Project no. 033104

MultiMatch

Technology-enhanced Learning and Access to Cultural Heritage
Instrument: Specific Targeted Research Project
FP6-2005-IST-5

D1.4 Functional Specification of the Second Prototype

Start Date of Project: 01 May 2006
Duration: 30 Months

Organisation Name of Lead Contractor for this Deliverable: UNED

Version: Final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Table of Contents

Executive Summary

This document focuses on the definition of the Functional Specifications for the second prototype of the MultiMatch system. This second prototype is scheduled to be integrated and tested at month 27.
In this document, we describe what the system is expected to do, rather than how, excluding the following topics from the specification of MultiMatch functionalities:

- User requirements, which have been input for this deliverable.
- Software engineering aspects such as the architecture of the system (discussed in WP3)
- Techniques, algorithms and data structures (WP4, WP5 and WP6)
- User interface design issues (WP6)
- Evaluation issues (WP7)

The document is based on five main sources:

1. The **MultiMatch project description** [4].
2. The **revision of the MultiMatch *common vision***, which is the result of an internal project discussion that solves some ambiguities and issues which were left open in the technical description of the project, in order to reach a common consensus within the consortium members on the goals to achieve.
3. The **user requirements of the project** [2], which are the result of an extensive user survey among experts from three communities related to cultural material: experts from cultural heritage (CH) organisations, from educational institutions and from the tourism sector.
4. The **Functional Specification of the First Prototype** [1].
5. The **Review Report** for the first project period provided by the European Commission, which provides MultiMatch Consortium with a useful set of guidelines and recommendations to take into account for the second prototype.

The MultiMatch functional specifications attempt to be a reasonable intersection of these five sets of input. Using this information (i) we have identified the main functionalities expected for the final MultiMatch search engine, and (ii) we have specified which functionalities should be improved or discarded based on the lessons learned from the first prototype. Due to a conflict in [4] between the timing for evaluation of Prototype 1 (months 15 to 18) and the drafting of the Functional Specifications for Prototype 2 (months 13 to 15) we have not been able to use the results of the evaluation when formulating the second prototype functional specifications. However, we have taken into consideration all the feedback received from the extensive discussions had with various members of the user communities during the development of the first prototype.

During this second period the MultiMatch Consortium will explore novel approaches and techniques to enhance and improve cultural heritage information indexing, retrieval, accessing and visualization. These can be considered as research challenges (some of them still under discussion) and, as a consequence, there is not yet a commitment about which of them will be finally integrated. The aim of the Consortium is first to test the viability of such approaches (i.e. developing small applications and/or laboratory experiments) and then decide about their integration on the final system.
The structure of the document is the following. First we review the achievements on the first prototype. The functional specifications tables for the first prototype are revisited checking all the items which have been finally integrated or moved to the second prototype. Secondly we review the MultiMatch Complete Picture introduced in [1]. This common vision has been re-discussed and adapted based on the lessons learned from the first prototype and the new research goals for this second period. Finally, we describe the specific functionalities for the second prototype. This document has also two annexes where functional specifications for the second prototype and the achievements of the first prototype are listed in a tabular format.

1 Achievements of the First Prototype

The MultiMatch first prototype has faced the basic problems and research issues related with the development of a domain specific search engine. For this reason, within this first period, efforts have been focused on providing the system with a robust and reusable baseline which offers accurate indexing and retrieval mechanisms rather than on providing the system with novel research approaches.

The main achievements of the project so far can be checked in Appendix II, where functional specifications for the first prototype have been summarized in tabular format. These tables show the functionality implemented in this first period and that moved to the second prototype or discarded.

2 MultiMatch Second Prototype. The Complete Picture

This section shows the revised common vision of MultiMatch initially introduced in [1]. Some points have been kept while others have been modified and revisited to best suit the new MultiMatch needs identified for this second period.

MultiMatch is intended to be a Web search engine in the cultural domain. To specify its functionality it is necessary to address the following issues:

- **What is MultiMatch going to index?** Which are the types of web sources that will be accessed via the MultiMatch search engine? How large and representative will the volume of data indexed within the scope of the project be?
- **What are the retrieval functionalities of MultiMatch?** This is the core of the MultiMatch functionality. In what ways will the system provide access to the indexed data?
- **What types of user/programming interfaces will be made available on MultiMatch search services?** We will list here the interfaces that will be provided by the system (a user interface and an application program interface) and the set of intended facilities they should provide.

The second part of this deliverable describes the specific set of functional specifications planned for the second prototype and also presents them as a check-list with priorities ready to be used and refined in the development of Deliverable 3.3 (Task 3.4-Detailed Specification of Second Prototype) and the implementation of the second prototype.

2.1 What will MultiMatch Index?

The **main source** of information stored in the final system will be publicly accessible web material related to cultural heritage. Specifically, MultiMatch will focus on crawling and indexing material under URLs from:

- Cultural heritage sites (such as museums or cultural institutions).
- Educational sites related to cultural heritage (such as universities).
- Tourism information sites (The indexing of these sites into MultiMatch will be optional).
- Encyclopaedic sources (such as *Wikipedia*[1]).

[1] http://www.wikipedia.org

- IPR Protected cultural heritage materials owned by specific organizations such as Alinari, Sound and Vision and Biblioteca Virtual Miguel de Cervantes.
- OAI[2] compliant resources.
- External sources accessible via proprietary APIs such as Google, Yahoo, Blinkx, etc.
- RSS feeds from authoritative cultural web sites (such as museums or cultural sections of main European newspapers). These feeds will include podcasts and, when possible, vodcast.

However, we will apply some restrictions:

1. The crawl will focus on English, Spanish, Dutch and Italian sources, plus German and Polish in the second prototype.
2. Any web page with cultural heritage contents will be suitable to be indexed by MultiMatch although we will focus primarily on material created or supported with public funds, which should consist of high quality contents.
3. The system will attempt to identify and index images, videos and audio sources with a cultural value from crawled web content, such as portraits, photographs of artists and works of art, educational videos, etc. Logos, navigational icons, etc. will automatically be discarded from document indexing as much as possible.
4. Crawling will be sufficiently extensive to test the validity of the MultiMatch approach in real-life searching scenarios.

2.2 Which Retrieval Functionalities Are Expected?

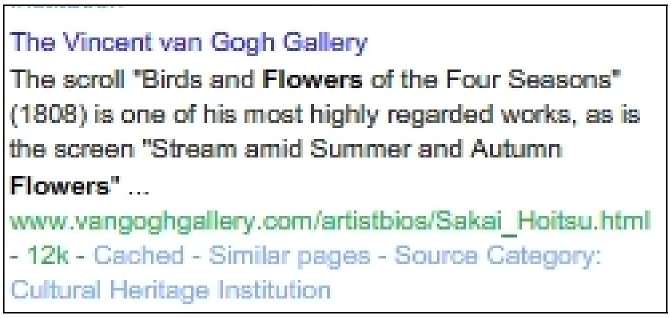
MultiMatch is expected to retrieve what can be called **cultural objects**. A cultural object can be defined as an information unit which refers to any item of society's collective memory including print (books, journals, newspapers), photographs, museum objects, archival documents, audiovisual material. This piece of information can be built and displayed in different ways which will be studied and discussed by the Consortium during this second period.

For instance, it could be possible to distinguish between **atomic cultural objects** and **compound digital objects.** An atomic cultural object could be defined as a single image, a web page, an audio file or a video file which can be located through an URL, while a compound digital object could be defined as a more complex piece of information composed by information combined from different sources. In this last case, compound digital objects should provide the user with summarized information about a cultural entity (such a painter, a sculptor, a place, a painting, etc.) and, when possible, some URLs to specific atomic cultural objects where this information could be expanded. The MultiMatch metadata schema designed for the first prototype points in this direction, making it possible to build this information based on information extraction techniques which will be explored during this second period.

Dealing with cultural objects in MultiMatch also implies that the way of defining how items are viewed could depend on the user profile, the type of search performed or the personal MultiMatch customization.

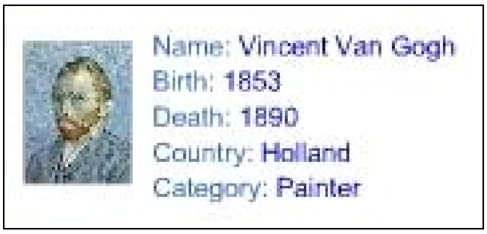
Figures 1 and 2 show two alternative views of cultural objects in MultiMatch. Figure 1 shows an atomic cultural object (which is rendered showing an snippet plus some extra information such as the source category) which links with a web page about *Van Gogh* works, while Figure 2 shows a compound cultural object which summarizes some information about *Van Gogh*.

[2] Open Archive Initiative: http://www.openarchives.org/ Example: MICHAEL (http://www.michael-culture.org/)



The Vincent van Gogh Gallery
The scroll "Birds and **Flowers** of the Four Seasons" (1808) is one of his most highly regarded works, as is the screen "Stream amid Summer and Autumn **Flowers**" ...
www.vangoghgallery.com/artistbios/Sakai_Hoitsu.html - 12k - Cached - Similar pages - Source Category: Cultural Heritage Institution

Figure 1. Atomic cultural object about a Van Gogh's resource created using title plus snippet plus extra information such as the page source (cultural heritage site)



Name: Vincent Van Gogh
Birth: 1853
Death: 1890
Country: Holland
Category: Painter

Figure 2. Compound cultural object which describes Van Gogh based on extracted metadata.

Users will be provided with a browsing capability allowing them to navigate the MultiMatch collection using, among others, a web directory-like structure based on the MultiMatch ontology.

All search facilities will be **translingual**, i.e. the user will formulate queries in a given language and retrieve results in all languages covered by the prototype. According to the user's language profile, results in unknown languages will be returned in a way that is interpretable by the user, e.g. with a document surrogate or associated keywords in the user's preferred language. Multilinguality has several implications from the point of view of the design of the user interface, which constitutes a further research objective in itself and is not discussed here (see WP6 documentation for details on multilingual features addressed for the second prototype). However, the list of specifications for the second prototype (at the end of this deliverable) includes basic specifications which are intended to cover the most general functionalities for the system interface that will be discussed in WP6.

Apart from **multilinguality** and **multimediality**, the most prominent feature of MultiMatch is that its search capabilities will go beyond simple resource matching and ranking, and will involve **information extraction** and **text mining** techniques. These techniques will allow MultiMatch to build and index the *cultural objects* described above. One of the main goals of this document is to determine what types of information needs (from the broad spectrum covered by users of CH information) will be covered by the MultiMatch prototype as a proof-of-concept on the use of such techniques for web search.

From the expert users survey [1] we can conclude that experts commonly tend to classify searches for information about creators (authors) and creations (works of art and masterpieces) as their most common search tasks. Therefore, in MultiMatch we have initially decided to focus two types of specialized searches on **creators** and **creations**, although specialized searches focused on other relevant categories can also be explored and discussed during this second period.

Search Interaction Levels in MultiMatch

In MultiMatch we can differentiate between two main **search interaction levels**:

- **A default/overview MultiMatch search level** where no assumption is made on the user's query and MultiMatch retrieves information from all indexed materials. Given a general query, MultiMatch will retrieve all the cultural objects which best suit the user's needs. Ranking and classification of these results will be also performed by the MultiMatch system. The default search level can be understood as a level where the user wants to have a general idea of what the MultiMatch system can offer for a general query but does not need to perform specialized operations such as video or audio browsing, image similarity search, etc. In this way, the default search level will provide the user with an initial set of atomic and compound cultural objects which could be deeply explored and refined using the specialised levels.
- **A set of specialized interaction levels**. These allow the user to query MultiMatch specific search services (for instance, video, audio or image search) and retrieve all the relevant information available via the selected search service. The use of these specialised levels give the user access to specialised features which can be used to drive specific media browsing or search results refinements that are not allowed in the default search level. The specialized interaction level can be understood as a level more focused on expert users which not only want a set of search results but also a set of tools/functionalities to explore and browse the retrieved information.

Both search levels will include two different search modes to drive users' queries:

- **A default mode** where MultiMatch will adjust and select the main search parameters for the user (i.e. source and target languages, query types, filters, etc.).
- **An advanced mode** where MultiMatch will allow the user to select and adjust all the search parameters and customize his profile. This advanced mode will also provide the user with a set of tools to perform metadata based searches which allow the user to query MultiMatch using any combination of the MultiMatch metadata schema fields (i.e. such as digital libraries query systems).

The first prototype has been focused on the development of the basic indexing and search functionalities (such as robust solutions for multilingual text, image, audio and video retrieval) but it has not faced the creation of compound cultural objects. In this second period MultiMatch will use these results as a baseline to address more complex issues, mainly related with the indexing and retrieval of compound cultural objects to improve search results and to provide the user with much more information rendered and displayed in sophisticated ways.

Default Interaction Level

As described above, the default interaction level will allow the user to query MultiMatch using its different search services at the same time. Search results will be combined and rendered to properly show the retrieved information to the user in an appropriate fashion.

The default search level must be understood as a way for the users to express their search needs when they want an overview of the MultiMatch contents but they do not want to deal with complex search functionalities associated to a specific media. This interaction level involves the retrieval of atomic and compound cultural objects which will be shown to the user using different visualization paradigms. Rankings used to order this information may consider different sources of information: **query-sensitive** such as content matching (i.e. frequency of the query terms in the retrieved documents), **absolute criteria** such as popularity (i.e. most cited creators, links to pages about the creator, most queried creator) or **freshness** (i.e. presence in the most recent news feeds).

Specialized Interaction Level

MultiMatch will also provide users with a specialized interaction level which will allow them to directly query specific search services such as creators/creations search engine, audiovisual search or browsing service. Retrieved results will be rendered according to the search service queried, and shown to the user using specialized interfaces.

The following subsections describe the MultiMatch approach to support the specialized search in the final version.

- **Creators/Creations Search.** The general idea of this specialized search level is that, for a given type of cultural entity (creator/creation) plus (optionally) free text**,** the user can query the MultiMatch system to retrieve all the information available about it (e.g. the user can query the MultiMatch system about *Van Gogh* and then retrieve all the information available about the painter). As described before, MultiMatch architecture is focused on indexing not only standalone resources, such as web pages, images, audio or video (atomic cultural objects), but also on compounding and indexing complex cultural objects based on mined and extracted information from different sources. In this way, creators/creations search exploits these complex indexes to provide the user with a more understandable representation of the retrieved information which can be used by the user to browse and explore relevant information. Although it is still under discussion and it can be considered a research issue, creators/creations specialized search should provide the user with the following information:

    - **MultiMatch cultural objects** (atomic and/or compound) information related to the query which could be also organized according to certain categories such as biographical data (i.e. resources with information about a creator's life), news/events (i.e. RSS feeds related to a creator) or works of art (i.e. the text in case of a poem, an image if it is a painting, sculpture, building, etc.), or reviews (i.e. texts in which the artwork is the subject) or news (i.e. related RSS feeds). Also, for creators it should be desirable to provide a list of creations and cultural heritage sites related with the creator and for creations, a list of cultural heritage sites where the creation is exhibited or reviewed.
    - **Alternative ways of exploring creator/creation related information.** MultiMatch should create relations between cultural objects to allow the user to browse and discover information related with his current search. A representation of the relationships between the current creator/creation queried and all the closely related neighbours should improve knowledge about the creator/creation's context. Term clouds can be used to provide the user with relevant information about terms/phrases more or less strongly related to the creator/creation queried. Finally, timelines can provide the user with valuable information about creator/creation impact across time.

- **Audiovisual Search.** This type of information can be considered as multimodal, which implies that pure visual contents (images and videos) are also related with spoken contents, associated metadata, and texts describing the contents. Searching for multimedia content will thus often require fusion and multimodal merging of features at various levels of search. The MultiMatch system will provide three different specialized searches on multimedia contents:

    - **Image Search.** MultiMatch will offer the possibility of retrieving still images and video keyframes based on text and image queries using multimodal searching. Also image relevance feedback can also be used to locate relevant images based on visual content. The retrieved image list will give access to: image thumbnails, original images, sources of the images and its relevance level.

  - **Video Search.** MultiMatch will offer the possibility to search for video contents using text queries and also image queries. As for the image search, multimodal search will be applied to improve the retrieval performance and the user will be provided with a set of representative keyframes used to either identify relevant video for playback or as a simple static summary which may be sufficient to satisfy their information need. We envisage that it will also be possible to search video based on associated ASR-generated transcripts of the video's soundtrack.
  - **Audio Search**. Users will be able to perform audio search to retrieve audio documents and also video documents by way of their speech tracks. An index built from these transcripts will make audio search possible.

From the user perspective, audiovisual search will operate as follows. The user will submit a free text query. Audiovisual documents relevant to this query (i.e. containing the query in the speech recognition transcript) will be shown, allowing the user to start playing the audio or video document before the occurrence of the first query word. The speech recognition transcripts will not be displayed to the user since practice has demonstrated that users find speech recognition transcripts (typically error ridden) to be more confusing than helpful. Also we will explore the possibility of showing snippets from the audio transcript. In the second prototype, the speech recognition vocabulary will be domain-adapted using information from either the production metadata of the audio or from the Internet context (html or rss feeds) in which the audio occurs. Also planned for the second prototype is post-recognition error correction. Both methods hold promise for improvement of audio search precision and recall.

- **Browsing Functionality.** Users will be able to directly explore the MultiMatch indexes without posing queries, by using a number of browsing facilities:

  1. Creators and creations will be classified according to the MultiMatch ontology. In this way, it is expected that the user can navigate and explore a cultural heritage knowledge space, accessing atomic and compound cultural objects. For instance, using this approach, the user could easily find all the authors related to a specific art period.
  2. Authors and works of art will also be listed alphabetically to allow the user to deal with MultiMatch contents in an authority list fashion.
  3. As a complementary way of browsing within MultiMatch, users will be able to navigate and explore the contents using the alternative representations attached to the specialised searches.

When interacting with a retrieval engine, users are normally perusing a particular need for information, rather than merely randomly gathering information on a topic. Bearing this in mind, browsing a fixed document structure or one adapted to previous search history or user preferences may be inefficient for the user in satisfying their current information need. They may be required to follow many irrelevant links (since they cannot know with certainty where the link will lead based on anchor text, thumbnail images, etc) or to view large amounts of non-relevant content within individual documents or objects (where the content fulfilling their information need is embedded with a large amount of non-relevant material). MultiMatch proposes to explore a search-based novel approach of focused browsing. Analogous to focused crawling, where the content collected for indexing is selected based on topically relevant criteria, focused browsing will bias the selection and presentation of content towards material more likely to be relevant to the user's information need. Selection and presentation will be based on the current search query and the users' interactions with content presented to them so far for this search (explicit indications of relevance or based on links followed while browsing the available content). Thus while browsing within content organised using the MultiMatch ontology or other content organisation, the user may be presented with summary snippets of documents based on the focus of their search thus far and links to other related content may be highlighted based on the likely relevance of the destination of the link to the information need. Such summaries and link markings will be updated dynamically based on the user's continuing search activity as they continue to explore the available information. By selecting content for presentation to the user and highlighting links to other content based on their relationship to the users' current search query, focused crawling aims to improve the efficiency of user access to relevant material and their overall satisfaction with their search experience.

The MultiMatch entry page can also be used to display alternative classifications of the MultiMatch contents using, for instance, popularity criteria. Rankings can be made with both user logs and index size. Again, a focused browsing approach could customize this MultiMatch home page according to the user's needs and profile.

Log in and User Profile Specifications

MultiMatch is planned as a web search engine to be used without any **log in** requirement. In this way, users can directly connect to MultiMatch and perform their queries as they do with general search engines.

However, in order to provide the user with more accurate search results and browsing facilities, MultiMatch will implement **log in** and **create account** features. **The create account** facility will allow the user to register in MultiMatch defining a set of features (such as preferred languages, user profile, search preferences, filtering, etc.) which could be used by default in MultiMatch to properly tailor search results to the user needs (for instance a university lecturer may register as an educational user and perform searches oriented to this specific profile). Logging in to MultiMatch will also activate some search specific functionalities such as:

- **Search history**. This stores previous user searches. It is still under discussion if the system will automatically store queries performed by the users (allowing them to make a post filtering to discard those not relevant) or whether users will manually perform this activity at search time saving only those queries really relevant for them.
- **Workspace facility**. This allows the user to save pointers to relevant resources and cultural items retrieved on previous searches for future references.

Not logging in to MultiMatch does not imply a restricted access to its search services (the user will be able to perform all the types of searches described above) but will have a reduced possibility of customizing search results.

Proprietary indexed contents will be provided to the users as low quality files (i.e. a low resolution image in the case of images, a short and low resolution video in the case of videos and a fragment of the whole text in the case of texts) with the corresponding link to the vendor site, allowing them to buy or to access the full contents using the specific selling or accessing policies of each content provider.

Search Facilities

Besides the default text box, the MultiMatch search engine will feature advanced search facilities which include those most highly rated in the user survey:

- **Boolean search facilities**. To include in the results only those resources that fit into the Boolean expression. This feature can be easily adapted to both types of queries defined above.
- **Field search**. As users dealing with cultural heritage are familiar with database searches, this functionality will allow the users to type their queries using fields similar to when searching databases.
- **Relevance Feedback**. This allows users to launch new searches based on a query automatically built using the text information stored in those documents considered to be relevant to the user. The MultiMatch second prototype will also enhance and improve the basic image relevance feedback implemented for the first prototype.

Other advanced search features will be set either as search preferences or at query time:

- **Filter by language.** Any combination of processed MultiMatch languages.
- **Filter by type of file**. This will retrieve only those resources which are of a specific file type (e.g. html, pdf, jpg, ppt, gif, avi, etc)
- **Filter by date.** This will retrieve only those resources which fit into a specific publishing date range (e.g. from 12-10-2005 to current date)
- **Filter by size.** This will retrieve only those resources which fit into an specific size range (e.g. not greater than 500 kb)

Some additional features will be set only as search preferences relating to the user profile. The most relevant features are the **user's language skills**: native language(s), active/passive/unknown languages, preferred query language(s), etc.

Meta Search Functionalities

According to the user survey, general search engines are one of the most used resource finding strategies for CH experts. Therefore, for both search interaction levels we will aggregate results from these sources. Live results from Google, Yahoo or other sources will be combined and clustered using specialized techniques that will exploit the internal MultiMatch indexes to organize the search results.

2.3 Interface Functionalities

MultiMatch will have several interfaces:

- **A web-based user interface.** All the MultiMatch functionalities will be shown via a full web application which will be highly configurable and where registered users should be able to organize MultiMatch search services in their own way (suppressing and adding result boxes or search fields from all main search interfaces). There will be default interface configurations for education, CH and tourism search profiles.
- **Set of simple web interfaces.** It is also planned to provide a set of simple web applications which exploit specific MultiMatch functionalities as standalone services, being accessible through different URLs. A clear isolation of web services, allowing individual and very specific tasks, will make this task possible. Independent interfaces not only provide users performance with simpler applications but also ease the evaluation of specific functionalities within the complex MultiMatch architecture during the development process.
- **An API (Application Programming Interface) providing web services for third-party applications.** As one of the main goals of MultiMatch is to disseminate the findings and results of the project, it is planned to provide the external users with a set of APIs which ease the creation and personalisation of MultiMatch services in their own web sites.

3 Functional Specifications of the Second MultiMatch prototype

This section summarizes the set of new functionalities expected for the second prototype in order to support the goals proposed in the previous section. A complete list of specifications with their priorities can be found on Appendix I of this document.

3.1 Content Selection and Preparation

The second MultiMatch prototype will use the source contents crawled and indexed for the MultiMatch first prototype enhanced with new contents provided by third party content providers, aggregated from external CH sources and extracted from news feeds as well. It is also planned for this second prototype to extend the number of languages covered by MultiMatch to **German** and **Polish**.

Specifically, the set of contents that MultiMatch will be able to store and retrieve will be the following (see also [3] for a detailed explanation):

- **Web Pages.** Web contents crawled and indexed for the first prototype will be reused and updated if necessary. Focused crawlers will also spider the web to automatically identify and index new web contents on the Cultural Heritage domain. Specifically, for this second prototype, web page crawling will target the 500,000 pages listed as success indicator in the DoW [4]. This will include the extended white list crawl, and information gathered from RSS feeds in English, Spanish, Dutch, Italian, German and Polish languages.

- **Wikipedia.** Wikipedia contents crawled and indexed for the first prototype will be reused and updated if necessary. Also, the wiki pages in the new languages (German and Polish) and their images will be crawled and added to the MultiMatch indexes.

- **Proprietary contents provided by Alinari, Sound and Vision and Biblioteca Virtual Miguel de Cervantes partners.** Proprietary text, image, audio and video contents indexed for the first prototype will be reused. Also, each content provider will increase the number of resources provided to the consortium as shown in Table 3.1
- **OAI Compilant Resources.** MultiMatch will also index Open Archive Initiative compliant contents. Specifically we will focus on MICHAEL (Multilingual Inventory of Cultural Heritage in Europe) and TEL (The European Library) contents.
- **News Feeds.** As a new and experimental information source for the second prototype, MultiMatch will index a set of cultural heritage news feeds identified from authoritative cultural heritage sites and cultural sections of newspapers. Domain specific news feeds are a dynamic information source which can provide users with updated information about cultural events (e.g. exhibitions, book releases, conferences, etc.). Although including news feeds into MultiMatch is an exciting challenge, it also has some technical drawbacks mainly related with crawling, identification, indexing and classification of such items which implies identifying towards relevant RSS sources able to provide authoritative and non redundant information. For this reason, the first task to accomplish RSS integration will be a study of currently available sources and its feasibility to be integrated within the search engine.
- **External Resources.** As a last information source, external search resources will be integrated into MultiMatch. The idea is not only to provide MultiMatch users with search results taken from MultiMatch indexes, but also with a set of resources retrieved from external sources which are not previously indexed and classified by the system in order to complement the information delivered to the user. Again, this is a challenge as there are many ways of managing this information and the way in which it is going to be aggregated to the original MultiMatch result set is still under discussion. As candidate search sources to be considered we can note the following: General Search Engines (such as Google or Yahoo!), current and

under development initiatives (such as the Dutch CATCH[3] program) or media specific meta search engines (such as Blinkx).

The first prototype development included the design of a metadata schema to meet the needs of MultiMatch and which covers the main aspects to be taken into account for the indexing, retrieving, organizing and browsing tasks. During the first period all the efforts were focused on the development of this schema and its global coherence, not being possible to fully populate it (i.e. only simple mappings were performed from proprietary data to the MultiMatch metadata). One of the main challenges for the second prototype is to define the conventions that will dictate how automatically generated metadata is going to be added and accommodated to the MultiMatch metadata schema in order to describe the set of CH objects in the MultiMatch system. The MultiMatch metadata schema will be revised, if necessary, to incoporate the new requirements derived from new content sources such as news feeds.

File formats supported by the 2nd MultiMatch prototype will be the following:

- **MIME types:** plain, html, xml. For text documents.
- **MIME types:** mpeg, xwav. For audio documents.
- **BMP, JPG, GIF, TIFF and PNG.** For still image documents.
- **MPEG1.** For video documents.

Table 3.1. New contents provided by MultiMatch content providers

Alinari	5,000 still images (jpg)
BVMC	7,000 texts (articles, books, etc.) 1,500 web pages
Sound and Vision	900 videos (mpeg-1. ~300 hours) metadata of the programma catalogue in XML thesaurus of names, places and keywords in RDF Between 5.000-9.000 texts from the Beeld en Geluid wiki

3.2 Multimedia/Multilingual Indexing and Information Extraction

In the second prototype, a crawl of cultural heritage material on the Internet will be carried out that will serve to expand the collection of material harvested from the Internet for prototype one. This crawl concentrated on collecting internet content whose information value is independent of time, i.e. static material. The first prototype used a semi-manually generated white list to drive the crawler to collect material from specific domains related to cultural heritage. In the second prototype, this crawl will be extended to collect additional material. For the purpose of the project, this material will be

[3] Continuous Access to Cultural Heritage

considered to be persistent across time, which seems a reasonable assumption for CH-domain material.

In addition to crawling additional (conventional) Internet content in the CH domain, the second prototype will also extend the crawling activities to gathering RSS-feed-based content. RSS-feed based content in the CH domain can be expected to contain time-sensitive information, such as information about exhibitions and news. RSS-feed based content will be stored in supplementary time-sensitive indexes. Regular updates will be scheduled. A mechanism to keep the full index fresh can always be added in a post-project exploitation phase, and does not have a specific research value itself.

Regarding indexing, the MultiMatch second prototype will provide the following functionalities:

- **Text Indexing:**
  - Standard CLIR indexing methods (tokenization, language dependent morphological normalization techniques such as lemmatization or stemming, language independent techniques such as character n-gramming).
  - Improvement and augmentation of techniques used in the first prototype. For instance, it is planned to explore the use of n-gram indexing as a means of examining language independent indexing and retrieval.
  - New indexing resources will be needed for additional languages.
  - Classifiers will be trained to classify web documents putting pages into categories by subjects. The main topics on which MultiMatch will concentrate on will be periods/styles of art and creators. These topics will be drawn from previously existing lists, in particular the **Getty** thesauri. Classifiers will be also trained to discover relations between web pages and Wikipedia articles and they will steer the focused crawler.
- **Image indexing:**
  - Based on image feature extraction and metadata indexing.
  - External images coming from web pages or wikipedia will be also indexed. Image annotation based on related text should be possible, e.g. related text from a page or anchor texts pointing to web images.
- **Speech indexing:**
  - ASR applied to generate transcripts from speech. For this second prototype, only German will be considered as a new language to be indexed as speech.
  - Metadata associated with speech (if available) will be indexed.
  - To deal with noisy transcripts we will use sub-word indexing terms, including syllables and n-grams
  - It is expected to apply text tiling techniques to audio, so that audio segments homogeneous with respect to topic can be retrieved.
  - It is also expected to include indexing mechanisms to allow rapid audio browsing (may be using term clouds).
  - The MultiMatch second prototype will focus on Sound and Vision video material and on CH themed podcasts on the Internet.
  - Internet audio is characterized by sparseness of associated metadata. We will enrich metadata by training classifiers that can take sparse metadata and speech transcripts and produce class labels. Class labels will be topics, and we will especially focus on creators and periods/styles. Additionally, for dynamic audio sources (podcast) we

would like to detect and track emerging themes, especially those related to creators and to cultural heritage events such as exhibitions.
  - As for text indexing, classifiers will be trained to function on speech recognition transcripts, making it possible to relate documents of different media types.
  - Exploratory work will be carried out to alleviate the effects of out of vocabulary words in the speech recognizer by exploiting information contained in the RSS feeds.

- **Video indexing:**
  - Representative keyframes identified and indexed using their visual features..
  - Shot boundary detection.
  - Indexing of metadata associated with video and transcriptions to allow standalone text retrieval over visual resources and multimodal retrieval.

Semantic enrichment of content will consist of generating class labels for documents and finding links between them. Extensive discussions about which are the best class labels to focus on have been taken and are still ongoing. As a first approach it is planned to focus our attention on creators and periods/styles labels although place/time, art objects/works, type/genre and exhibitions are also considered.

The MultiMatch second prototype will focus on linking web pages with Wikipedia and thesauri such as **ULAN**, this being the initial ground to provide the user interface with a basic set of browsing and retrieval functionalities. As far as possible all use of linkage information should be at indexing time to minimise use of computation resources at retrieval time. Linkage such as anchor text can potentially be used to annotate images and improve descriptions of other documents.

Also, it is planned to provide MultiMatch with a new index, which will store a **profile** related with the cultural object retrieved and that will provide the user with an overview about its main features (i.e. name, related dates, period/style, etc.)

3.3 Multimedia/Multilingual Information Retrieval

With respect to Multimedia IR, the second MultiMatch prototype functionalities will be the following:

- Improvement of still images and video (this last based on representative keyframes) retrieval. Retrieval based on image content and context will be available and techniques will be developed to perform the fusion of these search results. Multi-modal information will be exploited in a consistent manner via information fusion techniques allowing a combined understanding of the content and thus better retrieval. Multi-modal fusion will be performed at multiple levels encompassing the information indexing stages and the interactive search and retrieval phases
- In order to make image content based retrieval more effective from the user perspective we will provide an image similarity navigation functionality. Users will start with an initial query as seed (a keyword base query can also be used) and they can use the images in the results to refine the query interactively navigating by image similarity.
- As a complement to pure image relevance feedback (providing only a set of images as relevant and not relevant), this second prototype will provide multimodal relevance feedback based not only in the use of images as queries but also text.

- With regards to video, the user will be able to retrieve relevant portions of a video clip based on ASR transcript and visual metaphor tools.

Multilingual functionalities for the second prototype will be the following:

- Improvement of translation strategies with respect to the first prototype. Specifically we will focus on the improvement in quality and coverage of translation dictionaries, in disambiguation methods, in phrase translation resources. The second prototype will also explore how to locate and use comparable corpora to improve translation accuracy and integrated hybrids of machine translation and dictionaries.
- The improvement of the presentation of results of expansion and retrieval to users not familiar with document language will be considered.
- In order to allow the user interface to accommodate multilingual results, it is necessary that retrieval components are able to provide it with some extra information or services such as translation of retrieved snippets, a link to translated web pages or the display of alternative definitions (where applicable) with some sort of back-translation or pictorial representation.
- Textual relevance feedback will be improved with respect to the first prototype. The use of enhanced feedback will be explored, e.g. based on automated document summaries, term proximity in documents, and multilingual relevance feedback, i.e. potentially relevance feedback information from documents in different languages.

Cross-media aggregation functionalities will be the following:

1. Cross-media retrieval based on a unique query (textual or visual).
2. Different types of ranking depending on document matching scores and user preferences.
3. The presented document set will also be dynamically adaptable (via relevance feedback) for further exploration of the current search request based on user feedback for information relevance, and updated preferences for media and language sources.

Finally, the implementation of dynamic summarization will be studied for the second prototype. Ideas such as the exploitation of links, document similarity comparison, novelty analysis, documentation segmentation and automatic summary composition, and taking account of the user query to provide query-biased output are still under discussion.

3.4 User Interaction and Interface Design

The set of functionalities that should be included in the interface of the final prototype are described below. These are also designated as research themes, which means that the best way of presenting and implementing them will be based on empirical studies.

- Faceted browsing: Using the common MultiMatch ontology to navigate material by semantic relationships (e.g. location, creator, time) in order to enable individuals to browse in a focused manner for content based on one or more facets (exact nature of which to be determined.).
- Use of ontologies and semantic information to explore relations between entities. Identifying relations between entities and displaying these in some graphical visualization.
- Enhanced and expanded collection overviews.
- Use of *profiles* to give a brief summary of key information about a person/creation/institution

- Means of navigating by place and time (interactive timelines and/or maps). Enabling users to browse material temporally and/or spatially (e.g., by arranging items on a timeline based on associated dates, or displaying objects on a map to allow the examination of the material by associated location.)
- Cross-media fusion. For example, providing related results in all types of media (not just one) and allowing seamless navigation across types
- More advanced means of clustering and refining results. Enabling query refinement in the case of ambiguous queries or queries yielding large results sets. Also, clustering to avoid duplication of the same result coming from different sources.

4 Bibliography

[1] Functional Specification of the First Prototype. Deliverable 1.3.

[2] Analysis of User Requirements. Deliverable 1.2.

[3] Selected and Harmonized Content. Metadata. Deliverable 2.3.1

[4] MultiMatch Project. Annex I. Description of Work.

5 Annex I. Second Prototype Specification Tables

5.1 Contents, Crawling and Indexing Functionalities

Functional Specification 1.1. Languages Supported

Reference No.	Document Languages	Priority[4]
1.1.1	English, Dutch, Italian, Spanish, German, Polish	High

The task of identifying and getting content in such languages is high priority.

Functional Specification 1.2. Document Types

Reference No.	Document Type	Document Format(s)	Priority
1.2.1	Audio	MIME types audio/mpeg, audio/x-wav	High
1.2.2	Still Image	bmp, jpg, gif, tiff, png	High
1.2.3	Text	MIME types: text/plain, text/html, text/xml	High
1.2.4	Video	mpeg1	High

The task of identifying and getting content in such formats is high priority.

Functional Specification 1.3. Document Sources

Reference No.	Document Source	Description	Priority
1.3.1	Web pages	>500,000 success factor (~50 feeds, ~100 hour audio minimum)	High
1.3.2	RSS Feeds, podcasts and vodcasts		Medium
1.3.3	Alinari Collection	+5000 still images	High
1.3.3	BVMC	+7000 Texts +1500 Web pages	High
1.3.4	Sound and Vision	900 videos (mpeg-1. ~300 hours) metadata of the programma	High

[4] Where "high priority" indicates that this functionality is considered as essential and it is our intentino to implementi t: "medium priority" indicates that this functionality is considered as desirable and we hope to implement it.

		catalogue in XML thesaurus of names, places and keywords in RDF Between 5.000-9.000 texts from the Beeld en Geluid wiki	
1.3.5	Wikipedia	Wiki pages in the new languages (German and Polish)	Medium
1.3.6	OAI Compilant Resources	MICHAEL and TEL contents	High
1.3.7	External Resources	External Search engines to be queried by MultiMatch	Research Issue (No commitment for implementation)

Functional Specification 1.4. Crawling

Reference No.	Document Source	Description	Priority
1.4.1	Web pages crawling	One crawl scheduled for the 2nd period	High
1.4.2	RSS Crawling	One crawl is scheduled for the 2nd period and regular updates	Medium

Functional Specification 1.5. Indexing protocols

Reference No.	Document Set	Description	Priority
1.5.1.	For all textual documents	Pre-processing: Language dependent techniques will be applied for stop words removal, stemming and synonym matching.	High
		Testing of new weighting formulas apart of tf-idf such as BM25 or Okapi	Medium
		Improvement and Augmentation of techniques used in 1st prototype	High
1.5.2	For all images and also video keyframes	Indexing using standard low-level visual feature extraction (i.e. colour, histogram, texture) and encoding with agreed metadata mark-up.	High
		Image descriptions and associated metadata will be indexed to allow text querying for image retrieval.	High
1.5.3	For all video content	Video indexing using shot boundary and automatic	High

		keyframe detection.	
		Automatic Speech Recognition (ASR) for those languages for which it will be available.	Medium
		MultiMatch will index the provided transcriptions (when available) or the ASR output (when available).	High
		Keyframe extraction for still image indexing.	High
		Indexing of metadata associated with video and transcriptions	High
		Mechanisms to ease video browsing	High
1.5.4	For all Spoken content	Automatic Speech Recognition (ASR) for those languages for which it will be available (hopefully for EN, DE, ES, IR, NL).	High
		Metadata associated with speech	High
		Techniques to deal with noisy transcripts	Research Issue (No commitment for implementation)
		Text tiling techniques	Research Issue (No commitment for implementation)
		Mechanisms to ease audio browsing	Medium
1.5.5	For all metadata	Metadata mapping to the MultiMatch metadata Schema	High
		Metadata indexing and association with the textual or audiovisual resource if available	High
		Provision of mechanisms to support indexing and retrieval using native schemas	Research Issue (No commitment for implementation)

Functional Specification 1.6. Information Extraction and Classification

Reference No.	Document Set	Description	Priority
1.6.1.	Generation of class labels to semantically	Creators and periods/styles	High
		Place/time, art objects/works,	Medium

	enrich contents	type/genre and exhibitions	
1.6.2	Linkage between contents at indexing time	Web pages with wikipedia and thesauri	High
		Between other resources	Medium

5.2 Search Functionalities

Functional Specification 2.1. Search levels

Reference No.	Search Level	Description	Priority
2.1.1.	Default	MultiMatch will provide a combined search facility for free text, image and audiovisual retrieval and metadata based retrieval as well	High
2.1.2.	Specialized	MultiMatch will provide standalone search facilities for text, image and audiovisual retrieval and metadata based retrieval as well It will also include a general browsing facility based on categories and extracted terms.	High

Functional Specification 2.2. Search modes

Reference No.	Search Mode	Description	Priority
2.2.1.	Default	MultiMatch will perform monolingual retrieval using the user's native language as a default mode.	High
2.2.2.	Advanced	Search customization facilities: • Multilingualism (retrieved results in different languages will be presented as different document sets) • Filtering of search results under several criteria	High

Functional Specification 2.3. Multimedia/Multilingual IR

Reference No.	Search Facility	Description	Priority
2.3.1	Multilingual	Improvement of translation strategies and services	High

		Improvement of text relevance feedback	High
		Improvement of boolean operators implemented for 1st prototype	High
2.3.2	Multimedia	Improvement of still image and video retrieval.	High
		Improvement of image relevance feedback	High
		Multimodal fusion techniques	Medium
		Multimodal relevance feedback	Medium
2.3.3	Cross-media Aggregation	Cross-Media retrieval based on a unique query	High
		Different types of ranking	Medium
2.3.4.	Dynamic Summarization	Still under discussion	Research Issue (No commitment for implementation)

5.3 User Interaction and Interface Design

Functional Specification 4.1. Interfaces provided

Reference No.	**Search Level**	**Interface**	**Priority**
4.1.1.	Default	Web based Interface for default search level	High
		Application Programming Interface for default search level	High
4.1.2.	Specialized	Web based interface for specialized search Level (images, video and browsing).	High
		Application Programming Interface for specialized search level.	High
4.1.3		Set of standalone interfaces to test individual functionalities of the system	Medium

Functional Specification 4.2. Accessing functionalities

Reference No.	**Accessing Level**	**Description**	**Priority**
4.2.1.	Anonymous	Perform searches on any of the defined search levels but without advanced features such as search history.	High

4.2.2.	Registered	Search history	Medium
		Workspace	Medium
		Customization	Medium

Functional Specification 4.3. Browsing functionalities

Reference No.	**Functionality**	**Priority**
4.3.1.	Visualization of the words most frequently appearing in the collection and most related to a set of search results (e.g. term cloud facility).	Medium
4.3.2.	The user will be able to explore the collection via overview (e.g. a display or collage of items randomly chosen from the collection)	Medium
4.3.3.	The user will be able to browse the collection, when possible, by categories, based on pre-existing metadata information	High
4.3.4	Navigation based on creator/creation profile	Medium
4.3.5	Navigation by place and time	Research Issue (Not commitment for implementation)
4.3.6	Navigation across cross-media fused results	Medium
4.3.7	Dynamic clustering of search results	Research Issue (Not commitment for implementation)

6 Annex II. First Prototype Achievements

6.1 Crawling and Indexing Functionalities

Functional Specification 1.1. Languages Supported

Reference No.	Document Languages	Status
1.1.1	English, Dutch, Italian, Spanish	Done

Functional Specification 1.2. Document Types

Reference No.	Document Type	Document Format(s)	Status
1.2.1	Audio	MIME types audio/mpeg, audio/x-wav	Done, but not all formats covered
1.2.2	Still Image	bmp, jpg, gif, tiff, png	Done, but not all formats covered
1.2.3	Text	MIME types: text/plain, text/html, text/xml	Done
1.2.4	Video	mpeg1, mpeg4, mpeg7	Done, but not all formats covered

Functional Specification 1.3. Document Sources

Reference No.	Document Source	Description	Status
1.3.1	200 cultural heritage institutions and their web sites	Materials from white list crawl focused on museums. **Crucial:** 10,000 pages UK and Ireland **Additional:** 10,000 pages from Spain, Italy and Holland pages for a total of 40,000 pages	Done
1.3.2	Wikipedia contents	Wikipedia contents and wikipedia metadata. Focused on Wikipedia **artists** and **museums** categories. **Crucial:** Content in four languages (at least 10,000 items) **Additional:** Images	Done
1.3.3	Alinari collection	Image documents with metadata.	Done
1.3.4	BVMC	Text documents with metadata.	Done
1.3.5	Sound and Vision videos	Audiovisual contents with metadata.	Done

Functional Specification 1.4. Indexing protocols

Reference No.	Document Set	Description	Status
1.4.1.	For all textual documents	1.4.1.1 Pre-processing. Language dependent techniques will be applied for stop words removal, stemming and synonym matching.	Done
1.4.2	Subset of indexed documents from 200 CH institutions	1.4.2.1. Document indexing using text indexing techniques	Done
		1.4.2.2. A manual CIDOC-compliant annotation on a subset of the 200 cultural heritage institutions will be made.	Not Done
1.4.3	Wikipedia contents	1.4.3.1. Document indexing using text indexing techniques.	Done
		1.4.3.2. Items parsing to generate simple metadata.	Done
1.4.4.	Biblioteca Virtual Miguel de Cervantes	1.4.4.1. Document indexing using text indexing techniques.	Done
		1.4.4.2. Items parsing to generate metadata.	Done
1.4.5	For all images crawled, extracted from image libraries and also video keyframes	1.4.5.1 Indexing using standard low-level visual feature extraction (i.e. colour, histogram, texture) and encoding with agreed metadata markup. Also image descriptions and associated metadata will be indexed to allow text querying for image retrieval.	Done
1.4.5	Allinari Collection	1.4.5.1 Image indexing using image indexing techniques.	Done
		1.4.5.2 Metadata contents indexing using text indexing techniques.	Done
		1.4.5.3. Image parsing to generate metadata.	Done
1.4.6	Sound and Vision videos	1.4.6.1. Video indexing using shot boundary and automatic keyframe detection.	Done
		1.4.6.2. Automatic Speech Recognition (ASR) will be employed for those languages for which it will be available. When available, MultiMatch will make use of provided transcriptions. 1.4.6.3. Keyframe extraction for	Done

		still image indexing.	

6.2 Search Functionalities

Functional Specification 2.1. Search levels

Reference No.	Search Level	Description	Status
2.1.1.	Default	2.1.1.1 MultiMatch will provide a combined search facility for free text, image and audiovisual retrieval.	Done
2.1.2.	Specialized	2.1.2.1 MultiMatch will provide standalone search facilities for image and audiovisual retrieval. It will also include a general browsing facility based on categories and extracted terms.	Done

Functional Specification 2.2. Search modes

Reference No.	Search Mode	Description	Status
2.2.1.	Default	2.2.1.1 MultiMatch will perform monolingual retrieval using the user's native language as a default mode.	Done
2.2.2.	Advanced	2.2.2.1 Search customization facilities: • Multilingualism (retrieved results in different languages will be presented as different document sets) • Filtering for file type and size	Done

Functional Specification 2.3. Search facilities

Reference No.	Search Facility	Description	Status
2.3.1.	Relevance feedback	2.3.1.1. Text relevance feedback	Done
		2.3.1.2. Image relevance feedback	Done
2.3.2.	Boolean search	2.3.2.1 Typical of these boolean operators would be “AND”/ “OR” searches	Partially Done. Finalization moved to 2nd prototype

2.3.3.	Visual search	2.3.1.1 Using low-level features (e.g. colour histograms) only available for the indexed document collection (not for the external images).	Done

6.3 Retrieval Functionalities

Functional Specification 3.1. Retrieved sources

Reference No.	**Type of Retrieval**	**Sources**	**Status**
3.1.1.	For text retrieval	3.1.1.1. Cultural heritage web sites indexed	Done
		3.1.1.2. Wikipedia items indexed.	Done
		3.1.1.3. Biblioteca Virtual Miguel de Cervantes contents indexed.	Done
3.1.2.	For still image retrieval	3.1.2.1. Still images indexed from web pages and/or Wikipedia	Done
		3.1.2.2. Alinari still images repository	Done
		3.1.2.3. Sound and Vision video keyframes indexed as still images	Done
3.1.3.	For video retrieval	3.1.3.1 Sound and Vision audiovisual contents indexed	Done

Functional Specification 3.2. Retrieved results

Reference No.	**Search Level**	**Description**	**Status**
3.2.1.	Default	3.2.1.1. A list of documents, ranked according to some relevance criteria. Text documents will be summarized using query-biased snippet generation techniques which display the text either in original language or in English.	Done. Snippets translation to English has been moved to 2nd prototype
		3.2.1.2. A list of still images, ranked according to some relevance criteria. When possible, the retrieved information will be combined with metadata to provide the user with more comprehensive information about the images.	Done

		3.2.1.3. A list of videos, ranked according to some relevance criteria When possible, the retrieved information will be combined with metadata to provide the user with more comprehensive information about the videos.	Done
3.2.2.	Specialized	3.2.2.1. The image search service will provide the user with a image list, ranked according some relevance criteria Direct access to image thumbnails, original images and sources of the images will be provided	Done
		3.2.2.2. The video search service will provide the user with a video list, ranked according to some relevance criteria,	Done

6.4 Interface Functionalities

Functional Specification 4.1. Interfaces provided

Reference No.	Search Level	Interface	Status
4.1.1.	Default	4.1.1.1. Web based Interface for default search level	Done
		4.1.1.2. Application Programming Interface for default search level	Done
4.1.2.	Specialized	4.1.2.1.Web based interface for specialized search Level (images, video and browsing).	Done
		4.1.2.2. Application Programming Interface for specialized search level.	Done

Functional Specification 4.2. Accessing functionalities

Reference No.	Accessing Level	Description	Status
4.2.1.	Anonymous	Perform searches on any of the defined search levels but without advanced features such as search history.	Partially Done. Finalisation moved to 2nd prototype

4.2.2.	Registered	4.2.2.1. Search history	Partially Done. Finalisation moved to 2nd prototype
		4.2.2.2. Manual setting of language preferences.	Partially Done. Finalisation moved to 2nd prototype
		4.2.2.3. Automatic setting of user's preferred languages	Partially Done. Finalisation moved to 2nd prototype

Functional Specification 4.3. Browsing functionalities

Reference No.	Functionality	Status
4.3.1.	Visualization of words most frequent in the collection (e.g. term cloud facility).	Moved to 2nd prototype
4.3.2.	The user will be able to explore the collection via overview (e.g. a display or collage of items randomly chosen from the collection)	Moved to 2nd prototype
4.3.3.	The user will be able to browse the collection, when possible, by categories, based on pre-existing metadata information (i.e. this facility will be only available for specific contents indexed in MultiMatch first prototype).	Partially done, Finalisation moved to 2nd prototype

Functional Specification 4.4. Language facilities

Reference No.	Functionality	Status
4.4.1.	The user will be able to retrieve documents in the language of the query as well as others (CLIR)	Done
4.4.2.	The user's query will be translated into the other languages of the project and suitable matches will be identified	Done
4.4.3.	The user will be able to view possible translations (in the case of an ambiguous term) and select the preferred version	Done
4.3.4.	Users can view results in languages besides their preferred language. The presentation mode can be based on preference (e.g. user can elect to display a summary of the document in both original language and translated version, if necessary.)	Moved to 2nd prototype

Functional Specification 4.5. Presentation and organization facilities

Reference No.	Functionality	Status
4.5.1.	Results will be displayed in an arrangement that is deemed to be preferred by users (e.g., in a grid, slideshow format, etc.)	Partially done, Finalisation moved to 2nd

		prototype
4.5.2.	The user will be able to control or re-arrange the results display	Done
4.5.3.	The user will access text results as a summary showing the query in context (e.g. the query will be highlighted wherever it appears)	Partially done, Finalisation moved to 2nd prototype
4.5.4.	The user will be able to sort or re-organise the results in a simple way (e.g. by size/file type)	Done
4.5.5.	As another form of relevance feedback, the user will be given the option of selecting relevant objects or placing them into a "workspace" to signal relevance feedback.	Done
4.5.6.	The user will be able to group or cluster the results based on simple criteria (e.g. colour)	Moved to 2nd prototype
4.5.7.	The user will be able to read a translation of foreign language results; this translation will be displayed to the user based on his/her preferred presentation style.	Moved to 2nd prototype
4.5.8.	The registered user will be able to perform manual annotation of results.	Moved to 2nd prototype