



# Spatially-Aware Information Retrieval on the Internet



SPIRIT is funded by EU IST Programme  
Contract Number: IST-2001-35047

## *Evaluation of SPIRIT prototype following integration and testing*

<b>Deliverable number:</b>	D31 7301
<b>Deliverable type:</b>	PU
<b>Deliverable nature:</b>	Final
<b>Contributing WP:</b>	WP 7
<b>Contractual date of delivery:</b>	30/06/05
<b>Actual date of delivery:</b>	14/06/05
<b>Authors:</b>	Bénédicte Bucher, Paul Clough, David Finch, Hideo Joho, Ross Purves, Awase Khimi Syed
<b>Keywords:</b>	Evaluation, relevance, test collection, user tests

**Abstract:** The generic evaluation methodology defined in D20 has been refined to apply to SPIRIT final prototype. It includes user tests to evaluate how well users perform with SPIRIT, possibly as compared with other similar systems. It also includes performance testing based on a specific Test Collection.

# Contents

<b>1. INTRODUCTION.....</b>	<b>5</b>
<b>2. USER CENTERED EVALUATION.....</b>	<b>5</b>
<b>2.1. The limits of detailed scenarios .....</b>	<b>5</b>
<b>2.2. New protocol.....</b>	<b>6</b>
<b>3. SYSTEM CENTRED EVALUATION .....</b>	<b>7</b>
<b>3.1. Absolute performance and comparative performance .....</b>	<b>7</b>
<b>3.2. Test collection in IR.....</b>	<b>8</b>
<b>3.3. Document collection .....</b>	<b>9</b>
<b>3.4. Topics.....</b>	<b>9</b>
<b>3.5. Relevance judgement scheme.....</b>	<b>11</b>
<b>3.6. Building relevance judgements.....</b>	<b>12</b>
<b>4. RESULTS .....</b>	<b>13</b>
<b>4.1. Available functionalities.....</b>	<b>13</b>
<b>4.2. Results for user-centred evaluation.....</b>	<b>14</b>
<b>4.3. Results for system-centred evaluation .....</b>	<b>16</b>
<b>5. CONCLUSION .....</b>	<b>16</b>
<b>6. REFERENCES.....</b>	<b>17</b>
<b>7. APPENDIX A : MAIL SENT TO USERS FOR A FIRST CONTACT .....</b>	<b>17</b>
<b>8. APPENDIX C : QUESTIONNAIRE .....</b>	<b>18</b>
<b>QUESTIONNAIRE ON THE SPiRiT FINAL PROTOTYPE .....</b>	<b>19</b>
<b>9. APPENDIX D : DOCUMENTS RETURNED BY USERS .....</b>	<b>22</b>
<b>9.1. User 1. ....</b>	<b>22</b>
<b>9.2. User 2 .....</b>	<b>26</b>

**SPIRIT project**

*Evaluation of SPIRIT prototype following integration and testing*

IST-2001-35047

D31 7301

<b>9.3.</b>	<b>User 3 .....</b>	<b>30</b>
<b>9.4.</b>	<b>User 4 .....</b>	<b>33</b>

# Executive Summary

This deliverable presents the evaluation protocol and results of SPIRIT final prototype. It is composed of the user-centred evaluation and the system-centred evaluation.

User-centred evaluation aims at measuring SPIRIT effectiveness and user satisfaction through the performance of the main user task "retrieving documents relevant to a query". We proposed a protocol adapted to the specific nature of relevance. Results are rather consensual. Users criticise the low speed and the lack of ergonomics of SPIRIT. And they praise the underlying concepts and the user interface design.

System-centred evaluation aims at measuring the system recall and precision. We have designed a specific test collection for this purpose. Results highlight a best ranking method : angle distributed. They also show improvements brought by spatial awareness in precision and recall.

**Acknowledgements** : The authors wish to thank Frédéric Hubert, Lars Harrie, Nicolas Renault and Andrew Faulk for their contribution.

# D31 7301

## ***Evaluation of SPIRIT prototype following integration and testing***

### **1. Introduction**

A generic evaluation methodology for SPIRIT prototypes has been proposed in [Aumaitre et al. 04a]. It is composed of two major steps :

- The system validation, i.e. evaluating how well the system works, with its technical specifications as a reference, relies on two tests : One is to compare the functionalities that appear in the system specification and the functionalities that are really available in the system. The other is IR performance analysis based on measuring of two criteria: "Recall" (probability that a relevant document is retrieved) and "Precision" (probability that a retrieved document is relevant). This analysis is based on a Test Collection.
- The user validation, i.e. checking if the system fits user requirements, is based on an opportunity for users to use SPIRIT. There are three main aspects to be considered: effectiveness, efficiency and user satisfaction. In the general methodology, two tests are proposed.
  - In the first test, users are asked to perform search tasks that are explained in test scenarios. When the searches are completed, they are asked to fill in a satisfaction questionnaire. The test ends with an interview session.
  - The second test is a comparative usability test in which users compare SPIRIT's interface with other search engines' interfaces.

In this deliverable, we describe the specific evaluation protocol for the final prototype and the results of the evaluation of the final prototype following this specified protocol. Section 2 describe the new user tests protocol. Section 3 present the design of a Test Collection to measure precision and recall. Section 4 presents evaluation results.

### **2. User centered evaluation**

#### **2.1. The limits of detailed scenarios**

The generic evaluation methodology we must specify for evaluating SPIRIT final prototype had already been specified and applied to the evaluation of the interim prototype [Aumaitre et al. 04b]. This evaluation had highlighted limits in user tests. Users mentioned they could hardly be sincere when evaluating SPIRIT after the protocol proposed to them. This protocol was a set of scenarios leading them to accomplish user tasks through SPIRIT interim prototype.

During the project, we had other opportunities to organise *user sessions* somehow related to the main user task, that is finding relevant documents with respect to a user query.

- During user requirements reassessments, we used fixed scenarios illustrated through “snapshots” of a mock up to explain to users how SPIRIT should support spatially aware information retrieval.
- During relevance scheme test, assessors were asked to judge the relevance of documents, retrieved in the context of a specific scenario we gave them, after a proposed relevance judgement scheme.

It is important to underline that, on each of these occasions also, several users or assessors complained about the difficulty of assessing document relevance with respect to a topic they had not defined themselves. In other words, the notion of relevance is not formal enough for people to assess it sincerely in the context of too detailed scenarios.

[Borlund 03] extends upon the complex nature of document relevance in information retrieval as being dynamic, i.e. it evolves in time, and cognitive. The author recommends the use of simulated task situations made up of an objective and a strategy to achieve it through the interface of the interactive system to be evaluated. Yet, in this context, the objective described is not far from the real objectives of the user. In SPIRIT context, using simulated task situations where the objective is close to the user objectives is not possible unless we recruit a set of specific users whose needs for information fit in the limits of the SPIRIT document collection and geographical data.

## **2.2. New protocol**

We decided to specify the user tests protocol in a different manner than that used for the interim prototype. In order to put the emphasis on respecting this specific nature of "relevance of a document with respect to a query", this new protocol is not based on detailed scenarios and puts less control on the use context.

### **Protocol**

In the new user-centred evaluation protocol for the final prototype users are proposed to try out freely any search they want on the SPIRIT final prototype interface described in [Purves and Yang 05]. A help document is available on-line that contains :

- explanations for SPIRIT misbehaviours due to limited data and limited document collection,
- guidance elements to use the graphical interface,
- an example scenario the prototype has documents and data for.

It is important to keep users in a context close to the context in which they use search engine. In this new protocol, they don't have to try out SPIRIT in a specific place and at a specific time. They are given a period of time (8 days) and proposed to get a hot-line if they specify when they will try SPIRIT out.

8 users have been contacted, different from those that evaluated the interim prototype. They are given some background information about SPIRIT and a brief description of the testing protocol (see Annexe A).

During one week –initial duration which has been extended up to a month- they could log on SPIRIT prototype search engine and perform any query they wanted. The document they could browse to get help information is in Annexe B. The questionnaire is in Annexe C. The Annexe D lists the documents returned by users. One wished to answer in French.

We have contacted users with various nationalities (French, English, Scottish, Sweden and Belgian), aged between 27 and 60. A third of them are researchers because we expected researchers to have more curiosity and patience to try out a research prototype and keep in mind this is not commercial software.

At the time of writing this deliverable, not all of them have returned the questionnaire and the profiles of these who returned it are less distributed. They are all men, aged between 29 and 45 and three of them are researchers.

**Not scenario-based**

Fixed scenarios like “you are an English teacher going on holidays in the Swiss Alps with your wife and two children...” were an impediment to users being sincere. Yet, they were supposed to fulfill several functions in SPIRIT user tests :

- 1) Scenarios let users go through every important user task.
- 2) Scenarios prevent users from facing SPIRIT prototype misbehaviours due to documents and data limits, which is likely to happen if they try any query. For instance, if someone is looking for "kangaroos near Beinwill", the placename Beinwill happens not to be in the ontology, and there are no documents about this topic in the collection.
- 3) Scenarios set the context of a query so that results from several users are more homogeneous and can be aggregated.

Not using scenarios means these functions should be fulfilled another way or not fulfilled at all.

Concerning the first function of scenarios, “to let users go through every important task”, there is only one main user task on a spatially aware search engine : “to find relevant documents to their query”. Users are familiar enough with search engines to know the subtasks of this main task, i.e. to know the big cover story : to express a query, to send it, to explore the results. Moreover, in the new protocol, they may browse a document available on SPIRIT interface that contains an example scenario for them to explore these tasks.

Concerning the second function, the new protocol explains to the user the possible misbehaviours of the system, and suggests to him a scenario we know the prototype has geographical data and documents for.

Concerning the third function, a counterpart of respecting the subjective nature of relevance of a document to a user query is to possibly get heterogeneous context.

**Comparison with other systems**

The preceding evaluation did not include comparison with similar systems. For the evaluation of the final prototype we included some comparative evaluation elements. Users were asked to try out the same tasks on several applications :

<http://local.google.co.uk> : a Web application dedicated to searching a business and services directory by theme and by location.

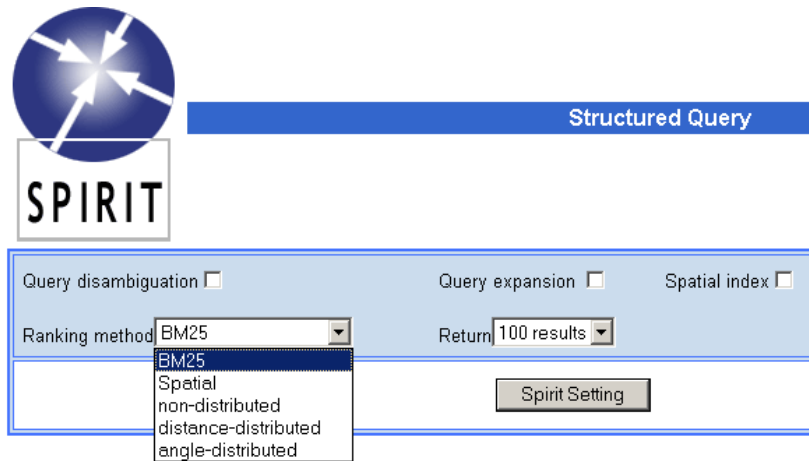
<http://www.yell.co.uk> : the on-line yellow pages of UK

<http://www.map24.com> : a mapping and routing Web application.

### **3. System centred evaluation**

#### **3.1. Absolute performance and comparative performance**

An important feature of SPIRIT final prototype is the possibility of selecting, for some of its components, one method among several that are implemented in the component. This feature is available through the link “preferences” on a user interface especially developed for this purpose and illustrated on Figure 1. It is possible to use a textual index only or to use the spatial index. It is possible to choose between the classical BM25 ranking method and several specific spatial ranking methods.



2004 Spirit

*Figure 1. Access to parameters settings through a user interface dedicated to technical iterative development and to evaluation.*

In the evaluation, a first task consists in comparing performance measures obtained by the different implemented methods that contribute to precision and recall, namely the index and the ranking methods.

Another task consists in measuring SPIRIT prototype performances with its optimal settings.

### **3.2. Test collection in IR**

In the generic methodology, system-centred evaluation is mainly based on using a test collection to measure system precision and recall [Aumaitre et al. 04a]. An IR test collection consists of the following components:

**A document collection** - a set of documents representative of a selected domain.

**Topics** - a set of typical user information needs based on the document collection.

**Relevance assessments** – manually assessed for the documents of the collection with respect to each topic. This is usually a list of relevant documents per topic.

Such a collection is used as a reference to measure a search engine precision and recall. These are measured by running a search engine on the document collection, with a topic, and keeping the TopN returned documents. Precision is the proportion of documents in the TopN that are relevant. Recall is the proportion of relevant documents that are retrieved. This is illustrated on Figure 2.

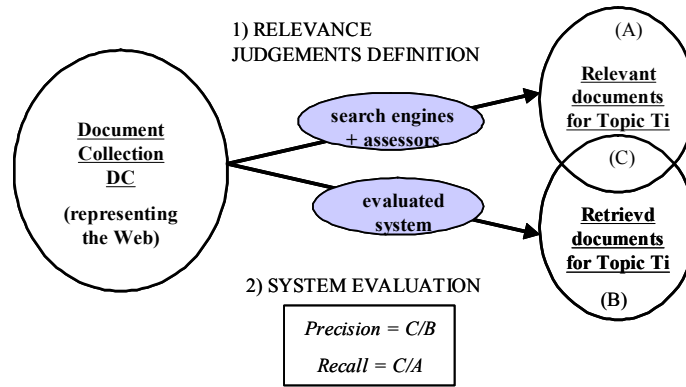


Figure 2. Measuring a search engine precision and recall with a Test Collection.

In another SPIRIT deliverable, [Sanderson and Joho 03] describe the various methods to build a test collection for SPIRIT.

In IR, standard test collections are built so that measures of precision and recall for several search engines can be compared. In SPIRIT, we tried to provide elements that could further contribute to building a standard test collection for spatially aware search engines.

### 3.3. Document collection

The aims in creating a document collection are firstly, to produce a representative sample of the whole set of documents which may be queried and secondly, to build a collection small enough that it is possible that relevance judgements may be performed for queries on every document in the collection. In the case of the SPIRIT, the Test document collection is the same as the final prototype document collection.

Building this document collection was done early in SPIRIT. It is the input for the documents annotation and for building the spatial index. The method used to build it is described in [Joho and Sanderson 04].

### 3.4. Topics

Topics to be used in the Test Collection should correspond to "something related to somewhere".

#### Topics for which classical search engines fail

The most important requirement expressed by users was to have efficient IR functions, i.e. to actually find documents that were relevant with respect to their query. The nature of queries should test the capabilities of GIR which are not available in standard IR systems and where spatial awareness is required. These are listed on Figure 3.

- The name of a location is also a more commonly used word (e.g. *Battle*)
- The location name is ambiguous (e.g. *London*, Ontario or *London*, England)
- The location name refers to an imprecise region (e.g. The *south of France*)
- The location name is unlikely to be referred to in relevant web documents (e.g. *blogs near Vélieux*)
- The theme of the query includes a (spatially) irrelevant location (e.g. *Elgin Marbles in Athens* – Elgin is also a town in the north of Scotland)
- The query includes a non-containment based spatial relationship (e.g. *golf courses north of Aberdeen*)

Figure 3. Definition scheme for topics testing spatial awareness in SPIRIT test collection.

#### Topics within the documents and data limits of SPIRIT final prototype

Classically, topics must be picked within the "coverage" of the document collection. In SPIRIT, not only should the topics be picked within the "coverage" of the collection of geo-tagged documents but also within the "coverage" of SPIRIT spatial awareness, i.e. the spatial extent and granularity of geographical data that have been acquired by the SPIRIT consortium. The objective of SPIRIT was indeed not to propose a spatially aware search engine with the largest geographical scope of awareness but to propose a prototype that demonstrates new concepts and methods.

#### Favouring topics number over "objective" judgements

As proposed by [Sanderson and Joho 03], we favoured the number of topics over the objectivity of relevance judgements. In other words, given the limited number of assessors resources, we won't do inter-assessor relevance judgements. Consequently, each assessor defined himself his topics and provided a brief description of them -but not an imaginary scenario- .

#### Final list of topics

Topics for the test collection were defined after the topics definition scheme defined in Figure 3 and with respect to the specific coverage of the document collection. It is the following list :

- Hotel near Horgen
- Museum near Trinity (in Edinburgh, UK)
- Museum near Rapperswil (in St Gallen)
- Skiing south of Zürich
- Castles east of Edinburgh
- Horse racing near Edinburgh
- Cycling near Bath
- Hotels in Caerdydd
- Museum near Riverside (in Cardiff)
- Fishing west of Cardiff
- Caravanning outside of Cardiff
- Castles within 50km of Cardiff

**3.5. Relevance judgement scheme**

An initial scheme was proposed to score both the thematic and spatial relevance of a document with respect to a query of the form “something related to somewhere”. A global judgement is made up of two scores : thematic relevance score and spatial relevance score. Within each type of relevance, three levels of relevance may be distinguished : highly relevant (score 1), relevant (score 2) and non relevant (score 3). The description of the initial scheme is detailed below :

Thematic relevance :

- Score 1 means the document contains relevant information about the concept queried AND on its own allows you to form a judgment about the document (i.e. requires no external knowledge).
- Score 2 means the document is relevant, since it points to a resource MENTIONING the concept, but you must consult further pages referenced by the document to perform a judgment.
- Score 3 means the document does not provide information about the concept provided.

Spatial relevance :

- Score 1 means the document refers to a location that is/near the query location AND you think that the location in the document has sufficient detail for you to find it on a local map of the area
- Score 2 means the document refers to a location that is in/near the query location BUT you think that there is insufficient information for you to find that location on a local map of the area
- Score 3 means the document does not fall within the query location

Spatial relevance is intended to address issues of granularity that is the detail, as well as the relevance, of spatial information in a document. Importantly, it has some theme dependence, since the resolution of the needed local map corresponds to the spatial granularity of the theme. For instance, for the query “churches in \*” the spatial relevance will be high (1) if a full address is given within a city, but still high if only a village name is given for a small village. On the other hand, if only a town or a county name are given we would consider the spatial relevance to be (2).

Since this scheme was new it was decided to test its usability. A total of 10 documents were retrieved for each of 5 queries (Table 2), and 11 subjects judged each document for its spatial and thematic relevance (giving a total of 1100 judgements). The scheme was illustrated through the use of examples to the relevance assessors.

Caving in Derbyshire (UK) Castles in Wales (UK) Skiing near Glencoe (UK) Art festivals in Edinburgh (UK) Music in Montreux (Switzerland)
--

Table 2: Queries used in testing relevance scheme

A simple questionnaire was devised to investigate how well users understood and could apply these schemes. In general, users found the schemes both easy to understand and suitable, but they were less confident in applying the spatial scheme.

We believe this difficulty results when a user is not familiar with an area, and clues to spatial relevance are not simply the place name in the query. For instance, a ski area near Glencoe exists but is called the „White Corries“. A Swiss or French relevance assessor is unlikely to have this detailed level of local knowledge and thus finds it hard to judge the spatial relevance

of the page returned. However, if the user is presented with results and a map showing the location of the website, then the spatial relevance of the document is confirmed. This has two implications for relevance assessments in GIR. Firstly, relevance assessment of documents with respect to spatial aspects is best done by assessors with local knowledge of the area under query. Secondly, care should be taken in measuring spatial relevance since users unfamiliar with an area are likely to classify spatial relevance using contextual clues (e.g. locational query words in the document).

We further assessed agreement between assessors for spatial and thematic relevance using a multirater Cohen's Kappa for all 50 documents across the 5 topics. For both thematic and spatial relevance the agreement between annotators was found to be significant across the entire set of tests at the 95% level. However, many assessors also commented on this measure of thematic relevance being unnecessarily cumbersome and expressed a preference for a binary scheme. Thus, the adopted scheme for the evaluation was eventually a standard binary scheme for thematic relevance and the ternary scheme exposed here for spatial relevance.

**Thematic relevance :**

**Score 1** means the document is relevant since it points to a resource mentioning the query concept.

**Score 2** means the document does not provide information about the query concept.

**Spatial relevance :**

**Score 1** means the document refers to a location that is/near the query location AND you think that the location in the document has sufficient detail for you to find it on a local map of the area

**Score 2** means the document refers to a location that is in/near the query location BUT you think that there is insufficient information for you to find that location on a local map of the area

**Score 3** means the document does not fall within the query location

*Figure 4. Relevance judgement scheme adopted for the evaluation of the SPIRIT final prototype.*

### **3.6. Building relevance judgements**

For each topic, relevance judgements were built by one assessor only using “interactive searching and judging” (ISJ) method presented in [Aumaitre et al. 04a]. Besides, when the assessor was not familiar with the place the topic was about, he could use paper maps or mapping web sites to assess the spatial relevance of a document.

A total of five people (called assessors) were involved in the generation of relevance judgments. The assessors spent between 30 minutes to 3.0 hours to find as many relevant documents as possible for a topic. A total of 2228 documents were judged for 15 topics with the average of 148 documents per topic. The number of documents judged varies between 40 and 347 across the topics, due to the resource available to each assessor. The detail of the outputs from our ISJ work is shown in Table 1.

Topic	Spatial			Thematical		No. of docs judged
	1	2	3	1	2	
SP001	8	6	26	9	31	40
SP002	36	11	3	42	8	50
SP003	40	13	10	42	21	63
SP004	43	14	28	44	41	85
SP011	221	43	40	233	71	304
SP012	274	46	25	245	100	345
SP013	271	16	13	234	66	300
SP014	139	7	2	118	30	148
SP015	259	59	29	309	38	347
SP016	97	22	12	83	48	131
SP031	25	27	3	50	5	55
SP032	46	24	14	44	40	84
SP033	70	46	34	81	69	150
SP034	66	10	7	57	26	83
SP035	23	8	12	20	23	43
Total	1618	352	258	1611	617	2228

*Table 1 Summary of ISJ*

As can be seen, the number of non-relevant judgements is similar across the two types of relevance suggesting that the relevance should not be skewed towards one side. For all topics except Topic SP035, a greater number of the score 2 was found than the score 3 for the spatial relevance.

The Trec\_Eval software from the Lemur Toolkit (<http://www.lemurproject.org/>) will be used to do measure precision and recall. This software runs on binary judgements only. We generated several binary judgements out of SPIRIT compound judgements :

- 2\_3: this set consists of documents that are judged as the score 2 in the thematic relevance and the score 3 in the spatial relevance
- 2\_2: this set consists of documents that are judged as the score 2 in the thematic relevance and the score 2 in the spatial relevance
- 2\_0: this set consists of documents that are judged as the score 2 in the thematic relevance
- 0\_2: this set consists of documents that are judged as the score 2 in the spatial relevance
- 0\_3: this set consists of documents that are judged as the score 3 in the spatial relevance

## **4. Results**

### **4.1. Available functionalities**

In the following, we merely list all the functionalities that were specified for SPIRIT final prototype and if they are actually available in the final prototype :

Textual interface (what-where-how) :	Yes
Spatial relationships (“in”, “North Of”, “East Of”, “South Of”, “West Of”, “Near”) :	Yes
Graphical interface :	Yes
Sketching interface :	No (not foreseen on final prototype, available on a demonstrator)
Fuzzy geographical areas :	Yes (only certain types of fuzzy areas)
Textual result presentation :	Yes
Graphical result presentation :	Yes (on the advanced interface)
Result presentation choice :	Yes (on the advanced interface)
Usual relevance ranking :	Yes
Ranking respect to spatial relation :	Yes
Relevance ranking choice :	Yes
Search refinement :	Yes (on the advanced interface)
Choice of number of results to display :	Yes
Query disambiguation (spatial) :	Yes
Query disambiguation (non spatial) :	No
Query expansion :	Yes
Text indexing :	Yes
Spatial indexing :	Yes
Use of Google web API :	No
Logging feature :	Yes

Some functions are available on an advanced interface only because they are java applets and we thought they would have constrained too much user tests. Graphical result presentation has been refined on the advanced interface.

The Google Web API has not been used in the prototype but has been used during SPiRiT work on describing fuzzy areas.

#### **4.2. Results for user-centred evaluation**

The results presented here are compiled after documents returned by users listed in annexe D. When a user circled a number on the scale, we had to change it because the scale was incorrect (the number 2 was missing). When users used the visual scale (i.e. actually circled numbers or used the number 6 during their marking), we interpreted its position instead of the value (4 means 3 on a scale from 1 to 5).

	user1	user2	user3	user4	average
<b>overall reaction to the system</b>					
terrible/wonderful	3	2	4	3	3
frustrating/satisfying	2	2	2	2	2
dull/stimulating	4	4	3	5	4
difficult/easy	3	5	4	5	4
rigid/flexible	3	5	3	5	4
<b>specific items</b>					
characters hard/easy to read	5	5	5	4	5
info on screen confusing/very clear	3	5	4	4	4
sequence of screens confusing/very clear	3	5	4	5	4
use of terms inconsistent/consistent	4	2	4	4	3,5
useful messages never/always	3	4	3	3	3
error message unhelpful/helpful	2	4	-	-	3
help messages unhelpful/helpful	4	4	3	-	4
learning difficult/easy	5	5	4	5	5
exploring difficult/easy	4	4	3	3	3,5
straight-forward never/always	3	4	4	4	4
supplemental material confusing/clear	4	3	3	-	3
system speed too slow/fast enough	2	1	2	-	2
system unreliable/reliable	2	3	3	-	3

*Figure 5. Marks given by users to SPIRIT prototype (on a 1 to 5 scale).*

At the time of writing this deliverable, we have only four questionnaires returned but since some marks are rather consensual, these might be meaningful.

SPIRIT effectiveness is rather good. The results returned by SPIRIT to users query were relevant in most cases.

SPIRIT is perceived as rather frustrating and rather stimulating, which we think go together. Users put great expectations on SPIRIT concepts and are all the more frustrated when not all of them are fulfilled in SPIRIT current prototype.

SPIRIT is rather flexible and easy to use, which is very positive because it was a challenge to provide a simple interface above complex concepts.

Screen information is readable, learning to use the system seems rather easy.

The system is very slow.

Last users were often puzzled to assess the reliability of SPIRIT. They explain this by the limits of the documents collection as compared to the Web.

Besides its low speed and the limits of the documents collection it is working on, users pointed out a negative point of SPIRIT to be its ergonomics : the interface is clear but not ergonomic enough. For instance, if a user wants to replay a query changing only one word in it or the spatial relationship, he has to type the entire query again. Another example of lack of ergonomics is that results are initially displayed on the country map, whose appropriateness varies according to query granularity.

The most positive aspect about SPIRIT pointed out by users is the interface design : they like the aestheticism and the simplicity of it. The introduction of spatial relationships in query expression and the link between results and the map was quoted next.

As compared to other GIR systems, SPIRIT main differences were perceived to be :

- The introduction of spatial relationships in the query.
- The search of Web documents without thematic restriction.
- SPIRIT low speed.

**4.3. Results for system-centred evaluation**

To compare the results obtained with various settings, SPIRIT was run on the set of topics of the Test Collection with several settings. We wanted to test various methods for the two components that contribute to precision and recall : the index and the ranking.

Comparing results obtained with only text indexing in combination with BM25 ranking with results obtained with spatial settings show that spatial settings permit the retrieval of documents that would not be retrieved with pure textual methods, such as for the query "museum near Riverside", where Riverside is a placename. They also show that spatial settings often improve upon classical techniques for the topics requiring spatial awareness, e.g. when the query uses a non-containment relationship. Last, they globally improve "precision at 10", i.e. the number of relevant documents in the first ten documents returned.

Query	Textual (p10)	Angle distributed (p10)	Distance distributed (p10)	Textual (Recall)	Angle distributed (Recall)	Distance distributed (Recall)
<i>Hotels near Horgen</i>	0	0.2	0.1	1/27	6/27	6/27
<i>Hotels in Caerdydd</i>	0	0.3	0.3	0/30	6/30	6/30
<i>Museums near Riverside, Cardiff</i>	0	0.4	0.4	0/48	5/48	7/48
<i>Caravanning outside of Cardiff</i>	0	0.2	0.2	1/3	2/3	2/3
<i>Castles east of Edinburgh</i>	0	0.2	0	1/54	7/54	0/54

*Figure 6 Example precision at 10 and recall results for spatial queries*

Figure 6 illustrates precision at 10 and recall results for the pure textual indexing ("Textual") method in comparison with results for spatio-textual indexing using angle distributed and distance distributed ranking respectively. Four of the selected queries are those that involve a spatial relationship other than "in" (for which pure text indexing methods can be expected to work reasonably well) and one is a query that uses the Welsh name for Cardiff (i.e. Caerdydd) in combination with the "in" spatial relationship. In each case the pure textual indexing query failed to find relevant documents in the top 10 while relevant results were found using the spatio-textual indexing methods.

**5. Conclusion**

User evaluation gave consistent results with SPIRIT priorities on innovative features. Yet, users explain a feeling of frustration which highlights that their requirements are beyond SPIRIT achievements and that there is still more work to be done in this area.

Measuring precision and recall has demonstrated the value of spatial indexing with with results are usually improved by the precision at ten measure. Thus spatial indexing is important to back up the textual index for types of query that involve spatial operators other than "in" and when searching with rarely used alternative names for a given place. This is reflected in specific types of topics, as was foreseen by the methodology. This has two consequences. The

first one is that an important clue for building a standard test collection for spatially aware search engines is the topic definition scheme. The second one is that the effort when building the spatial index should be put on the geographical domain where spatial awareness is needed.

Besides evaluating the final prototype, and evaluating the various methods to support spatially aware information retrieval in SPIRIT, this task has yielded several results.

One result is setting the path for building a standard test collection to evaluate spatially aware search engines in the future. Further work should in particular focus on the relevance judgement scheme. During this evaluation, we learned that IR methods to evaluate system performance do not apply as such very well on a spatially aware search engine. Measures need to be added for instance to test not only ranking but also clustering after spatial criteria.

Another result is the experimentation of two different protocols for organising user tests to evaluate a geographical IR system. We may conclude that each protocol has its pros and cons. The protocol adopted for the interim prototype was more adapted to having users explore the user interface but did not respect the subjective and cognitive aspects of document relevance. The protocol adopted for the final prototype did not hinder user sincerity and spontaneity. Yet, it was not rigid enough to have them explore the full functions of the user interface during the test. To merge the pros of both protocols we see two methods. One is to organise both types of user tests. Another is to take care, when building of the document collection, to ensure that there is some consistency between the selection of a domain of topics that is appropriate to the system functionality and the recruitment of users. This would probably put many constraints on the system but allow to write scenarios the objectives of which meet the users concerns and which actually test the implemented methods.

## 6. References

- [Aumaitre et al. 04a] Guillaume Aumaitre, Sandrine Balley, Stephen Levin, Evaluation Methodology, SPIRIT deliverable D19 7201, July 2004
- [Aumaitre et al. 04b] Guillaume Aumaitre, Sandrine Balley, Bénédicte Bucher, Stephen Levin, Evaluation of the Interim Prototype, SPIRIT deliverable D20 7202, July 2004
- [Borlund 03] Pia Borlund, The IIR evaluation model: a framework for evaluation of interactive information retrieval systems, *Information Research*, Vol.8, No.3, april 2003
- [Joho and Sanderson 04] Hideo Joho, Mark Sanderson, The SPIRIT Collection: an overview of a large web collection. *SIGIR Forum*, 38(2), 2004
- [Purves and Yang 05] Ross Purves, Bisheng Yang, Graphical query and presentation interface, SPIRIT deliverable D26 4301, june 2005
- [Sanderson and Joho 03] Mark Sanderson, Hideo Joho, Test collection formation methods, SPIRIT deliverable D11 2102, december 2003

## 7. Appendix A : mail sent to users for a first contact

Content of the mail sent to users :

Hello,

If you receive this mail, it means you have been once approached by a member of the SPIRIT team or that your name has been proposed by a member, for instance by Ross Purves. We hope you will take some time to read this message and try out the interface of our prototype. This would be very valuable to us.

SPIRIT is a European research project. It has developed a prototype search engine with a spatial awareness. SPIRIT prototype is for instance able to retrieve documents relevant to a query even if the words used in the query to express the spatial aspects are not to be found in the document.

If you agree, we would be very grateful if you take some time next week -we estimate it around 1 hour- to try out our on-line prototype and fill in a questionnaire after that. You will need an Internet Explorer and a java virtual machine.

Then, anytime you want between Monday June 6th and Monday June 13th, use Internet Explorer to log on spirit prototype :  
<http://cheese.geo.unizh.ch:28080/tmp/visinterface/index.jsp>  
(the url may be working before that time but it will be in a draft state until Monday morning) On this prototype, please try out some searches. During this, you may find useful the on-line help document. It accounts for some "puzzling behaviours" of SPIRIT due to documents and data limits. It also gives some indications about how to use the interface. Feel free to test any query. Yet, given data that were bought for the prototype, you are likely to get better results on areas around Cardiff and Edinburgh.

When you have finished, please fill in the questionnaire that can be downloaded from the site and send it by mail to  
<mailto:benedicte.bucher@ign.fr>

If you think you will need some "hot line" support, please write a mail to the same person (<mailto:benedicte.bucher@ign.fr>) with your phone number, location and period of time you will be using the prototype so that I can get someone from the SPIRIT team available on-line and possibly calling you when you are testing the prototype.

We hope you will agree to test this prototype and find this experience interesting. We will keep you informed of evaluation results in general.

Best regards.

--

Bénédicte Bucher  
Institut Géographique National  
Laboratoire COGIT  
2 av Pasteur  
94 165 St Mandé Cedex - France  
+ 33 (0)1 43 98 80 03

## **8. Appendix C : questionnaire**

Nota bene : This is the questionnaire that users filled. It has errors in the section "give us marks". The scales are missing the number 2, which should bias the results. This has been handled during the results interpretation.

## **Questionnaire on the SPIRIT final prototype**

You may decide not to answer a question and still answer the rest of the questionnaire.

Thank you for your help.

### **Who are you?**

**Name :**

e.g. Madonna

**Job. Background :**

e.g. Singing, Acting

**Age :**

e.g. 45

**Nationality, Gender:**

eg American woman

**How well do you speak English ?**

(please choose only one)

Native - Fluent - Basic

**How often do you use Internet search engines (e.g. Google, AltaVista, Yahoo)?**

(please choose only one)

Never - Occasionally - Once a week - Several times a week - Daily

**How often do you use commercial online search engines (e.g. Dialog, Lexis-Nexis)?**

(please choose only one)

Never - Occasionally - Once a week - Several times a week - Daily

**How often do you perform searches on computerized library catalogues (e.g. your library)?**

(please choose only one)

Never - Occasionally - Once a week - Several times a week - Daily

**How often do you perform searches on mapping websites (e.g. multimap.co.uk, streetmap.co.uk)?**

(please choose only one)

Never - Occasionally - Once a week - Several times a week - Daily

\* \* \* \* \*

### **Express yourself**

Please use this section to say what you think about SPIRIT and to tell us what you did in your own words.

**What queries did you try?**

**Were the *first* documents retrieved relevant to your query? If not, could you find relevant documents?**

**Did anything about SPIRIT puzzle you, or make it hard to complete your search?**

**What do you like about SPIRIT ?**

**What do you dislike about SPiRiT ?**

\* \* \* \* \*

## Give us marks

For each of these pairs of properties, where would you locate SPiRiT, on a scale from 1 to 5 :

**A: Overall reaction to the system**

Terrible 1 3 4 5 6 Wonderful

Frustrating 1 3 4 5 6 Satisfying

Dull 1 3 4 5 6 Stimulating

Difficult 1 3 4 5 6 Easy

Rigid 1 3 4 5 6 Flexible

**B : Specific items**

Characters on the computer screen are :  
hard to read 1 3 4 5 6 easy to read

The organisation of information on screen is :  
Confusing 1 3 4 5 6 very clear

Sequence of screens is :  
Confusing 1 3 4 5 6 very clear

Use of terms throughout system is :  
Inconsistent 1 3 4 5 6 consistent

Useful messages told you what was happening :  
Never 1 3 4 5 6 always

Error messages were :  
Unhelpful 1 3 4 5 6 helpful

Help messages were :  
Unhelpful 1 3 4 5 6 helpful

Learning to operate the system was :  
Difficult 1 3 4 5 6 easy

Exploring new features by trial and error was

Difficult 1 3 4 5 6 easy

Tasks can be performed in a straight-forward manner :

Never 1 3 4 5 6 always

Supplemental reference materials is :

Confusing 1 3 4 5 6 clear

System speed is :

too slow 1 3 4 5 6 fast enough

System is :

Unreliable 1 3 4 5 6 reliable

**Do you want to comment on why you gave some marks?**

\* \* \* \* \*

## Comparisons

**Please, if you have some more time, try out one of your queries on some of the following Web sites :**

<http://local.google.co.uk>

As compared to SPiRiT,..  
Were the retrieved documents more relevant?

Was writing the query / refining of the query easier ?

Was results presentation better (mapping, ranking)?

<http://www.yell.co.uk>

As compared to SPiRiT,..  
Were the retrieved documents more relevant?

Was writing the query / refining of the query easier ?

Was results presentation better (mapping, ranking)?

<http://www.map24.com>

As compared to SPiRiT,..  
Were the retrieved documents more relevant?

Was writing the query / refining of the query easier ?

Was results presentation better (mapping, ranking)?

**Please, mail this form to :**

[benedicte.bucher@ign.fr](mailto:benedicte.bucher@ign.fr)

or post it to :

**Bénédicte Bucher  
IGN - COGIT  
2 avenue Pasteur  
94165 Saint Mandé Cedex  
FRANCE**

\*\*\*\*\*

**Thank you very much for your contribution.  
We will communicate to you a summary of the results of this  
evaluation.**

\*\*\*\*\*

**9. Appendix D : Documents returned by users**

**9.1. User 1.**

Who are you?

*Name : **Hubert Frédéric***

*Job. Background : **Postdoc***

*Age : **29***

*Nationality, Gender: **French man***

*How well do you speak English ? **Basic***

*How often do you use Internet search engines (e.g. Google, AltaVista, Yahoo)? **Daily***

*How often do you use commercial online search engines (e.g. Dialog, Lexis-Nexis)? **Never***

*How often do you perform searches on computerized library catalogues (e.g. your library)?*

***Occasionally***

*How often do you perform searches on mapping websites (e.g. multimap.co.uk, streetmap.co.uk)?*

***Once a week***

*\* \* \* \* \**

Express yourself

**What queries did you try?**

Church west of Cardiff, Church outside of Cardiff, Restaurant near Edinburgh, Restaurant near Cardiff, Restaurant in France, Restaurant in Paris France, Restaurant outside of Paris, School outside of Paris, Castles east of Edinburgh

**Were the *first* documents retrieved relevant to your query? If not, could you find relevant documents?**

Lorsque des résultats étaient proposés, les documents correspondaient à ce que je cherchais. Par contre, je ne sais pas, s'il s'agissait des meilleurs documents disponibles en premier, puisque je ne connais pas tous les documents disponibles... donc c'est difficile à évaluer. Dans certaines des requêtes, aucun résultat n'était disponible, ou les résultats retournés étaient étranges comme pour "Restaurant near Cardiff", où le premier résultat pointe sur un restaurant en Allemagne (Hamburg).

**Did anything about SPIRIT puzzle you, or make it hard to complete your search?**

Les limitations des données disponibles (que je ne connaissais pas) ne permettent pas de juger pleinement de la pertinence de la recherche.

Il n'est pas évident de comprendre à quoi correspondent les résultats fournis. On a un lien avec un titre, non nécessairement explicite et c'est tout. En tant qu'utilisateur, une vision même sommaire du contenu du document serait appréciée.

**What do you like about SPiRiT ?**

La formulation et l'exécution d'une requête sont simples. L'interface contenant la page et les liens est également simple d'utilisation. Cela ne demande un effort énorme pour comprendre avec la documentation à quoi correspondent tous les éléments et les fonctionnalités de l'interface. C'est appréciable.

L'association de la carte avec des documents multimédias tel que des pages Web avec des images est une excellente idée.

L'exploitation de termes relatant des relations topologiques dans un système de formulation est novateur (selon moi) au niveau des interfaces de requêtes sur le Web en dehors des systèmes basés sur la langue naturelle.

**What do you dislike about SPiRiT ?**

Je trouve que le pouvoir d'attraction du prototype est faible. Connaissant un peu le projet, je trouve évidemment que cela est intéressant, mais à qui s'adresse le système ? Quelqu'un qui n'y connaît pas grand-chose, risque de ne pas être satisfait.

\* \* \* \* \*

## Give us marks

For each of these pairs of properties, where would you locate SPiRiT, on a scale from 1 to 5 :

**A: Overall reaction to the system**

Terrible 1 3 **4** 5 6 Wonderful

Frustrating 1 **3** 4 5 6 Satisfying

Dull 1 3 4 **5** 6 Stimulating

Difficult 1 3 **4** 5 6 Easy

Rigid 1 3 **4** 5 6 Flexible

**B : Specific items**

Characters on the computer screen are :  
hard to read 1 3 4 5 **6** easy to read

The organisation of information on screen is :  
Confusing 1 3 **4** 5 6 very clear

Sequence of screens is :  
Confusing 1 3 **4** 5 6 very clear

Use of terms throughout system is :  
Inconsistent                      consistent

1 3 4 **5** 6

Useful messages told you what was happening :  
Never                                      always

1 3 **4** 5 6

Error messages were :  
Unhelpful                              helpful

1 **3** 4 5 6

Help messages were :  
Unhelpful                              helpful

1 3 4 **5** 6

Learning to operate the system was :  
Difficult                                      easy

1 3 4 5 **6**

Exploring new features by trial and error was  
Difficult                                      easy

1 3 4 **5** 6

Tasks can be performed in a straight-forward manner :  
Never                                      always

1 3 **4** 5 6

Supplemental reference materials is :  
Confusing                                      clear

1 3 4 **5** 6

System speed is :  
too slow                                      fast enough

1 **3** 4 5 6

System is :  
Unreliable                                      reliable

1 **3** 4 5 6

### **Do you want to comment on why you gave some marks?**

#### **Pour la section A:**

Globalement, je suis un peu déçu par ce que j'ai vu et testé. Je suis frustré car je trouve que l'idée générale du projet est bonne et que je ne crois pas spécialement que le prototype mette suffisamment en avant les idées sous-jacentes au projet. Je suis également frustrée, car je m'attendais à voir plus de choses sur l'interface utilisateur! Une plus grande intuitivité et une facilité pour la recherche de l'information désirée.

#### **Pour la section B:**

Pour l'organisation de l'information sur l'écran, je trouve que c'est clair, mais qu'il y a encore des ajustements à faire au niveau de la mise en forme de la zone de formulation de requêtes. Il reste trop d'espace inutilisé... on pourrait agrandir la carte par exemple, ajouter d'autres fonctionnalités. Et puis, le Next 10 (Previous 10) est mal positionné. Il devrait être en dessous des 10 résultats proposés.

La séquence des pages Web est gênante par moment. Notamment lorsqu'on clique sur le bouton "return" de la page informant de l'impossibilité de trouver une requête, la page retournée n'est pas celle d'où on vient et en plus, on perd les informations sur la requête effectuée.

Pour les messages d'erreurs, j'ai obtenu un message d'erreur provenant d'Apache suite à une de mes requêtes. J'aurais préféré obtenir un autre type de message (requête: restaurant in France).

Il manque des messages d'utilisation de l'interface de formulation de la requête, malgré qu'elle soit simple.

La visualisation des résultats sur la carte n'est pas optimale, dans le sens où on voit en premier lieu le pays en entier. On demande une ville ou une région et on devrait avoir une représentation cartographique se focalisant sur cette ville ou région et ainsi une meilleure vision des résultats.

Le temps de réponse de la première requête est d'environ 45 secondes. Je trouve cela plutôt long. Et ensuite, en utilisant les zooms sur la carte, cela prend encore un certain temps.

Concernant la fiabilité, j'ai obtenu quelques messages d'erreurs Apache suite à des requêtes su type "restaurant in France" et j'ai eu droit à un plantage général de Internet Explorer suite à l'ouverture et à la fermeture d'un document à partir de l'interface. Ceci m'a conduit à redémarrer Internet Explorer.

\* \* \* \* \*

## Comparisons

**Please, if you have some more time, try out one of your queries on some of the following Web sites :**

<http://local.google.co.uk>

As compared to SPIRIT,..

Were the retrieved documents more relevant?

Google trouve plus rapidement des informations (pages Web) avec davantage d'information sur le contenu avec des numéros de téléphone ou des adresses de localisations. La pertinence n'est pas nécessairement présente, car elle n'est pas évidente à évaluer.

Was writing the query / refining of the query easier ?

Oui, la requête s'écrit aussi facilement que sur SPIRIT. Par contre, il n'est pas possible de spécifier des relations topologiques (near, west of). Un affichage sur une ligne pourrait faciliter la lecture de la requête côté SPIRIT.

Was results presentation better (mapping, ranking)?

Pour l'ordre des documents affichés, il est difficile d'évaluer la pertinence. Il est cependant évident que les résultats sont affichés par proximité par rapport au centre de la localisation. Pour leur présentation, elle est mieux que sur SPIRIT, même si on ne voit pas toutes les informations sur la page d'un seul coup d'œil. On dispose de plus d'information sans avoir à atteindre la page Web liée au résultat.

Concernant la carte, la qualité de la carte est meilleure à petite échelle sur Google. Et la navigation se fait mieux pour visualiser et atteindre les résultats.

<http://www.yell.co.uk>

As compared to SPIRIT,..

Were the retrieved documents more relevant?

Oui, le site permet de retrouver l'information, mais il est difficile de savoir quel est la pertinence de proposition des pages Web.

Was writing the query / refining of the query easier ?

L'écriture de la requête est simple et fonctionne comme pour SPiRiT, mais sans les relations topologiques.

Was results presentation better (mapping, ranking)?

Le problème avec Yell est qu'il n'y a pas de cartes localisant directement les résultats. Pour l'ordre de présentation des pages, c'est pareil ! Il y a beaucoup d'informations pertinentes, mais on ne sait pas comment les sites sont classés.

<http://www.map24.com>

Were the retrieved documents more relevant?

Non, la fiabilité du système de recherche laisse à désirer. J'ai mis plus de temps pour réussir à faire fonctionner le système et trouver comment effectuer mes requêtes. Il me trouve bien des résultats, mais il m'est difficile d'évaluer la pertinence de ce qui est proposé.

Was writing the query / refining of the query easier ?

Non, ce n'est pas aussi évident qu'avec SPiRiT au premier abord. J'ai eu des problèmes pour effectuer mes requêtes, à cause de mon appartenance au sol Canadien et que le map24 existe en Canadien. J'ai dû faire des ajustements et j'ai des problèmes pour effectuer les requêtes.

Was results presentation better (mapping, ranking)?

Au niveau de la carte, il n'y a pas grand-chose à redire. C'est vraiment très bien avec toutes les informations désirées et leur localisation exacte en passant sur les informations référées. Concernant la présentation des résultats et leur classement, il n'est pas évident de juger. Il manque des informations pour comprendre ce qu'est exactement le résultat.

**Commentaire global: Il n'est vraiment pas évident de juger de la pertinence du moteur de recherche SPiRiT comparativement aux sites précédents, que les sites Web et documents indexés sont considérables pour ces derniers.**

## **9.2. User 2**

### Who are you?

**Name :**

*Lars Harrie*

**Job. Background :**

*Teacher, researcher in GIS*

**Age :**

36

**Nationality, Gender:**

Sweden, man

**How well do you speak English ?**

Fluent

**How often do you use Internet search engines (e.g. Google, AltaVista, Yahoo)?**

Daily

**How often do you use commercial online search engines (e.g. Dialog, Lexis-Nexis)?**

Occasionally

**How often do you perform searches on computerized library catalogues (e.g. your library)?**

Occasionally

**SPIRIT project**

*Evaluation of SPIRIT prototype following integration and testing*

IST-2001-35047

D31 7301

**How often do you perform searches on mapping websites (e.g. [multimap.co.uk](http://multimap.co.uk), [streetmap.co.uk](http://streetmap.co.uk))?**  
Occasionally

\* \* \* \* \*

## Express yourself

Please use this section to say what you think about SPiRiT and to tell us what you did in your own words.

### **What queries did you try?**

- 1) Firth of Forth north of Edinburgh
- 2) The Royal Observatory in Edinburgh
- 3) Castles west of Edinburgh

### **Were the *first* documents retrieved relevant to your query? If not, could you find relevant documents?**

- 1) No document found.
- 2) Yes, very relevant.
- 3) No. But there were other relevant documents (especially 5 and 6)

### **Did anything about SPiRiT puzzle you, or make it hard to complete your search?**

No, the interface was quite clear.  
A question is of course how the spatial operators (like "West of") are defined.

### **What do you like about SPiRiT ?**

The design of the interface.

### **What do you dislike about SPiRiT ?**

That you have to rewrite the whole query if the server does not find any document. So if you have used an erroneous spatial operator you have to rewrite all your text ...

The supplementary material (help) was not detailed enough.

Would like to be able to zoom out.

Lacks printing facilities.

\* \* \* \* \*

## Give us marks

For each of these pairs of properties, where would you locate SPiRiT, on a scale from 1 to 5 :

### **A: Overall reaction to the system**

Terrible                      Wonderful  
4

Frustrating                      Satisfying



Unreliable                      reliable  
3

**Do you want to comment on why you gave some marks?**

**The answers are based on just a few questions and might not be representative.**

\* \* \* \* \*

## Comparisons

**Please, if you have some more time, try out one of your queries on some of the following Web sites :**

<http://local.google.co.uk>

As compared to SPIRIT,..  
Were the retrieved documents more relevant?  
**No.**

Was writing the query / refining of the query easier ?  
**Yes. But more limited possibilities.**

Was results presentation better (mapping, ranking)?  
**No**

**+ Faster**

<http://www.yell.co.uk> + <http://www.map24.com>

As compared to SPIRIT,..  
Were the retrieved documents more relevant?  
**No**

Was writing the query / refining of the query easier ?  
**Yes. But more limited possibilities.**

Was results presentation better (mapping, ranking)?  
**No**

**+ Faster**

**- Did only include "commercial" site**

### **9.3. User 3**

## Who are you?

**Name : Nicolas Regnauld**

**Job. Background : Research computer science - GIS**

**Age : 35**

**Nationality, Gender: French bloke**

**How well do you speak English ? fluent franenglish**

**How often do you use Internet search engines (e.g. Google, AltaVista, Yahoo)?** Several times a week

**How often do you use commercial online search engines (e.g. Dialog, Lexis-Nexis)?** never

**How often do you perform searches on computerized library catalogues (e.g. your library)?** occasionally

*How often do you perform searches on mapping websites (e.g. multimap.co.uk, streetmap.co.uk)?*

*Once a week*

\* \* \* \* \*

## Express yourself

### **What queries did you try?**

Titanic near Southampton

Sex shop in Edinburgh

Pubs east of Edinburgh

Tourism near Edinburgh => NO DOCUMENTS FOUND!!

Horses east of Edinburgh

### **Were the *first* documents retrieved relevant to your query? If not, could you find relevant documents?**

For the titanic, only irrelevant documents.

For the sex shop, the 4<sup>th</sup> and 5<sup>th</sup> gave me some information about my former "local", which I was quite please about. The rest was mostly irrelevant. The 4<sup>th</sup> and 5<sup>th</sup> were actually the same links, with same titles.

For the pubs, 2 responses. The first one is a list of pubs in Scotland, the second is a specific pub in the centre of England!!! Et there are both located on the map east of Edinburgh, not sure why...

For the horses, I had mostly relevant links (only 6 though).

### **Did anything about SPiRiT puzzle you, or make it hard to complete your search?**

Its very slow.

### **What do you like about SPiRiT ?**

I like the spatial options, but I am not convinced that it works well as a filter for the results...

### **What do you dislike about SPiRiT ?**

I haven't really understood how it works... It seems to find documents which have very little to do with the keywords.

I don't understand how the documents relate to their location on the map...

\* \* \* \* \*

## Give us marks

For each of these pairs of properties, where would you locate SPiRiT, on a scale from 1 to 5 :

### **A: Overall reaction to the system**

Terrible 2                      Wonderful

Frustrating 2                      Satisfying

Dull 4                              Stimulating

Difficult 5                        Easy

Rigid 5                             Flexible

### **B : Specific items**

Characters on the computer screen are :

hard to read                      easy to read  
6

The organisation of information on screen is :

Confusing                      very clear  
6

Sequence of screens is :

Confusing                      very clear  
5

Use of terms throughout system is :

Inconsistent                      consistent  
3

Useful messages told you what was happening :

Never                      always  
5

Error messages were :

Unhelpful                      helpful  
5

Help messages were :

Unhelpful                      helpful  
5

Learning to operate the system was :

Difficult                      easy  
6

Exploring new features by trial and error was

Difficult                      easy  
5

Tasks can be performed in a straight-forward manner :

Never                      always  
5

Supplemental reference materials is :

Confusing                      clear  
4

System speed is :

too slow                      fast enough  
1

System is :

Unreliable                      reliable  
4

\* \* \* \* \*

## Comparisons

<http://local.google.co.uk>

As compared to SPIRIT,..

Were the retrieved documents more relevant?

Yes

Was writing the query / refining of the query easier ?  
No, they don't have the "near" or "east" etc.

Was results presentation better (mapping, ranking)?  
Yes, especially the mapping

<http://www.yell.co.uk>

As compared to SPiRiT,..

Were the retrieved documents more relevant?

Yes and no. Relevant but not very informative, apart from contact

Was writing the query / refining of the query easier ?  
Lack flexibility on the spatial side (near, east, etc.)

Was results presentation better (mapping, ranking)?  
Better map, but not nice general presentation

<http://www.map24.com>

Were the retrieved documents more relevant?

Seems to be only for addresses, can't look for something

In a city it asks for a postcode!

Was writing the query / refining of the query easier ?

Was results presentation better (mapping, ranking)?

**9.4. User 4**

**Who are you?**

**Name** : Andrew Faulk

**Job. Background** : Consultant, specialising in environment /economic development issues

**Age** : 39

**Nationality, Gender**: British man

**How well do you speak English ?** Native

**How often do you use Internet search engines (e.g. Google, AltaVista, Yahoo)?** Several times a week

**How often do you use commercial online search engines (e.g. Dialog, Lexis-Nexis)?** Never –

**How often do you perform searches on computerized library catalogues (e.g. your library)?**

Occasionally

**How often do you perform searches on mapping websites (e.g. multimap.co.uk, streetmap.co.uk)?**

Several times a week

\* \* \* \* \*

**Express yourself**

**What queries did you try?**

Transport, Edinburgh, Scotland

European Funding, Newcastle-upon-tyne, England (SPiRiT couldn't find Newcastle)

European Funding, Cardiff, Wales

Local Agenda 21, Edinburgh, Scotland (no documents found)

**Were the first documents retrieved relevant to your query? If not, could you find relevant documents?**

Transport question had some results, but seemed a bit random.

EU funding in Wales better, but included an article on Sweden(!!) – SPiRiT seemed to find lots of academic stuff, but not so much from government compared to google (see below)

**Did anything about SPiRiT puzzle you, or make it hard to complete your search?**

The idea is good. But results much less so. Also, by comparison with other enquiries (I tried the EU funds in Wales one in google) the results were a lot more focused...

**What do you like about SPiRiT ?**

As above – the idea is good!

**What do you dislike about SPiRiT ?**

Not so much the system; the information it searches is clearly limited just now.

\* \* \* \* \*

## Give us marks

For each of these pairs of properties, where would you locate SPiRiT, on a scale from 1 to 5 :

**A: Overall reaction to the system**

Terrible  
3                                  Wonderful

Frustrating  
2                                  Satisfying

Dull  
**5 – it was certainly interesting!**                                  Stimulating

Difficult  
5                                  Easy

Rigid  
5                                  Flexible

**B : Specific items**

Characters on the computer screen are :  
hard to read                                  easy to read  
5

The organisation of information on screen is :  
Confusing                                  very clear  
5

Sequence of screens is :  
Confusing                                  very clear  
6

Use of terms throughout system is :  
Inconsistent                                  consistent  
5

Useful messages told you what was happening :  
Never                                  always  
4

Error messages were :  
Unhelpful                                  helpful

## **SPIRIT project**

*Evaluation of SPIRIT prototype following integration and testing*

IST-2001-35047

D31 7301

Error messages were frustrating, more than anything else, and seemed due to lack of source material for the system to research

Help messages were :

Unhelpful                      helpful  
N/a

Learning to operate the system was :

Difficult                      easy

6 – very easy to start. Like any other search engine, it would take a while to find out how to work it.

Exploring new features by trial and error was

Difficult                      easy

4

Tasks can be performed in a straight-forward manner :

Never                      always

5

Supplemental reference materials is :

Confusing                      clear

?

System speed is :

too slow                      fast enough

limited by my computer & connection, I suspect

System is :

Unreliable                      reliable

As above.

**Do you want to comment on why you gave some marks?**

Already done that

;-)

\* \* \* \* \*

## **Comparisons**

<http://local.google.co.uk>

**As above – answers more relevant (in the sense of finding information I know is available and useful), query process pretty much the same.**