Mapping for Multi-Source Visualization: Scientific Information Retrieval Service (SIRS)

Dario Rodighiero[†], Matina Halkia[‡1], Massimiliano Gusmini[•]

[†]Arcadia S.I.T.
Via Mondovì, 4 I-20132 Milano (MI), Italy
fumoseaffabulazioni@gmail.com

[‡]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
matina.halkia@jrc.it

[•]Reggiani
Via Tonale, 133 I-21100 Varese (VA), Italy
massi@otolab.net

Abstract. This paper discusses the design process of a multi-index, multi-source information retrieval system (SIRS). SIRS provides comprehensive visualization of different document types for the JRC working environment. The interface design is based on elastic window management and on the Focus+Context method to browse large amounts of information without losing its contextual relevance. Source integration was achieved by mapping techniques, on which we applied methods, degree-of-separation and closure, to provide advanced relational context for objects.

Keywords: interface design, information visualization, mapping, multiple indexes, SIRS, adaptive interface.

1 Introduction

In recent years, many organizations have invested in semantic integration of information systems. Examples include the Terminology Services by OCLC, the mapping experiments by FAO, the CrissCross by the German National Library and Vascoda by the German Research Foundation [1]. This project also investigates the potential of document retrieval based on multiple indexes. In this case however, indexes are integrated through an adaptive interface as to enable browsing across

[1] The first two authors equally contributed to the paper.

different types of documents, reflecting the complex work environment of a public research organization.

The work was conducted at the Joint Research Centre (JRC), European Commission (2006-2009). The project aim was to establish a well-organized and comprehensive scientific information discovery system. The system is provided as a JRC Central Library service, and it is known as Scientific Information Retrieval Service, (SIRS). The technical development was completed in February 2009 while the content is still being expanded.

We will discuss this work using the Löwgren and Stolterman principles [2]. According to their principles, the design process is composed by the *Vision*, *Operative Image* and *Specification* phases. Löwgren and Stolterman believe that the interaction design process starts earlier than the traditional methodologies allow in information system development and software engineering. According to them, the "design of the design process," the questions chosen to be addressed at an early stage and requiring the "mind of a thoughtful designer," (how much time and importance is placed to the organization of the project, the users, the clients, the choice of technology: old and tested, or innovative and creative, etc.), greatly impact the end result.

The three Löwgren and Stolterman stages of design are:

- The *vision* which emerges early in the design process. Basic elements take shape: an idea, a function or an infrastructure. Intuition plays an important role in this early stage;
- The *operative image* is the first tangible expression of the vision. It takes the form of sketches, metaphors, analogies, scenarios and storyboards. By a process of cyclical refinement or an iterative dialectical process we obtain the final model;
- The *specification*, or final model, where all interface details are well-defined and ready to be integrated in the product construction.

To be sure, there is no clear division between design and construction, nor such definitive boundaries between the three phases. However, we have found the Löwgren and Stolterman approach to be very useful in organizing the development work on SIRS and in discussing it critically. Indeed, substantial time was spent in these three phases before construction started.

2 Designing SIRS

2.1 Vision

The underlying idea behind SIRS is to provide JRC users with a novel browsing experience by permitting control over three different knowledge sources, EU

legislation, JRC publications, and JRC Central Library holdings, separated in three different databases, but used interoperably in the working life of JRC scientists. Now, the SIRS system allows users to discover this information in one space – the SIRS interface – that accurately describes the complexity of the organization they serve.

In this phase, which is the investigative part of the design process, we identified the types of users, which led us to their needs. Then, we identified the type of documents they use in their daily working life, which in turn directed us to the indexes these documents refer to.

The design process started with these basic elements:

The *users*:

- The *researcher*, who should be able to retrieve the journal articles, the bibliographic references to books and journals, relevant scientific websites, and JRC publications in his specific field of interest;
- The *thematic programme leader*, who should be able to retrieve information about scientific institutions active in a specific field, relevant EU policy documents, and descriptions of the main JRC research themes in the relevant policy area or in the specific scientific field;
- The *JRC Action leader*, who should be able to retrieve European legislation, supported by JRC actions, and the internal publications output in support of the former.

The investigation of users' needs led us to identify three information *sources* whose items were designated as *documents*:

- The *EU legislation*, whose policy instruments indicate the present and future directions of the research in the JRC;
- The *JRC publications*, which comprises all JRC publications produced in the past;
- The *JRC Central Library holdings*, which includes books and both paper and digital journals, reflecting the scientific production outside JRC.

Since each *document* is classified according to its *source*, referring to a different indexing structure, we have three *indexes*:

- The *JRC Actions*, a taxonomy which reflects the internal JRC organization and provides the descriptors for profiling the JRC publications;
- The *Eurovoc Thesaurus* [3], a thesaurus that comprises more than six thousands index terms, developed specifically for organizing EU legislation, by the European Parliament and the Office for Official Publications of the European Communities (OPOCE);
- The *Dewey Decimal Classification* [4], an internationally applied decimal system of library classification with more than ten thousands captions, which index Library holdings.

2.2 Operative Image

Once we gathered the basic design elements, the information was translated in scenarios. In the process of writing these scenarios, we realized the importance of integrating the three different types of *indexes* and *documents*. Thus, when we started sketching the interface, the Focus+Context paradigm [5] was introduced as a natural way to manage the quantity of information while keeping it contextually relevant.

Two assumptions were made early on: the contextual relevance of information and the adaptability of the interface. On one hand, to keep the contextual relevance of information visible to the user, the SIRS interface space should include all *indexes* and all *documents* at all times. On the other hand, to provide an adaptive interface to the variable user focus, the screen organization should be elastic in order to accommodate the movement of users' attention. The first assumption implied a rational spatial organization of the screen in modules, and the second dictated the behaviour of the geometric characteristics of these modules, which we call *frames*.

Each *frame* is rectangular and tagged by a label, which indexes the content. The first division reflects the source composition: *index* on the left and *documents* on the right, according to the direction of western writing. Then, each *frame* splits horizontally or vertically as needed for inserting three *indexes* and three *document* sets. *Indexes* are aligned vertically and *documents* horizontally.

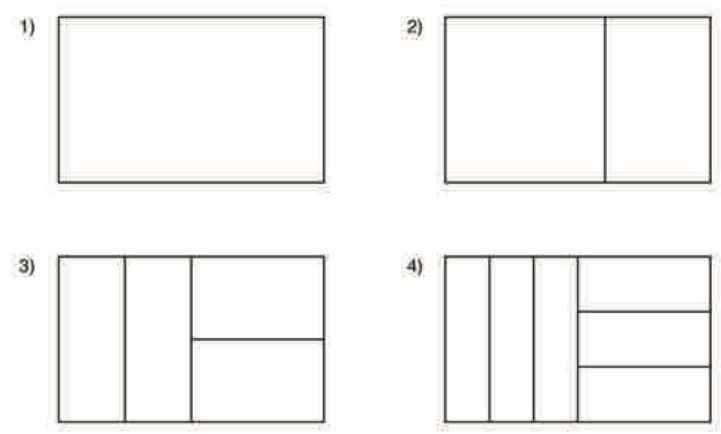

Fig. 1. Succession of *frames* accommodation.

To make the interface space elastic, we enhanced the SIRS interface with a mechanism to enable *frames*' size customization: in three actions, the users could change the *frames*' size according to the movement of their focus area. During the change from an arrangement to another, the visual continuity is ensured by the use of a smooth movement of all *frames*. The three actions are:

– *Enlarge*, which makes the focused *frame* double-sized when others are still present in the same area;
– *Maximize*, which makes the focused *frame* as big as possible;
– *Normalize*, which restores default size values.

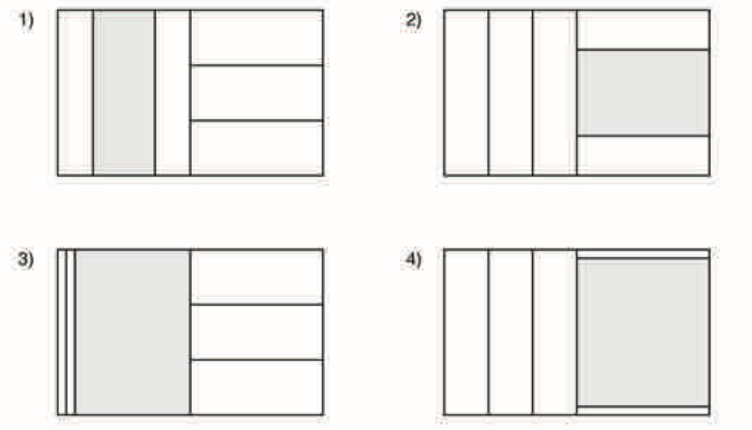

Fig. 2. *Frames* customization on 1) *index* enlargement, 2) *documents* set enlargement, 3) *index* maximization and 4) *documents* maximization.

Frames host objects. Objects belonging to an *index* are called terms. Vertical *frames*, which host indexes, display terms in two ways: as a multiple-level structure or as a simple list. Multiple-level structures allow browsing through *index* hierarchies. By default, the SIRS interface provides terms presented as a simple list.

Horizontal *frames* display *documents* only as a list. The *documents* are listed chronologically, both for EU legislation, and JRC publications, while it is alphabetical for Library holdings. Due to the fact that the list of *documents* could be very long (in the thousands, in the case of EU legislation), we decided to limit the number of visible *documents* to less than forty. The complete list of *documents* is available by an appropriate link at the bottom of the *frame*. This accommodates needs for different user experience.

2.3 Specification

Once the *operative image* has been discussed sufficiently we can move to the system *specification*.

The first decision we made for the specification of the interface was to use Adobe Flash technology. It is a standard tool for cross-browser compatibility, and offers advanced capabilities in terms of user experience.

Another important decision we took for managing the amount of data we had to manipulate was to access by query the remote data repositories through three different web services: EU legislation, JRC publications and Library holdings. We chose not to store the data locally using an ad hoc repository; only integrate information in the SIRS interface. Through web services, the system is able to retrieve *documents'* metadata and to point to relative websites for *documents'* reading.

The next problem we faced and seemed to be insurmountable was that there were no tools available to manage multiple indexes. In fact we had two major technical issues to solve: the first one was the necessity to manage more than one index, and the second was the requirement to create cross-index relations. To respond to these needs, we decided to build the tool ourselves. We call it Thesaurus Management Service (TMS).

TMS is an open-source application based on the free PHP technology. Oracle is used for implementation. In order to keep the tool compatibility with open platforms, the system is also available for MySQL. TMS is developed according to the British Standard for Thesauri Construction [6], which manages indexes as concepts and terms. Full TMS development being onerous, the JRC works in partnership with the FAO (Food and Agriculture Organization) and the BGS (British Geological Survey) to complete this project, which is also called Thesaurus Tool Code [7].

In order to enable multi-source browsing, we had to create a unique network composed of terms and *documents*. To unify different sets of *indexes* and *documents* in a single interface, we used the technique of *mapping*, which enables cross-index relations to connect terms located in different *indexes*.
The interface is based on multiple window management, more specifically the concept of *elastic windows*, which allows a variable screen real estate to be used according to the role or task performed [8]. As discussed earlier (2.2) there is a basic left-vertical-*index* to right-horizontal-*document* screen division. Both *indexes* and *documents* are considered *sources*. EU legislation is the backbone among *sources*. In fact the Eurovoc is placed in the middle of the three *indexes* as a bridge. The other two are JRC Actions and Dewey. On the left side, the Eurovoc is mapped to the JRC Actions. Each JRC Action is classified by a set of Eurovoc terms. The relations created through the Eurovoc terms enable cross browsing from EU legislation to JRC publications and vice versa. On the right side, the Eurovoc is mapped with the Dewey. This mapping, which is based on semantic similarity defined by Martin Doerr [9], allows the navigation from EU legislation and Library holdings in both directions.

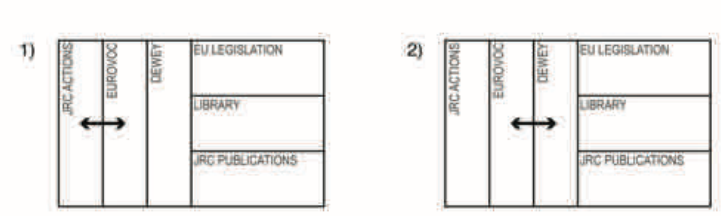

Fig. 3. The SIRS mapping is based on two types of cross-relations: 1) one between Actions and Eurovoc terms and 2) one between Eurovoc and Dewey terms.

Once the relations network was set up, we proposed a method for visualizing information based on graph navigation. For a selected term, we can navigate all relations in order to get the set of terms and *documents* at the first degree of separation. Moreover, we can use that set as input to receive a new set of terms and *documents* at the second degree of separation from the selected term. The degree of separation is the measure that we use to define the context.

The methodology consists in getting *indexes* at once, then in retrieving related *documents*. By selecting a term in the middle *index* the user follows the shortest route to the objects. The contextual distance is not perceived by the user, but is used to populate the interface. For example, if a Eurovoc term is selected, two degrees of separation are in place. At the first degree, the resulting set includes EU legislation *documents*, JRC Actions and Dewey objects (Fig. 4.1). At the second degree, we use only the two *indexes* (JRC Actions and Dewey) as input for retrieving *documents* from JRC publications and the Library holdings (Fig. 4.2).

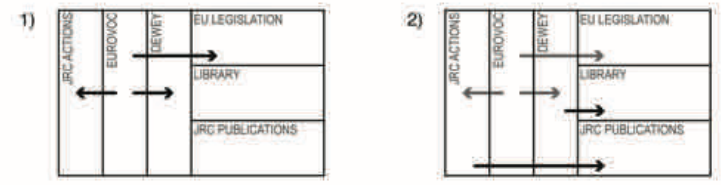

Fig. 4. SIRS composes the context from a Eurovoc term respectively using 1) the first and 2) the second degree of separation.

On the contrary, when a term in one of the two lateral *indexes* (JRC Actions and Dewey) is selected, three degrees of separation are in place in the resulting context. For example, if a JRC Action is selected, the first set of related nodes is composed by JRC publications and Eurovoc terms (Fig. 5.1). The second set, which uses only Eurovoc terms as input, includes EU legislation *documents* and Dewey terms (Fig. 5.2). The third set is obtained using Dewey terms and contains Library *documents* (Fig. 5.3).

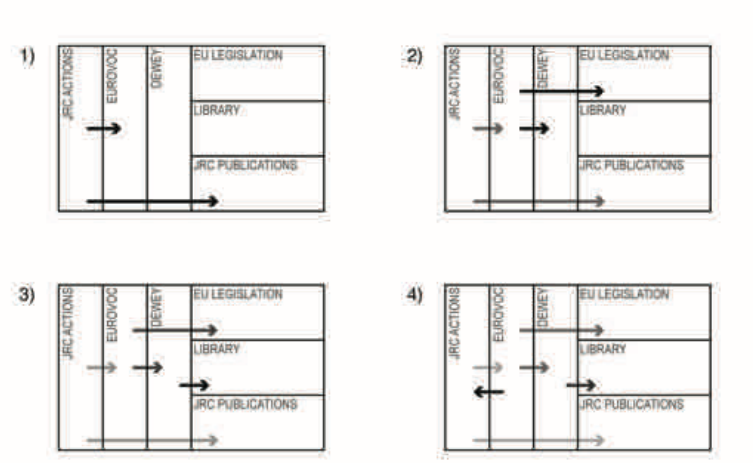

Fig. 5. SIRS composes the context from a JRC Action term respectively using 1) the first, 2) the second and 3) the third degree of separation. 4) The last picture represents the relation used for *closure*.

Moreover, in the second example, a behavior we call *closure* [10] can be observed. *Closure* is a way to suggest similar objects. For example, when we select a JRC Action, we obtain a set of Eurovoc terms. As EU legislation *documents* are retrieved from the latter, we can also retrieve other JRC Actions (Fig. 5.4). This kind of operation - similar to a boomerang's flight path - suggests a group of JRC Actions, which share at least one Eurovoc term with the initially selected JRC Action. *Closure* provides an interface with advanced relational context.

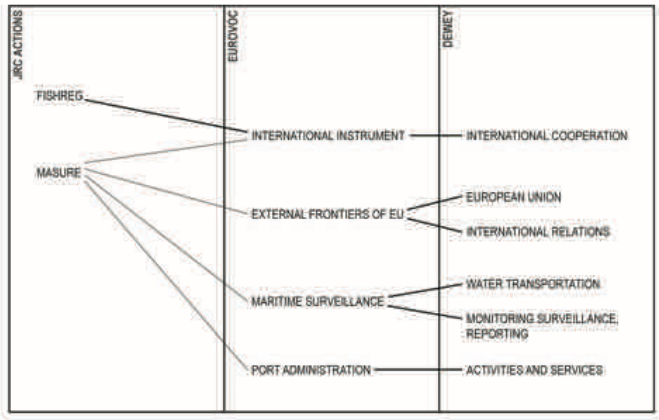

Fig. 6. This picture shows how the JRC Action MASURE is connected to JRC Action FISHREG by the sharing of a Eurovoc terms. *Closure* makes this relation available to users.

As seen in the previous examples the method has four types of parameters:

- The *input*, in the form of an array of sets;
- The *output*, an array of sets too;
- The *degree of separation*, an integer;
- The *selected term*, an ID number.

To avoid inserting the wrong numbers of parameters, we have to remember that pairs of input and output sets must have the same numerical occurrence as the degree of separation. For each degree of separation, a pair of input and output sets has to be declared. Practically this means that for two degrees of separation there are two pairs of input and output sets, for three degrees, three pairs and so forth.

3 Discussion

We have discussed multi-index, multi-source information retrieval in the context of a scientific knowledge management service. We have presented the design-process stages and discussed them according to the Löwgren and Stolterman levels of abstraction. Types of users, *documents* and *indexes* were introduced, as well as their contemporaneous organization on the screen area based on the Focus+Context paradigm by means of *frames* and objects. We then discussed the techniques of mapping, degree-of-separation and closure and how these were used to provide comprehensive, customized information visualization for the JRC working environment, as well as advanced relational context between objects. Finally, we hinted at the technical functions supporting the system.

Although the Löwgren and Stolterman approach has proven to be extremely useful, in fact the construction process starts at the moment technical issues are addressed, which can happen even at the earliest stages of the design process, and vice versa: technical issues reformulate the design process. In the three abstraction levels of their theory, we would add an additional one: construction.

Multi-index systems have been used for a while in information retrieval, including advanced mapping techniques. However, little has been done to translate these information structures in adaptive interfaces providing accessibility and transparency to the information networks underlying multi-index integration. We propose an interface for accommodating multiple *sources* of different type simultaneously on the screen, which enables information discovery by visualizing contextual relations between objects, by degree of separation and closure.

4 Acknowledgements

Carlo Ferigato contributed in many ways on SIRS, especially on the mapping specification. Daniela Panfili contributed to the interface design. Giovanni Salvagno

has been an enormous technical resource. Giuseppe Merlo, in the JRC Central Library, provided general background in information science. Finally, we wish to thank Marc Wilikens, under whose responsibility this project was undertaken and financed at the JRC.

Reference

1. Mayr, P., Petras, V.: Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: IFLA World Library and Information Congress. Québec (2008), http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf accessed on March 12, 2009
2. Löwgren, J., Stolterman, E.: Thoughtful Interaction Design: A Design Perspective on Information Technology. The MIT Press, Cambridge (2004)
3. Eurovoc thesaurus, vol. 1 Permuted alphabetical version (parts A and B), vol. 2 Subject-oriented version. Office for Official Publications of the European Communities, Luxembourg (2007)
4. Dewey, M.: Dewey Decimal Classification and Relative Index, edition 21. Forest Press, Albany (1996)
5. Spence, R.: Information Visualization: Design for Interaction, 2nd ed. Pearson/Prentice Hall, Harlow (2007)
6. Official development website for BS 8723, http://schemas.bs8723.org/ accessed on March 12, 2009
7. Thesaurus Tool Code, https://www.assembla.com/wiki/show/thesaurusToolCode/ accessed on March 12, 2009
8. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. Addison-Wesley, Harlow (1999)
9. Doerr, M.: Semantic Problems of Thesaurus Mapping. Journal of Digital Information, vol. 1(8) (2001), http://journals.tdl.org/jodi/article/viewArticle/31/32 accessed on March 12, 2009
10. Ferigato, C., Merlo, G., Panfili, D., Rodighiero, D.: Role of Thesauri in a Scientific Organization. In Networks of Design, Design History Society Conference 2008 (in press). Brown Walker Press, Boca Raton (2009)
11. Arms, W.Y.: Digital Libraries. The MIT Press, Cambridge (2000)