes and individuals. IRIS will provide for personalized "interest rooms" for individual scientists. Since scientist can receive constant updates on information matching their interest profiles, this feature could create a critical incentive for reviewers and scientists to keep their profiles updated in the system.

Step 5. Real time evaluation of scientific performance

In close collaboration with national and international partners involved in bibliometric analysis of scientific performance it is proposed to develop dedicated tools for internal and external evaluation of scientific institutes. Based on the cross-referencing technology used, such tools can eas-

ily be conceptualized and developed. However, serious discussion between the players is still needed to define the parameters and consequently the (fine-tuning of the) technology and to work out schemes that respect full privacy security of sensitive data related to persons or institutes. Therefore, these tools will be heavily protected and only made available to officials dealing with science policy.

3.5 Organizational Scope

National Research Councils and Science Management Organizations can participate in IRIS as co-owners of the initiative. For the initial phase, a small international administrative office (through NWO) will provide the necessary administrative support including help desk and provisions for training employees of local Institutes. NWO is prepared to assume responsibility for effective execution of IRIS during the first two years of its existence. An International Advisory Board will be established after the Pilot Phase has been evaluated and a decision to continue IRIS on a permanent basis will be made by all collaborating IRIS Partners. For further details, please contact the corresponding author via: iris@nwo.nl

4 References

Heisenberg, W. (1976) The uncertainty principle. Zs. F. Phys. 43, 172-98

Sveiby, K.E. (1997) The new organizational wealth: Managing and measuring Knowledge based assets. http://hallinternet.com/net_history_trends/71.shtml

Tuomi I, 1999 Data is more than knowledge, Implications of the Reversed KnowledgeHierarchy for Knowledge management and Organizational Memory, Proceedings of the Thirty-second Annual Hawaii International Conference on System Sciences. http://www.computer.org/proceedings/hicss/0001/00011/00011071abs.htm

Van Mulligen EM, Diwersy M, Schmidt M, Buurman H, Mons B. (2000) Facilitating networks of information, Proc AMIA Symp 2000: 868-72

5 Contact Information

Barend Mons
Nederlandse Organisatie voor Wetenschappelijk Onderzoek
WOTRO
P.O. Box 93120
2509 AC Den Haag
The Netherlands

e-mail: barend.mons@inter.nl.net