A federated search engine providing access to validated and organized "web" information within the World Agricultures Observatory

Philippe LEMOISSON[1], Audrey BONA[2], Thierry HELMER[3], Michel PASSOUANT[1]

1. CIRAD, UMR TETIS, TA C-91 / F, 34398 Montpellier Cedex 5 France
2. Ecole Nationale Supérieure de Cognitique, 146, rue Léo Saignat - Case 40, 33076 Bordeaux Cedex, France
3. CIRAD, DSI, TA 182/05, Avenue Agropolis, 34398 Montpellier Cedex 5 France

Abstract

As part of the international initiative "World Agricultures Observatory (WAO)", CIRAD aims at producing and capitalizing original knowledge in order to promote information exchange and to enable reflection and debate about the sustainability of production systems. A web portal will offer three classes of services: i) access to documentary resources organized and validated by scientists, ii) sharing of data managed by the WAO partners, starting with the sites where CIRAD conducts studies on rice, cotton, livestock ..., iii) access to a collaborative platform, to news, to a literature review ... This article discusses the 'web documentary resources' component.

We first introduce the concepts and technical features of the supporting tool 'Système d'Information Scientifique et Technique (SIST)' developed by CIRAD in the context of a project funded by the French Ministry of Foreign Affairs in West Africa. In a second part, we expose the cognitive approach adopted: analysis of the needs, design, implementation and evaluation of the solution and enter in the details of the analysis and design steps. In a third part, we describe the SIST-OAM prototype. We then evaluate the solution. Finally we put in perspective the overall approach and compare it with usual web search of information.

Résumé

Dans le cadre de l'initiative internationale « Observatoire des Agricultures du Monde (OAM) », le Cirad s'est donné comme objectif la constitution d'un espace de production et d'accumulation de connaissances originales, en vue d'échanges, de réflexions et de débats sur la viabilité des systèmes productifs. Ces connaissances seront accessibles à travers un portail web proposant trois classes de services: i) accès à des ressources documentaires validées et organisées par des scientifiques, ii) partage des données gérées par les partenaires OAM, à commencer par les terrains sur lesquels le Cirad conduit des études sur la riziculture, la production du coton, l'élevage …, iii) accès à une plateforme collaborative, des actualités, une revue bibliographique …

Cet article traite de la composante 'ressources documentaires web'.

Nous présentons dans un premier temps les concepts et caractéristiques du Système d'Information Scientifique et Technique (SIST) développé par le Cirad dans le cadre du projet de même nom réalisé en Afrique sur financement du Ministère français des affaires étrangères. Dans un second temps nous exposons l'approche cogniticienne adoptée pour le mettre en œuvre : analyse des besoins, conception, implémentation et évaluation de la solution, en entrant dans le détail de l'analyse et de la conception. Nous décrivons dans un troisième temps le prototype SIST-OAM réalisé. Puis nous l'évaluons. Finalement, nous mettons en perspective l'ensemble de l'approche en la comparant avec les méthodes classiques de recherche sur le Web.

Introduction

As part of the international initiative "World Agricultures Observatory, *in French: Observatoire des Agricultures du Monde (OAM)"*, CIRAD aims at producing and capitalizing original knowledge, in order to promote information exchange and to enable reflection and debate about the sustainability of production systems. A specific three years thematic action (2009-2011) has been launched with the two objectives of providing research work on the question of the viability of agricultures through the study of five contrasted cases, and of prototyping a set of methods and technical features aimed at supporting the international initiative.

The technical solution which is currently under construction consists in a web portal offering three classes of services:

i) access to documentary resources organized and validated by scientists ;

ii) sharing of data managed by the OAM partners, starting with five contrasted study sites chosen by the CIRAD team among the regions where research is ongoing : cotton in the south of Mali, transhumant livestock in Niger, food crops in Costa Rica, rainfed rice in the highlands of Madagascar, flooded rice in the Mekong Delta ;

iii) access to a collaborative platform, to news, to a literature review ...

This article discusses the 'web documentary resources' component; it is organized in five parts:

1. we introduce the concepts and technical features of the 'Système d'Information Scientifique et Technique (SIST) developed by CIRAD which plays the role of supporting tool for the 'web documentary resources' component.
2. we expose the cognitive approach adopted, which starts by an analysis of the needs and current practices, translates this analysis into an adequate setting of the SIST platform, and assesses the prototype in order to prepare future improvements.
3. we describe the SIST-OAM prototype.
4. we provide an assessment of the result
5. we put in perspective the overall approach and compare it with usual web search of information.

Presentation of the SIST platform

The SIST devices are used mainly to meet the needs of "information portals" or "thematic observatories" in order to "support effective information searching and analysis as well as to enhance communication and collaboration among researchers" (Chau, Huang et al. 2006). A portal can be built and made operational very easily and quickly through customization of the SIST modules by mean of the SIST toolbox.

We first present below the components of the toolbox, and then briefly browse through the few steps required by the customization.

The SIST platform: a federation of components addressing different types of information resources

The SIST platform provides a federated search engine that allows searching most of types of resources which are available on the Internet. The SIST platform does not limit its focus to websites built upon *static pages*; it also includes the exploration of the deep Web.

Thus the SIST platform has the ability to simultaneously search *online databases* (or any kind of information accessible via a *web form*), *websites*, *open archive repositories*, *RSS feeds*, *full text repositories* (physical directories containing the documents to the most common formats hosted by servers) as well as the *forums* and *Wiki* locally run on the server.

The whole system architecture is based on a standard Open Source (Linux, Apache, MySql and PHP) solution developed on top of SPIP-Agora.

Each category of searchable sources is addressed through a specific module. The activation and coordination of these different modules (*MAGELLAN*, *CYRUS*, *GUTEMBERG*) as well as the homogenization of results are performed by the federated search module *HUBBLE*.

When a final user requests for example "*rainfed rice*" to the SIST platform, each type of source is addressed in a specific way, although he/she cannot notice it. Two major searching strategies are available: i) crawling of "a periodic and incremental centralization of full-content indices of widely dispersed and autonomously managed document sources" (Simeoni, Yakici et al. 2008), ii) direct request upon a set of-preselected resources. One or the other strategy will come into play according to the type of resource:

- when addressing *static pages in websites* or *full text repositories*, the *MAGELLAN* module operates in two times, like most of search engines. It firstly scans (by spidering and crawling) at regular intervals all sources selected by the SIST administrators and builds its own indexes; doing so, its stores expressions like "*rainfed rice*" as well as links between these expressions and the original sources. Then, on a user's request, it retrieves "on the fly" through a full text interface the data supplied by the indexing engine and returns links to original sources;
- when addressing *open archive repositories*, the *CYRUS* module stores an image (metadata) of the original sources, in the format 'Open Archive Initiative's Protocol for Metadata Harvesting (OpenArchives.org 2008), and then uses it as an index. As soon as the SIST administrators have selected an Open Archive, the OAI module daily updates its storage of the metadata; those can be either queried in full text or retrieved in a structured way by combination of criteria on different fields.
- when addressing *RSS feeds*, the *GUTEMBERG* module browses through the sources, and returns only the RSS feeds articles that match the search criteria "*rainfed rice*". The research is only in full text. The contents of RSS feeds are not stored in advance by *GUTEMBERG*, but retrieved on the fly.
- when addressing *Wikis or fora*, the SIST platform browses through the pages and selects those providing an answer to the request;
- when addressing *web forms*, the *HUBBLE* module takes into account "on the fly" the html pages sent by the form; it is based upon mnoGoSearch[tm] web search engine software (http://www.mnogosearch.org/). A preliminary analysis of the web form reproduces the URL query embedding the user's search criteria. A mask is then applied on the results in order to extract the required metadata.

The *HUBBLE* module is also responsible for the federation: it distributes the search criteria on all other modules, controls their operations and aggregating and homogenizing the results returned by these modules. Finally, the *HUBBLE* module provides a customizable framework for the user interface allowing categorization of the different sources as well as useful features operating on the results (sort by relevance, view, export, create alerts).
On the user's request "*rainfed rice*" to the SIST, the results of all modules are displayed by *HUBBLE* in order of arrival; among those is the hypertext link allowing direct access to the original data.

It must be noted that in order to provide a consistent presentation of results, answers provided by the different modules have to be converted on the fly to a single format; the RSS format was chosen for homogenization of SIST results. It indeed offers the five essential metadata that describe a result regardless of the type of source which originated it; a section of a web page, a bibliographical reference, an OAI record, RSS feed, a directory sheet, a document, an intervention in a forum, can all be described in a notice containing i) a title, ii) an author, iii) a date, iv) a description and v) a link.

The SIST platform: technical steps required by the customization

As reported in (Chau, Huang et al. 2006), the approach for building a web portal goes through two steps: "(1) creating a domain-specific collection of Web content and (2) supporting searching of the documents and analysis of search results."

Thus, before constructing an "information portal" or a "thematic observatory" with the SIST toolbox, one must firstly operate a significant inventory of electronic resources available on the Web or on the network, in relation with the domain to be covered.
One must also define a classification system for these resources. It might be a thematic classification, a geographical classification, or otherwise.
To end with, it is necessary to define the granularity with which these resources will be scanned by the research tools of SITS.
To draw an example, one website consisting in several different sources can be viewed as a collection of web pages that are all indexed, or as a single resource. Finally, if the website contains a *web form* or if it produces a *RSS feed*, it can be considered as many different sources, each of which being described in the appropriate module of the SIST platform.
To complete the design and implementation of the online website, the SIST platform device also provides a website manager tool (CMS) that will orchestrate all the online modules.

The SIST toolbox provides two interfaces:

- a customizable “user interface” giving access to the selected sources through the federated search engine ;
- an “administrator interface” allowing to configure the prototype (addition / removal of sources ...), i.e. to customize the “user interface”, and to manage the informational content.

The cognitive approach

The data managed by the OAM partners is in first place associated to five contrasted study sites: cotton in the south of Mali, transhumant livestock in Nigeria, food crops in Costa Rica, rainfed rice in the highlands of Madagascar, flooded rice in the Mekong Delta.
The general idea was that the research scientists from CIRAD were the proper people for the identification and validation of useful information related to those study sites; it was therefore decided that the prototype would be firstly built in response to their needs, then eventually adapted or completed when more people would use it for their own search.
In this approach, the already existing SIST toolbox allows iterative and interactive setting of a tool adapted to the needs of the research scientists.

A first phase consisted in the analysis of the current methods and needs of the research scientists from CIRAD in order to identify “search scenarios” in terms of types of information sources and search criteria:

- analyzing search methods
- highlighting the common habits, criteria and procedures used by all the researchers

Three recurrent characteristics helped the building of the scenarios:

- the main search engine used is Google
- the documents sought are mainly scientific publications, articles, dissertations and bibliographic references, with a special interest for a set of Journals available through a secured access
- the search criteria are: the type of the document, the date of the document, the geographic area covered, the type of agriculture documented

A second phase consisted in gathering sources that had to be integrated into the SIST-OAM prototype. Thus, each researcher provided a list of different types of sources (websites, databases, RSS feeds ...) and posted it on the same collaborative workplace which would later host the SIST-OAM prototype.

The second phase (development of the prototype SIST-OAM) was then started, based on the scenarios and sources. The design took place in two distinct steps:

1. integration of the gathered sources, i.e. providing a proper description of each source, including the addition of “manually chosen” selection criteria aimed at completing the automatic indexation
2. system configuration, especially the choice of the tags grouping sources in an intuitive way; it was decided to have as many tags as “study sites”, with the help of the “theme selection criteria” automatically understood by the SIST platform. Thus, each type of agriculture is represented by a tab that lists all sources on it, and within each tab sources are organized by type. The label system allows refinement of the search by the use of the selection criteria “type of document” and “country”.

The last phase consisted in an evaluation of the first version of the SIST-OAM prototype.

The SIST-OAM prototype

This SIST-OAM user interface consists of four main tabs (not those which are visible in Fig.1.):

- the tab accessing the *federate search engine* (currently activated in Fig. 1.) ,
- the tab displaying the latest scientific news on specific topics,
- the tab displaying alerts on specific topics, according to user’s specifications
- the tab accessing the collaborative work platform including fora and wiki.

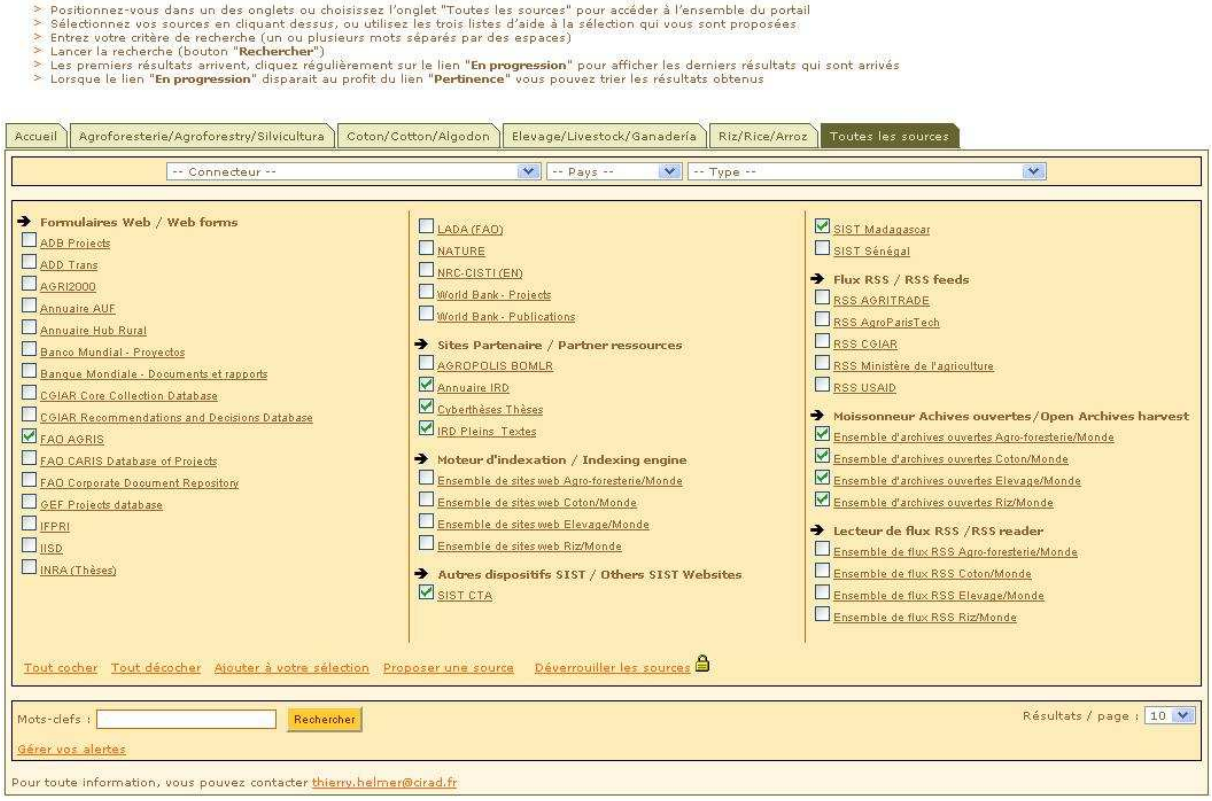

Fig. 1. the SIST-OAM prototype interface

Queries are performed using key words and their specification can be completed with the help of different labels. Results can be sorted by relevance. The *federate search engine* is organized in the following way (see Fig. 2.):

- the sources
- the labels allowing a pre-selection upon sources
- the thematic tabs
- actions on sources

- the search field where the user will indicate his request with key words (for example: “rainfed rice”)

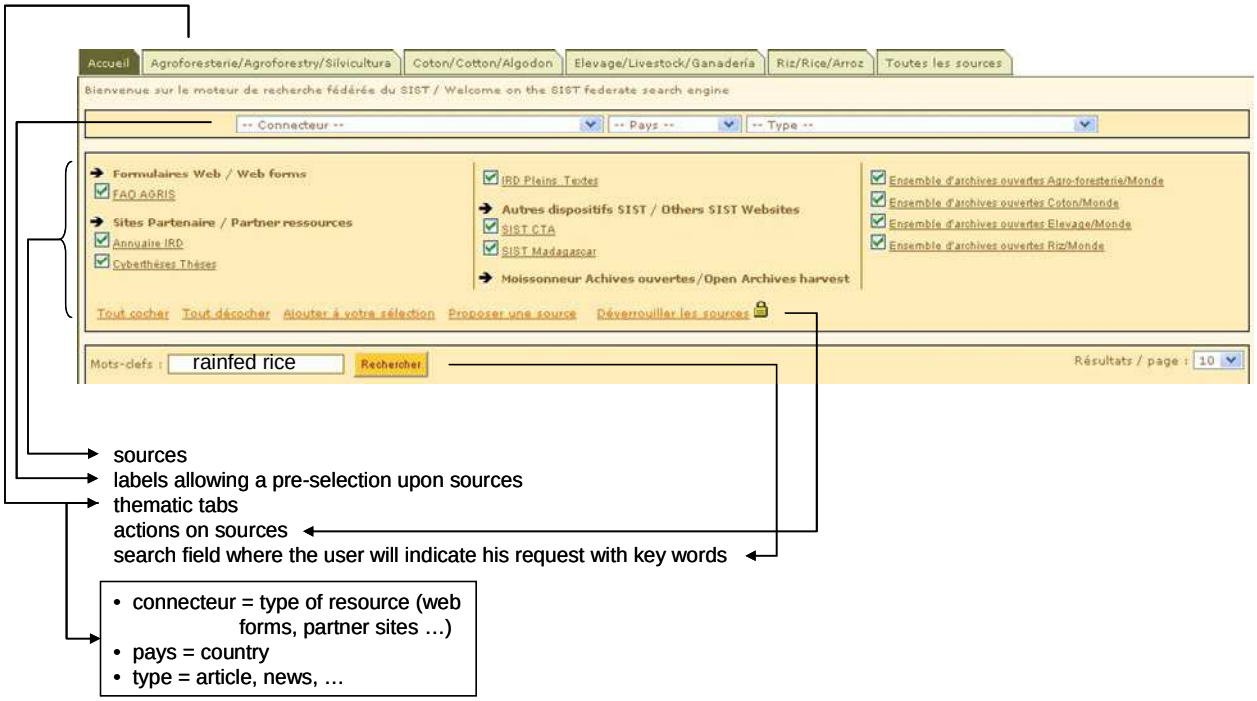

Fig. 2. the SIST-OAM federate search engine

The management and organisation of the SIST-OAM informational content is done through the “administrator interface”, which provides access to the different SIST modules, according to the different types of sources:

- the *MAGELLAN* module allows management of *websites* or *full text repositories*
- the *CYRUS* module allows management of *open archive repositories*
- the *GUTEMBERG* module allows management of *RSS Feeds*
- …

The addition of a *web site* through *MAGELLAN* is illustrated in Fig. 3. below :

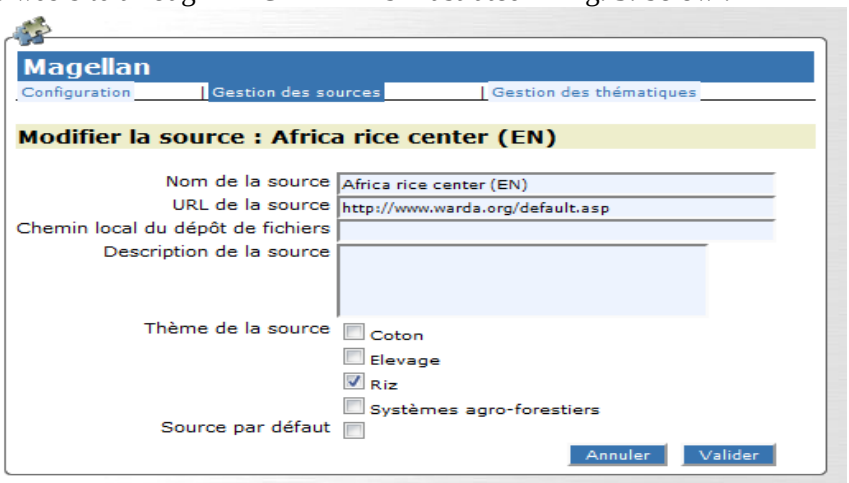

Fig. 3. the *MAGELLAN* module

In the *MAGELLAN* module, the source is described by its various characteristics including name, URL and linked to a thematic tag. Once the source referenced in *MAGELLAN*, it is visible from the search engine "websites" only (module Magellan)

The federated search engine *HUBBLE* allows the simultaneous interrogation of all types of sources.
If one wants a set of websites (sharing a same thematic) to be searchable from Hubble, it is necessary to create an aggregating source in Hubble with the following characteristics:
- Category: '*MAGELLAN* indexing engine'
- Theme (combo): Thematic shared by all the websites of the set.

For example, the source "All websites Rice / World":

- will be visible from *HUBBLE* in the theme "Rice / Rice / Arroz"
- belongs to the category '*MAGELLAN* indexing engine', thereby establishing the link with sources of *MAGELLAN*,
- gathers sources described in *MAGELLAN* as containing information about "Rice"

The indirect communication between the two modules is done by the thematic field, as illustrated in Fig.4. below:

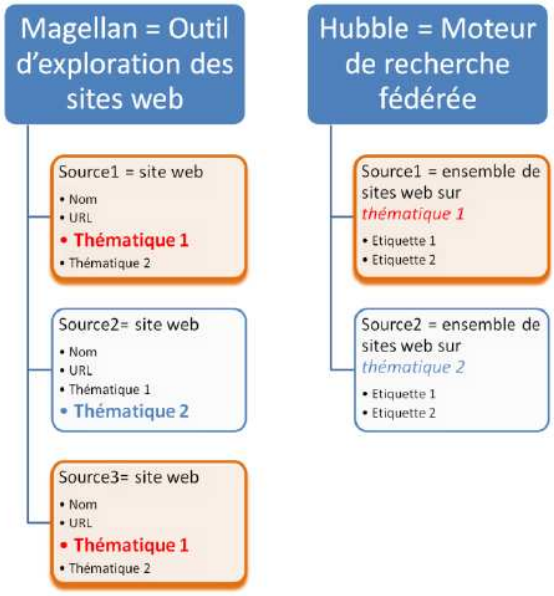

Fig. 4. connecting *MAGELLAN* to *HUBBLE*

Assessment of the SIST-OAM prototype

Ex-ante evaluation

The engineer in charge of the prototype was not familiar with the SIST toolbox. During the building phase, she had established a personal diagnostic highlighting some necessary improvements. According to this diagnostic, the system could be used "in the current state" by CIRAD researchers; it would require some training at the beginning but once taken in hand it would facilitate information retrieval. However, if it had to address a wider audience, part of its general ergonomics should be rethought.

Ex-post evaluation

A series of tests, implying the researchers who had provided the specifications were planned in order to confirm on reverse this diagnostic, and highlight the positive and negative aspects.
The tests were based on a series of questions associated to several scenarii.

- a positive aspect for the users was the easy access to a wide set of all kind of resources which had been pre-selected by trustworthy colleagues.

Two kinds of requested adaptations came out:

- most of them could be answered by ergonomic changes, like the size of the character set, the number of sources …
- a few were related to the internal settings of the prototype and need a deeper analysis, like the indirect communication between *MAGELLAN* and *HUBBLE* which does not correspond to the categorization habits of the experts.

If user interface is very important for users, it is often designed and implemented without method, which makes modifications difficult. The use of design model that allows for flexible development by formalizing the user interface prototype in any GUI environment should by tried.(Lee, Chang et al. 2007).

Recommendations

Since a brief analysis of the system implying only a small sample of users is not sufficient to lead an overhaul of the tool, the next steps before the integration of SIST –OAM in a future portal would be to go through a full ergonomic analysis. Such an analysis should imply a larger sample of users and should be based on specific evaluation grids.

Discussion

In this article we have presented an approach as well as a supporting toolbox for the management of 'web documentary resources' associated to a number of pre-defined "themes":

- pre-selection of web resources by specialists of the pre-defined "themes";
- customization of the modules of the toolbox.

Because of this pre-selection of web resources, this approach is significantly different from the mere use of a popular engine.

We know for example that the demonstrated trust in Google has implications for the search engine's tremendous potential influence on culture, society, and user traffic on the Web.(Pan, Hembrooke et al. 2007). As a counterpart of this power, Internet users are pelted with spam, hoaxes, urban legends, and scams - in other words, untrustworthy data.

The Internet is largely without any infrastructure to help users identify authoritative and trustworthy content. Indeed, the history of the Internet is littered with examples of how technologists have underestimated the crucial role that social trust and authority play in communication (McDonough 2004).

In the case of the SIST-OAM, the trust is not put in the engine but in the people who pre-select the sources.

It seems that this easy access to trustworthy web resources is much appreciated by the users. For cons, the ergonomics of the toolbox could be improved. The challenge seems therefore to keep the richness of the tool while providing a more intuitive interface.

Moreover, it might be interesting to give more power to the indexation process for knowledge extracted from texts, by building a terminological database around an Ontology for the concepts (Feliu, Giraldo et al. 2004).

References

Chau, M., Z. Huang, et al. (2006). "Building a scientific knowledge web portal: The NanoPort experience." Decision Support Systems 42(2): 1216-1238.

Feliu, J., J. J. Giraldo, et al. (2004). "The GENOMA-KB project: a concept based term enlargement system ". Retrieved 15/01/2010, 2010,
from http://www.upf.edu/pdi/dtf/teresa.cabre/docums/ca04fel.pdf.

Lee, C. M., O. B. Chang, et al. (2007). "Usage-centered interface design for quality improvement." Computational Science - ICCS 2007, Pt 2, Proceedings 4488: 1139-1146
1251.

McDonough, M. (2004). "In Google we trust?" Aba Journal 90: 30-+.

OpenArchives.org (2008). "The Open Archives Initiative Protocol for Metadata Harvesting." Document Version 2008-12-07T20:42:00Z. Retrieved 02/02/2010, 2010, from http://www.openarchives.org/OAI/openarchivesprotocol.html.

Pan, B., H. Hembrooke, et al. (2007). "In Google we trust: Users' decisions on rank, position, and relevance." Journal of Computer-Mediated Communication 12(3): -.

Simeoni, F., M. Yakici, et al. (2008). "Metadata harvesting for content-based distributed information retrieval." Journal of the American Society for Information Science and Technology 59(1): 12-24.