Role of thesauri in a scientific organization

Carlo Ferigato Joint Research Centre of the European Commission,
Institute for the Protection and Security of the Citizen
via E. Fermi, 2749 I-21027 Ispra (VA) Italy
carlo.ferigato@jrc.it

Giuseppe Merlo Joint Research Centre of the European Commission,
Ispra Site Directorate
via E. Fermi, 2749 I-21027 Ispra (VA) Italy
giuseppe.merlo@ec.europa.eu

Daniela Panfili via Rovereto, 11 I-20127 Milano (MI) Italy
daniela@piyosailing.com

Dario Rodighiero viale Cornaggia, 6 I-23807 Merate (LC) Italy
fumoseaffabulazioni@gmail.com

Abstract

We aim at describing the mediation language between users and indexers in a document retrieval system for a big scientific community intimately related to the European Union policies. We argue that this mediation can be represented at three levels by thesauri. At the first level, thesauri are sets of indexes apparently coordinating the possible searches by means of term to term relations like Narrower Term and Related Term. At a higher level, person to terms relations are consequent to the use of thesauri for indexing and retrieval, while person to person relations are embodied into a thesaurus via the implicit representation of the organization it serves.

In this way, thesauri constitute a network of mediation having historical, social and - because of the scientific community served - scientific and technological perspectives. These three perspectives are embedded in time, since changes in organization change the person to person relations, change in retrieval and indexing needs change the person to term relations. Changes in document types, in aims of the organization and progress in science change term to term relations. In particular, we want to analyse the network originally proposed by the Euratom Thesaurus (1966, 1967) and the network of relations - in the three perspectives above - assumed while it was designed.

Subsequently, we compare the results of this analysis with a more recent thesaurus designed for a community very close to the one originating the Euratom thesaurus. This latter thesaurus is part of an information retrieval system for the mentioned scientific community. This information retrieval system provides to its users a browsing mechanism through the relations above. Its interface is based on the Focus+Context and Elastic Grid concepts, allowing for a flexible graphical structure.

Introduction

The association of indexes to sets of documents for ordering them as well as the organization of indexes in *maps* as mnemonic tools is part of the history of mankind (Mangani, 2006; Serrai, 1997).*Indexes* are terms of a formal language allowing to *indicate* sets of documents. Here we use *indicate* with both the meaning of "giving name" or "attribute a quality". If the set of index terms is controlled in some way, it is said to be a

controlled vocabulary. Controlled vocabularies are a specific subject of investigation in Library Science. The importance of controlled vocabularies in Library Science can be expressed by quoting the introduction of a fundamental text in this field:

> This book deals with properties of vocabularies for indexing and searching document collections; the construction, organization, display, and maintenance of these vocabularies; and a vocabulary as a factor affecting the performance of a retrieval system. (Lancaster, 1972, p.VII)

A controlled vocabulary ordered by a set of relations between its index terms is called *thesaurus*. Graphically, a thesaurus can be displayed as a net whose vertices are indexes and meshes are formed by drawing as arcs the binary relations between indexes.

In this work, we deal with some specific thesauri designed for the *Joint Research Centre* of the European Commission (the JRC) and we will discuss their use from three perspectives in sections Networks of people, Networks of terms and people, and Networks of terms below.

By *network of terms* we mean the thesaurus with his history and the choices made while designing it. We claim that a thesaurus is more than a formal language used in Library Science as tool for the "manipulation of classes" of documents. It captures some characters of the community it serves.

As a network of terms, a thesaurus can be used in many ways: for keeping a collection of documents ordered; for measuring the consistence and the quality of a collection of documents; as a formal *mediation language* for information retrieval purposes.

Upon a thesaurus two more networks can be built. In section Networks of terms and people, the relations between people and descriptors are analysed. These relations link people to subjects of interest as in a *recommender system*. In section Networks of people we describe the organization for which the thesauri were designed. In this section, our aim is in describing the mutual relations between the structure of a thesaurus and the structure of the organization. In section Networks of terms, a thesaurus for nuclear science developed by a group of indexers and researchers in the same community (Euratom Thesaurus, 1966, 1967) is described in detail and the project for new thesaurus for the JRC is presented.

The practical outcome of our work is in the design of a new information retrieval system for the JRC based on the new thesaurus presented in section Networks of terms. This system, called SIRS (Scientific Information Retrieval System), is built by considering both the history of the JRC and the analysis of a thesaurus as a formal language as in section Networks of terms and people. While performing this task, we used as conceptual tools some of the *Communication Disciplines* introduced by Carl Adam Petri (1977, 2001) as analysis tools for computerized systems seen as "general medium for strictly organized information flow". Our work in designing SIRS is reported in section SIRS thesauri. Conclusions and ideas for continuing our work are reported in the last section.

Networks of people

By *network of people* we consider the professional relations among the members of a given scientific organization through its history. In particular, the Joint Research Centre of the European Commission - with attention to its Italian establishment - is the object of our work. The JRC was formally established by article 8 of the Euratom Treaty (1957) . We remark that, with the same article establishing the JRC, the mandate for establishing a *uniform terminology* concerning the object of work of the JRC was given.

> After consulting the Scientific and Technical Committee, the Commission shall establish a Joint Nuclear Research Centre. This Centre shall ensure that the research programmes and other tasks assigned to it by the Commission are carried out. It shall also ensure that a uniform nuclear terminology and a standard system of measurements are established. [...] (Euratom Treaty, 1957, art.8)

After a period of stability in the research themes from 1962 to 1971, a progressive enlargement of the fields of research occurred together with a strong dependence of the research themes from policy acts, multi-annual work-programmes approved by the Council.

Concerning the scientific activities of the Italian establishment of the JRC, its presentactivities are illustrated by the organizational chart in Figure 1. A comparison with the scientific activities as reported by Gueben (1962) shows deep enlargement of the interests.

> Physique des Réacteurs (Physique Mathématique Appliquée, Physique Neutronique, Essai Critiques, Automatisme et Régulation); Matériaux (Métallurgie; Chimie; Physico-Chimie); Engineering (Technologie; Echange Thermique); Bibliothèque et Documentation; Neutronique Expérimentale et Conversion Directe; Physique Sanitaire, Médecine et Santé. (Gueben, 1962, pp.3-4)

While the JRC research themes in 1962 and the Euratom thesaurus could be superposed completely, this operation can not be done for the present research fields. Consequently, while in the period 1962-71 the Euratom thesaurus was a complete representation of the JRC scientific interests, this is not true today.

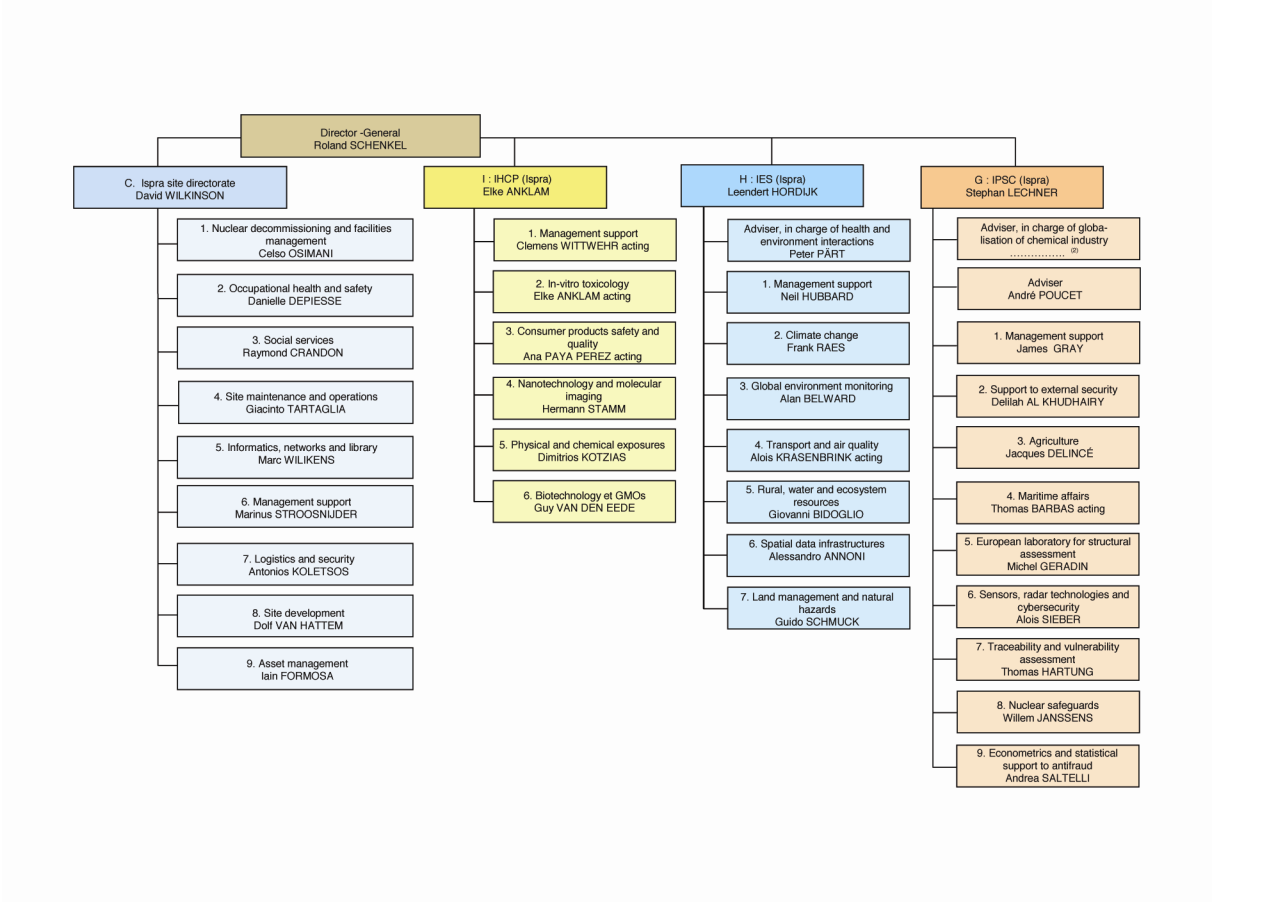

Figure 1. Organizational chart (2008) of the JRC excerpt for main activities at the Ispra site.

A proposal for a combination of thesauri allowing for an explicit representation of both the organization and the research activities are proposed in section SIRS thesauri below.

Networks of terms and people

We assume that coordination among people participating in an organization is performed via language. The role of the language in organizations and its use as a modelling tool is a deep subject of study. A survey of approaches based on linguistics, philology and pragmatics is reported in Bazermann & Paradis (Textual Dynamics of the Professions, 1991). From the perspective of computer science, models of conversations have been used as representatives of organizations since the design of the program *Coordinator* (Winograd & Flores, 1986), and originated a rich debate in the world of *Computer Supported Cooperative Work* - CSCW. This debate is well summarized by Giorgio De Michelis (2007). In these approaches, organizations are represented through the *conversations* of their members. This type of representation does not touch the individual terms used in conversations but some *high-level* representation of the conversations themselves. Applications using some sort of controlled vocabulary yet exist. These applications are called *recommender systems* (see the content-based methods in Adomavicious & Tuzhilin, 2005) where people are represented by index terms.

In this work, we assume the existence of a formal language as a *controlled vocabulary* endowed by binary relations. The resulting relational structure - the thesaurus - is our modelling tool. Our use of a thesaurus as a *linguistic* representation tool is a compromise between the explicit representation of conversations and the simple use of index terms for representing people.

A thesaurus can be seen as representative of an organization in the following three ways:

- As in Blair (1990) information retrieval is a *process of communication*:

  If we understand that the description of documents for retrieval, and the formulation of search requests are fundamentally linguistic acts, then it is no great conceptual leap to see that information retrieval is fundamentally a process of communication. Inquirers are trying to describe the information they need in a way that indexers would understand, and indexers (or automatic indexing procedures) are trying to describe the content and context of documents in the collection in ways that would be understandable to the inquirers. (Blair, 1990, p.188)

  Obviously, it is not a process of communication as in the explicit representation of *speech-acts* in the Coordinator application mentioned above since here the operations of indexing and inquiring are not synchronous.

- Rolling (1966), while remaining to a practical level of description, sees the thesaurus as a kind of *machine language* apt to mechanical processing.

  The intricacies of natural language, as it occurs in the texts of documents and abstracts, make it unsuitable for machine analysis. Computer retrieval requires a controlled vocabulary in which ideally each descriptor stands for one concept only. (Rolling, 1966, p.96)

  By extending this view to a wider game of language played between indexers and inquirers we simplify the interactions among people to the ones it is possible to play on a thesaurus seen as a kind of *terminological chessboard.*

- At least in a simple way, a thesaurus contains also a *structural representation of the organization* it serves.

  Consequently, the games of language played on it have multiple levels. The first level, a formal mediation language between library services and people using them, contains more than a list of terms and their mutual abstract relations. If well designed, a thesaurus pragmatically allows for games of language intimately related to the structure and the interests of the organization it serves. As a simple example, for looking at the interests of a group inside a big organization, it is normal practice to look at how are indexed the documents produced by that group. In our view, language games are played at three levels: inter-organization, organization-policy and organization-research.

The organization-policy level results clearly from the JRC history after 1971, with the progressive enlargement of research themes and their implementation via research *Actions*.

In section SIRS thesauri, below, we will use the previous considerations as grounds for a new controlled language and thesaurus for the JRC.

Networks of terms

The Euratom Thesaurus (1966, 1967) was designed by the *Center for Information and Documentation* (CID) following the mandate of the Euratom Treaty. The CID was created in 1961 by the European Atomic Energy Community (Rolling, 1966) and followed the merging of the European Coal and Steel Community, of the European Atomic Energy Community and of the European Economic Community into a single entity in 1967 (Brée, 1972). The Euratom thesaurus, whose first edition is published in 1964, was very advanced. Its *terminology charts* are still quoted in the library science literature (Atchison, Gilchrist & Bawden, 2000) as one of the first examples of *graphical display.*

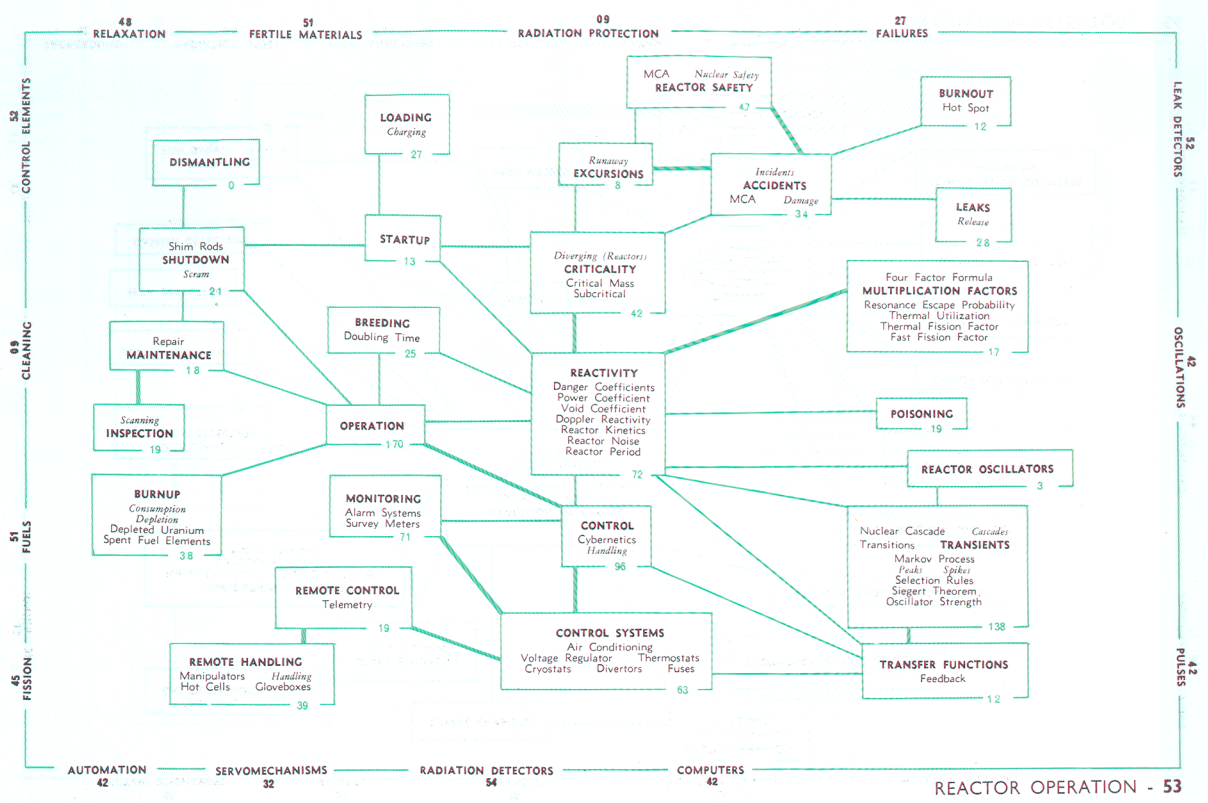

Figure 2. Terminological chart n. 53, "reactor operation" (Euratom Thesaurus, 1966, 1967).

The Euratom thesaurus served as a tool for subject control at the CID in order to supply scientists of the European Atomic Energy Community and industry in its member countries with documentary information on all aspects of nuclear energy. The terms were not limited to nuclear physics and reactor technology, but covered many related topics as radiation protection, isotope technology, fabrication and use of nuclear materials and instruments, radiochemistry, and radiobiology.

The graphic display of the Euratom thesaurus is realized by the terminology charts as in Figure 2. In these charts, terms are synthetically represented in clusters of *descriptors, non-descriptors* and *forbidden terms*. The strength of a relation between two terms is represented by the thickness of the arc. While the Euratom thesaurus is a tool well designed from many perspectives, it does not fit anymore to the retrieval needs of the JRC. A new thesaurus - indeed the combination of three thesauri - has consequently been considered while designing the Scientific Information Retrieval System of the JRC.

SIRS thesauri

The history of JRC, the evolution of its research themes and the progressive need of a close connection between the policy of the European Union and research activities requires a new controlled language for its information retrieval system. In designing it, we followed the interpretation given by Blair (1990) and assumed that a retrieval system models a *process of communication*. The *controlled language* used in this system is consequently a kind of *terminological chessboard* where *language games* - informing the *process of communication* - are played.

These language games have the general character of inquirer-indexer games played at three levels. The first level concerns the collection of documents, the traditional *extent* of a controlled language. The second level concerns the internal organization of the JRC and acts as a kind of *recommender system* since it groups around the same term different activities. The third level is intimately related to the *existence* of the JRC and associates its research themes to the general policy indications of the European Union.

In the following three sections we analyse in more detail these language games and propose three distinct thesauri for them. These three thesauri represent three distinct perspectives on the JRC, namely the perspectives of the *internal organization*, of the *policy* and of the *general research index*. Subsequently, we brief describe how these three thesauri are integrated.

Internal organization

The Actions thesaurus is a simple scheme derived from the scientific part of the organizational chart of the JRC. All of its terms have a direct connection with the organizational chart. The finest representation of the scientific activities is given by the so called *Actions*, small groups of 5 to 20 researchers. Actions are not represented in the organizational chart but have hierarchical dependence from its leaves. Few of the 120 JRC Actions active in 2008 are reported as an example in Table 1 where IPSC is the Istitute for the Protection and the Security of the Citizen and IES the Istitution for Environment and Sustainability (Fig. 1).

Actions	Unit	Institute
Maritime Surveillance	Maritime Affairs	IPSC
Community Image Data Portal	Agriculture	IPSC
Geo-Information Management and Control Methods	Agriculture	IPSC
Systematic Observations of Land and Ocean	Global Environment Monitoring	IES
Sustainable Transport	Transport and Air quality	IES

Table 1. JRC Actions active in 2008 and their hierarchical dependence from the organizational chart.

Policy

In 1981 the European Parliament and the Publication Office decided to establish a thesaurus for the organization of Parliamentary acts. A first edition of the thesaurus - now called Eurovoc - was published in

1984 in seven Community languages. The Eurovoc Thesaurus (2007) presently contains 6645 index terms, it is constantly updated and translated in 21 EU languages.

General research index

The Dewey Decimal Classification, hereinafter DDC, was conceived by Melvil Dewey in 1873 and published for the first time in 1876 (Dewey, 1996). DDC system is one of the most widely used library classification systems. This thesaurus is strictly based on a hierarchical structure supported by decimal numbers labelling.

Integration of thesauri

The overall thesaurus of the Scientific Information Retrieval System, SIRS, is the combination of the three thesauri above: Actions, Eurovoc and DDC. To create a unique relational system, the three thesauri must be joined. This operation is possible trough a mapping procedure creating two sets of cross-thesauri relations. One connects Actions terms to Eurovoc terms, other relates Eurovoc terms to DDC terms.

The mapping procedure is a task for the responsible of SIRS contents. When he works on cross-thesauri relations, he has two main duties: the first is to describe what an Action is by using Eurovoc terms. This operation is traditional indexing applied to Actions instead of books. The second is to connect each DDC term - in the subset already used for classifying the JRC Central Library holdings - to one or more Eurovoc terms.

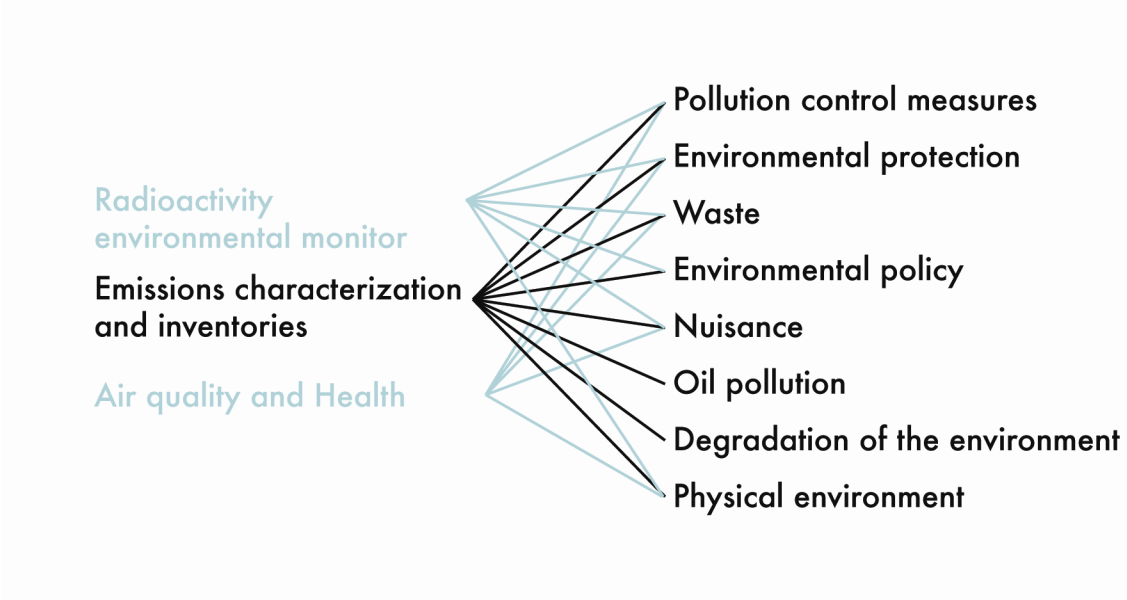

Figure 3. Action "Emission characterization and inventories" indexed by Eurovoc terms.

As an example, the cross-thesauri mapping of the Action: "Emissions characterization and inventories" with Eurovoc is shown in Figure 3. While the focus of SIRS is on this Action, it is possible to display as context other Actions. The context is computed on the number of common index terms in Eurovoc. In the example in Figure 3, Action "Radioactivity environmental monitoring" shares with "Emission characterization and inventories" six Eurovoc indexes. The two Actions are consequently mutually *close*. An example of the final display - including the DDC cross-thesauri relations - is in Figure 4.

The SIRS interface is dynamic and designed in agreement with the Focus+Context paradigm (Spence, 2007). When users' focus changes the displayed frames do not disappear from the interface, they are resized and lose definition.

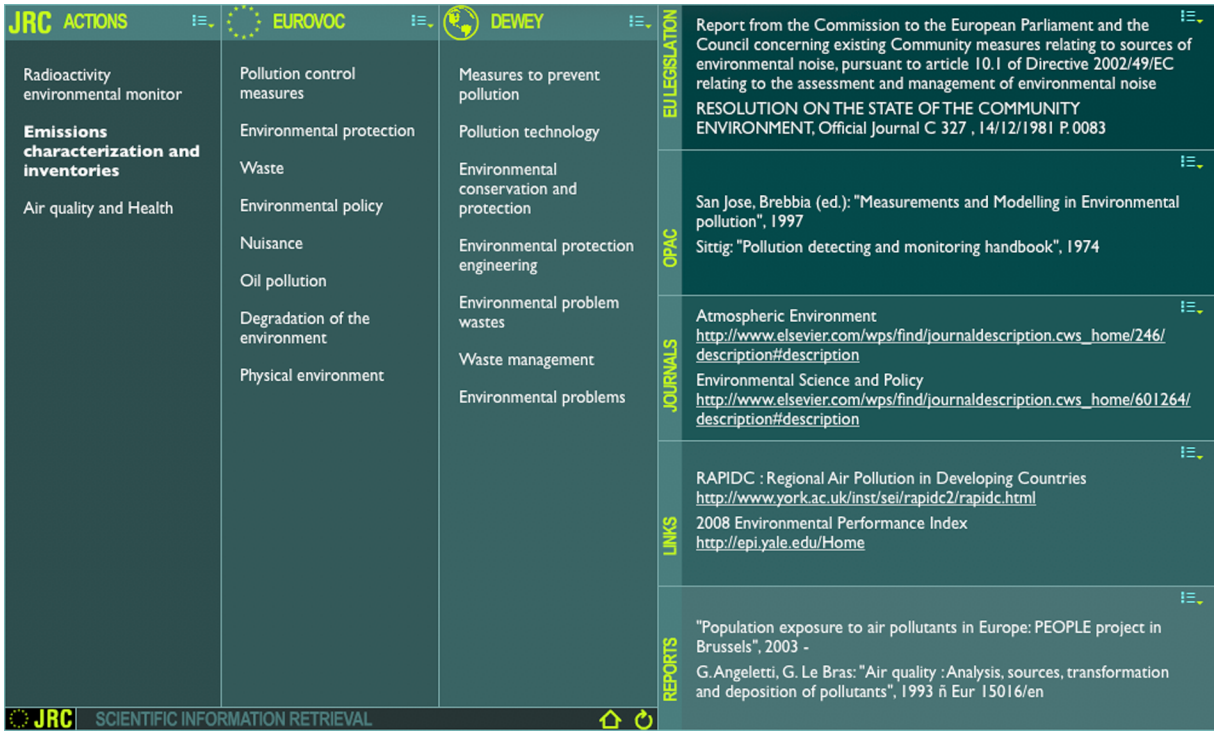

Figure 4. The SIRS interface. The three thesauri on the left. Retrieved items in the right frames.

Conclusion

We described our view on the role that a formal language for information retrieval and classification has in modelling an organization. We described this role in three layers:

1. Layer of *people*. How the history of an organization gives birth to a network of terms and changes it;
2. Layer of *terms and people*. From the plain proximity of people's interests reached through the common indexes to the more complex role of the network of terms as mediation language between inquirers and indexers;
3. Layer of *terms*. Index terms, their mutual relations and how a thesaurus can be seen as a formal linguistic tool.

Subsequently we used our analysis for the design of an information retrieval system. In designing it, we aimed at making explicit relations like *research themes* to *policy* or the mutual closeness of *research themes*.

While doing this work, we used conceptual tools coming from various disciplines: library science, pragmatics, information retrieval and computer science.

More work remains to be done since we have only shallowly touched the problem of a thesaurus as a mediation language. More precisely, its use as mediation language between both humans and non-humans and among non humans is still to be done. Moreover, a precise analysis of the *translation* (Latour, 1996) of the classification and retrieval disciplines to the computerized world is still to be done as well. We aim at continuing our work in this direction: by analysing the historical development of the first formats for

automatic library interchange of bibliographic data (Marc, 1975), the first tools for distributed indexing (Harvest, 1999) and reading their history in the light of thesauri seen as main component of formal mediation languages.

References

Adomavicious G., Tuzhilin A. (2005). Toward the Next Generation of Recommender Systems: A survey of the State-of-the-Art and Possible Extensions. *IEEE transactions on knowledge and data engineering*, 17(6), 734-749.

Aitchinson, J., Gilchrist, A., Bawden, D. (2000). *Thesaurus Construction and Use: a Practical Manual.* London: Aslib.

Blair, D. C. (1990). *Language and Representation in Information Retrieval.* Amsterdam: Elsevier.

Brée, R., (1972). European Community, Center for Information and Documentation. In Allen Kentet al. (Eds.), *Encyclopedia of library and information science*, vol. 8. New York: M. Dekker.

De Michelis, G. (2007). The contribution of the language-action perspective to a new foundation for design. In Erickson, T., McDonald, D. (eds.), *HCI Remixed* (pp.293-299). Cambridge MA: MIT Press.

Dewey, M. (1996). *Dewey Decimal Classification and Relative Index*, edition 21. Albany, N.Y.: Forest Press.

Euratom Thesaurus (1966, 1967). *Euratom Thesaurus: part I, Indexing terms used within EURATOM's nuclear documentation system*, EUR report 500.e. Bruxelles, Information and Documentation Center; *part II, Terminology Charts Used in Euratom's Nuclear Documentation System*, EUR report 500.e. Brussels: Information and Documentation Center.

Euratom Treaty (1957). *Treaty Establishing the European Atomic Energy Community*, 298 U.N.T.S. 140, as amended in Treaties Establishing the European Communities (EC Off'l Pub. Off. 1987).

Eurovoc Thesaurus (2007). *Eurovoc thesaurus, vol. 1 Permuted alphabetical version*, parts A and B, *vol. 2 Subject-oriented version.* Luxembourg: Office for Official Publications of the European Communities.

Gueben, G. (1962). *Ispra, Centre commun de recherche de l'Euratom*, EUR report 54.f. Originally published in Bulletin d'Information de l'Association belge pour le Dèveloppement pacifique de l'Energie Atomique, 37.

Harvest (1999), *Harvest Web Indexing Package.* Retrieved September 26, 2008, from http://sourceforge.net/projects/webharvest/.

Lancaster, F. W. (1972). *Vocabulary Control for Information Retrieval.* Washington, DC: Information Resources Press.

Latour, B. (1996). *Aramis: or the love of technology.* Cambridge, Mass.: Harvard University Press. (Original published in 1992).

Mangani, G. (2006). *Cartografia Morale.* Modena: Franco Cosimo Panini.

MARC (1975). Machine-Readable Cataloguing Program. In Allen Kent et al. (Eds.), *Encyclopedia of library and information science*, vol. 16. New York: M. Dekker.

Oraganisational chart (2008), *Organizational chart of the JRC in July 2008.* Retrieved August 27, 2008, from http://www.europa.eu/.

Petri, C. A. (1977). Communication Disciplines. In Shaw, B. (Ed.), Proceedings of the Joint IBM University of Newcastle upon Tyne Seminar (pp. 171-183). United Kingdom: University of Newcastle upon Tyne.

Petri, C. A. (2001). Cultural aspects of net theory. *Soft Computing*, 5 (pp.141-145). Berlin: Springer Verlag.

Rolling, L. N., (1966). A computer-aided information service for nuclear science and technology.*Journal of Documentation*, 22(2), 93-115.

Serrai, A. (1997). *Storia della Bibliografia [Volume VIII]: Sistemi e Tassonomie*. Roma: Bulzoni Editore.

Spence, R. (2007). *Information visualization: design for interaction*, (2nd ed.). New York: Addison Wesley.

Textual Dynamics of the Professions (1991). In Bazermann, C., Paradis, J. (Eds.).Madison, Wis.: University of Wisconsin Press.

Winograd, T., Flores F. (1986), *Understanding Computers and Cognition: a New Foundation for Design*. Norwood: Ablex.