Project no. 033104

MultiMatch

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

D1.3 Functional Specification of the First 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)

Document Information

Deliverable number:	D1.3
Deliverable title:	Functional Specification of the First Prototype
Due date of deliverable:	31 October 2006
Actual date of deliverable:	14 December 2006
Author(s):	Julio Gonzalo, Juan M. Cigarrán (UNED)
Participant(s):	All
Workpackage:	1
Workpackage title:	User Requirements & Functional Specification
Workpackage leader:	Julio Gonzalo UNED
Dissemination Level:	PU
Version:	Final
Keywords:	user requirements, user groups, education, cultural tourism, cultural heritage

Table of Contents

Executive Summary 4
1 MultiMatch System Overview 5
1.1 What will MultiMatch Index? 5
1.2 Which Retrieval Functionalities are Expected? 6
1.3 Information Stored in MultiMatch Indexes 14
1.4 Interface Functionalities 16
2 Functional Specifications for the First Prototype 17
2.1 Crawling and Indexing Functionalities 17
2.2 Search Functionalities 18
2.3 Retrieval Functionalities 19
2.4 Interface Functionalities 20
3 Conclusions 21
4 References 21
5 Acknowledgements 22
Annexes 23
First Prototype Specification Tables 23
Images 29

Executive Summary

This document focuses on the definition of the Functional Specifications for the first prototype of the MultiMatch system. This first prototype is scheduled to be integrated and tested at month 16. At a later stage, a second functional specification will be delivered to deal with deficiencies found in the first prototype and to address functionalities not covered in this first stage of the project.

For clarity, this document also includes all the expected functionalities for the final MultiMatch search engine, indicating which will be implemented in the first prototype and which are left for the second prototype. The functionality currently envisaged for the second prototype will be revised, according to the experience gained during the first stage of the project, and fully described in the functional specification for the second prototype (month 16).

We therefore 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 [1].
- 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)

In order to illustrate the functionality of the system, we have included illustrations of a possible user interface; these images should not be taken as prescriptions for the interface design; they serve as visual representations meant to clarify the intended system functionality.

The document is based on three main sources:

1. The MultiMatch project description.
2. The so-called 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 (see appendix to [1]).
3. The user requirements of the project [1], 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.

The MultiMatch functional specifications attempt to be a reasonable intersection of these three sets of input. Using this information we have (i) identified the main functionality expected from the final MultiMatch search engine, and (ii) specified which functionality will be implemented in the first prototype.

Functionality which is not covered in the first prototype is included to give a clearer picture of the project implementation plan. This is, however, subject to discussion and changes are possible during the course of the project; these will be reflected in the functional specification of the second prototype (due at month 16).

1 MultiMatch System Overview

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 WWW 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?
- **Which are the retrieval functionalities of MultiMatch?** This is the core of the MultiMatch functionality. In what ways will the system will provide access to the indexed data?
- **What information will be stored in the MultiMatch index(es)?** What information should be extracted, mined, and stored to enable the above retrieval functionalities?
- **Which are the meta-search functionalities in MultiMatch?** Besides retrieval functionalities on the internal index, we need to describe which external indexes will also be aggregated in MultiMatch search results and whether some type of organization and combination of results will be implemented within the system.
- **Which 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 initial set of intended facilities they should provide.

In the following sections we give a general overview of the functionalities expected for the final MultiMatch search engine, stating which ones will be delivered in the first prototype. The remaining functionalities are meant to give the reader a better understanding of where the project is heading; they should not be considered as binding requirements on the final MultiMatch search engine, and will be revised when preparing the functional specification of the second prototype.

The second part of this deliverable describes the set of functional specifications planned for the first prototype and also presents them as a check-list for the implementation of the first prototype.

1.1 What will MultiMatch Index?

The **main source** of information stored in the final system will be publicly accessible WWW 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
- Encyclopaedic sources (such as *Wikipedia*[1])
- 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

However, we will apply some restrictions:

1. The crawl will focus on English, Spanish, Dutch and Italian sources.
2. It will focus primarily on material created or supported with public funds, which should consist of high quality contents.

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

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

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 not be exhaustive, but will be sufficiently extensive to test the validity of the MultiMatch approach in real-life searching scenarios.

In the first prototype, we will limit ourselves to a white list crawl harvesting at least 10,000 pages from UK and Ireland and 10,000 pages from Spain, Italy and Holland, for a total of 40,000 pages. The white list will include museum sites. We will also include proprietary data samples provided by MultiMatch Cultural Heritage partners.

For the purposes of the project, indexed material will be considered to be of a persistent informational value across time. With the exception of temporary exhibitions and related material, this seems a reasonable assumption for CH-related material. Therefore, the index will grow along the project timeline, but the content on indexed pages will not necessarily be refreshed during the project. A mechanism to keep the index fresh can always be added in a post-project exploitation phase, and does not have any research value in itself.

As a **secondary source** and still under discussion, the second prototype may also index news feeds (RSS) from different cultural sites and news services related to culture (for instance, the *culture* sections of main European online newspapers). Using this source makes it possible to offer immediate information without requiring a refreshing mechanism for the whole index. Again, English, Spanish, Dutch and Italian will be the languages considered for indexing.

Other sources of cultural information will be accessed via meta-search facilities on external indexes; this is discussed in Subsection.2.2.4.

1.2 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 displayed in different ways, such as:

- Typical title plus snippet based description (possibly mixed with images and/or video and/or metadata).
- More complex structures which, using all indexed materials, are able to model specific information related to a concrete cultural heritage item such as a writer, a painter, an artwork and so on.

The first option can be created "on the fly" using the query to determine which snippet is best suited for the user's needs, while the second option will be, in general, a combination of the information in the MultiMatch ontology[3] (parsed from structured sources) and information mined and extracted from the MultiMatch collection. 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 different possible conceptualisations of cultural objects in MultiMatch. Figure 1 shows a typical title plus snippet cultural object (with some extra information such as the source category) which links with a web page about *Van Gogh* works (note that in this case the

[3] This ontology is now under discussion and will be defined in Deliverable 2.2.

cultural object has been rendered without any image), while Figure 2 shows a more detailed cultural object about *Van Gogh* which has been taken from the MultiMatch ontology.

Figure 1. Cultural object created using title plus snippet plus extra information such as the page source (cultural heritage site)

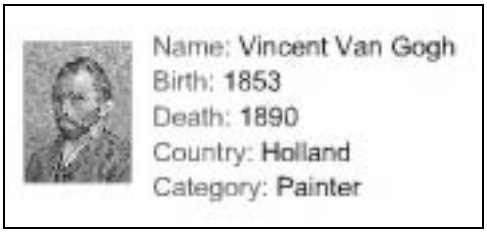

Figure 2. Cultural object created using extracted metadata

MultiMatch will additionally be able to perform simple searches (similar to Google or Yahoo!,) to retrieve standalone web pages and multimedia contents (in this case, things not considered as *cultural objects* in the MultiMatch sense), It is also expected to provide users 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 constitute a further research objective in itself and are not discussed here (see WP6 documentation for details). However, the list of specifications for the first 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 exactly 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, on average, experts 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 will also be considered. In this document we also propose a specialized cultural site search that can be understood as an extension to the initial creators and creations searches.

Search Interaction Levels in MultiMatch

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

- **A default MultiMatch search level** where no assumption is made on the user query, and MultiMatch retrieves information from all indexed materials. In this way, given a general query, MultiMatch will retrieve all the cultural objects, web pages and multimedia contents which best suit the query. Merging, ranking and classification of these results will be also performed by the MultiMatch system.
- **A set of specialized interaction levels**. These allow the user to query MultiMatch specific search services (for instance, video search) and retrieve all the relevant information available via the selected search service. In this way, MultiMatch will include standalone image, video and metadata-based searches (creators and creations), each with its own search fields, display and refinement options. It will also include a set of browsing capabilities to explore MultiMatch contents.

Every search interface will include two different search modes to drive users' queries: these will be available for both search interaction levels:

- **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.

The first MultiMatch prototype is intended to develop the basic search functionality expected for MultiMatch. This implies the focusing of development efforts on the design of robust solutions for multilingual text, image and video retrieval. Using these results as a baseline, the second prototype will address more complex issues (i.e. related with metadata-based retrieval, data mining and information extraction issues) to improve search results and to provide the user with much more sophisticated information rendered as a cultural object.

In this way, the first prototype will provide the system with a default search interaction level oriented to perform text and multimedia searches at the same time, and also two specialized search interaction levels based on standalone image and video retrieval.

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.

The default search level must be understood as a way for the users to express their search needs when they are not looking for information about a specific cultural item (such as a creator or a creation) or when they do not want information in a specific media but want to retrieve all relevant information related to their free text queries.

In this way, this interaction level involves the retrieval of not only cultural objects (i.e. creators and creations) but also web pages, images and videos related with the query. For instance, the query "flowers" should retrieve all cultural objects related with that topic, such as "Van Gogh" or "the sunflowers" artwork, but also a ranked list of web pages, a list of relevant news or a list of images and videos.

For the final prototype it is expected that MultiMatch will retrieve at least the following information items:

- **A ranked list of web pages** rendered as cultural objects in a title plus snippet plus image/video plus metadata.
- **A ranked list of the creators** most closely related to the query and rendered as cultural objects with specific metadata about them.
- **A ranked list of the creations** most closely related to the query and rendered as cultural objects with specific metadata about them.
- **A ranked list of news feeds** related to the query and rendered as cultural objects in a title plus snippet plus image/video plus metadata.

The first MultiMatch prototype will only retrieve web pages and proprietary contents, images and audiovisual material.

Figure 3 illustrates the behaviour for general searches using the example presented above. The user types in *flowers* and retrieves web pages (such as the announcement of a temporary exhibition at the *Tate Modern* entitled *Flowers and questions*), creators (such as *Van Gogh*), creations (such as *Sunflowers* by Van Gogh) and news (such as Van *Gogh exhibit focuses on wheat, harvest themes*). The first prototype will not include the creation/creator/news boxes.

Rankings may consider different sources of information: **query-sensitive** such as content matching (i.e. frequency of *flowers* in documents about *Vincent Van Gogh*, etc.), and **absolute criteria** such as popularity (i.e. most cited authors, links to pages about the author, most queried authors) and freshness (i.e. presence in the most recent news feeds).

Specialized Interaction Level

MultiMatch will also provide the user with a specialized interaction level which will allow him to directly query specific search services such as metadata-based search, image and video search or browsing. 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.

- **Metadata Based Search.** The general idea of metadata based search is that, for a given type of cultural entity (for instance, creators), the whole collection of web pages can be used to mine information about each particular entity that is not present in the individual documents. For instance, the set of all documents talking about Van Gogh can be used to create a profile of the terms most closely associated (i.e. co-occurring more frequently) with Van Gogh. This profile can be subsequently used to compare Van Gogh with other authors. The implication is that for each type of entity considered, MultiMatch must have an index containing such descriptions.
  Metadata-based search will only be implemented in the second MultiMatch prototype; this section is intended to make the transition from the first to the second prototype understandable within the scope of the whole project, and is still subject to changes.

    - **Metadata Based Search: Creators.** The first specialized metadata based search will allow the user to directly search for specific creators indexed in MultiMatch. The user is expected to type in an author's name plus (optionally) additional free text to retrieve relevant information related to the creator (i.e. artist, music composer, writer...). Once the author is set, the system will retrieve different types of information accessible for

the user using navigational items such as tabs.
The main types of information retrieved will be:

1. **Web pages** with title plus snippet description which will be classified in one of these categories: **biographical data** (i.e. web pages with information about the creator's life), **works of art** (i.e. web pages mainly discussing one of the creator's works), **reviews** (i.e. critics, scholar publications, etc), **news** (i.e. RSS feeds related to the creator) and **other** (i.e. other web pages that are related with the creator but not related to one of the previous categories.)
   In each category, entries will be ranked. Figure 4 illustrates this search functionality.
2. **A graphical depiction of the author's network** of relationships to other authors, the different strengths of such relationships, and the concepts/keywords/terms that characterize such relations (see Figure 5.) The idea of the author's networks is to allow the user to discover and explore those authors closely related with the queried author to improve his knowledge about the author's context. In this way, relations may be extracted using common terms or concepts shared by MultiMatch indexed authors.
3. **An author's tag cloud** with those keywords most representative of the author according to the indexed material (see Figure 5).
4. **A list of works of art related to the author**, rendered as cultural objects. Also, in Figure 6 we have added an optional functionality based on *image cloud*, built using the artwork illustrations sized proportionally to their importance. This option has to be discussed in detail and the feasibility of its implementation must be examined.
5. **A list of cultural heritage sites which host content related to the author**, rendered as cultural objects. Also, using the information stored in these cultural objects, in Figure 7 we have added a map with their geographic location (for physical CH sites). Again, this last feature should be considered as optional and MultiMatch partners have to consider the viability of implementing it using third-party commercial APIs (e.g. Google Maps' publicly available API).

In addition, the system will show a cultural object (called "profile"): a brief database-like summary with the name, portrait, description, etc. of the creator (see Figure 4.) MultiMatch will also offer additional views on the retrieved web pages, based on a dynamic clustering of the search results which will consider the terms in the user's query, the information in the author's network and author's tag cloud, the information in the retrieved pages, and the MultiMatch ontology. Figure 8 illustrates one possible way of creating and showing such clustered results.

- **Metadata Based Search: Creations.** The second metadata specialized search service will consist of a search service for works of art (i.e. creations) information. The user is expected to type in an artwork's title plus (optionally) additional free text. Once the artwork is set, the system will retrieve different types of information accessible for the user using navigational items such as tabs.

  The different types of information provided by the system will be:
  1. **Web pages**, with title plus snippet description and classified in one of these categories (see Figure 9): **source**. (i.e. the text in case of a poem, an image if it is a painting, sculpture, building, etc.), **reviews**. (i.e. texts in which the artwork is the subject), **news** (i.e. related RSS feeds) and **other** (i.e. for not categorized results).
  2. **A graphical depiction of the artwork's network of relationships to other works of art**, the different strengths of such relationships, and the concepts/keywords/terms that characterize such relations (see Figure 10). Its goal

is the same as in authors' advanced search but, in this case, relationships between most closely related works of art are shown.

3. **An artwork's tag cloud with those keywords most representative of the artwork** according to the indexed material (see Figure 10).

As in the case of author searches, MultiMatch will offer additional views on the retrieved web pages, based on a dynamic clustering of the search results which will consider the terms in the user's query, the information in the artwork's network and artwork's tag cloud, the information in the retrieved pages, and the MultiMatch ontology.

- **Metadata Based Search: Cultural Heritage Sites.** The final metadata-based advanced search facility is oriented towards retrieving all the MultiMatch information available on a specific *cultural heritage site* such as a museum. The general pattern for searches will be very similar to those described above, and information retrieved will be as follows:

  Main information about the cultural site rendered as a cultural object (cultural heritage site **profile**).

  1. **Web pages** associated to the site rendered as title plus snippet and classified.
  2. **A list of creators most closely associated with the site** and rendered as cultural objects.
  3. **A list of works of art most closely associated with the site** and rendered as cultural objects.
  4. **The network of cultural heritage** sites most closely associated to the site.
  5. **The tag cloud** of the terms which best describe its web contents.

  Figures 11 and 12 show an illustration of these site search facilities.

- **Visual 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 to describe the contents themselves. In this way, searching for multimedia content will thus implicitly often require fusion and multimodal merging of features at various levels of search. The MultiMatch system will provide two 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 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 (i.e. URL of the web pages, video files, image databases, etc.) and its relevance level.
  - **Video Search.** MultiMatch will offer the possibility to search only 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.
  - **Audio Search**. Users will be able to perform audio search to retrieve audio documents and also video documents by way of their speech tracks. For the first prototype, 20 hours of speech recognition transcripts will be generated from audio material in the cultural heritage domain drawn from the four MultiMatch languages Italian, Spanish, Dutch and English. 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 displayed as a hyperlink that allows the user to start playing the audio document 2 to 3 seconds 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 (by nature error ridden) to be more confusing than helpful.
For the second prototype, it is possible that snippets from the (corrected) audio transcript will be displayed.

In the first prototype, speech recognition will use a general-language vocabulary only. In as far as possible, however, the acoustic profiles (i.e. adapted models) specific for the indexed content will be used.

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 feed) 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.** 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 in a directory structured according to the MultiMatch ontology. In this way, it is expected that the user can navigate and explore a cultural heritage knowledge space (based on the MultiMatch ontology), accessing contents rendered as cultural objects. For instance, using this approach, the user could easily find all the authors related to a specific art period).
    2. Users will also be able to explore the author and artwork networks (as described and attached to the specialized search facilities). In this case, accessing the networks will be tied to a prior advanced query made by the user.
    3. Authors and works of art will also be listed alphabetically to allow the user to deal with MultiMatch contents in an authority list fashion.

  When interacting with a retrieval engine users will usually be 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 a topically relevant criteria, focused browsing will bias the selection and presentation of content within user browsing 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 user's 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 their 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 user's current search query, focused crawling aims improve the efficiency of user access to relevant 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 popularity and freshness criteria. In Figure 12, for instance, we have portrayed an entry page which displays the most recent CH-related news and "top ten" lists of the most popular authors, works of art, sites and queries in MultiMatch. 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.

For the first prototype we intend to deal with a general category extraction and classification task for a subset of cultural contents, which means that a simple browsing facility could be developed to ease user's navigation. The second prototype will improve this initial classification with more categories (final MultiMatch ontology) and the browsing strategies will be expanded to cover all the functionalities presented above.

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.
- **Annotation facilities**. This allows the user to annotate the search results with their own comments for future reference. This facility can be considered as optional and will be discussed by MultiMatch partners.
- **Basket facility**. This allows the user to save pointers to relevant images 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 URL 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.

A basic user account and log in facility will be implemented in the MultiMatch first prototype, mainly oriented towards defining the user's language preferences and relating to his search history. The second prototype will implement the whole set of features described.

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 pages which 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. It is also planned to implement image relevance feedback based on image features.

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

- **Filter web results by type of site.** According to the MultiMatch user groups, web pages can be filtered as belonging to a cultural heritage site, educational site, or a tourism site.
- **Filter by language.** Any combination of English, Spanish, Dutch and Italian.
- **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 an specific publishing date range (e.g. from 12-10-2005 to date)
- **Filter by size.** This will retrieve only those resources which fit into an specific size range (e.g. not grater 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 (particularly Google) 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. Figure 15 illustrates this aspect of the MultiMatch interface. This feature is planned for the second prototype.

1.3 Information Stored in MultiMatch Indexes

To support the retrieval functionalities described above, it is necessary to prepare different indexes with different information. Note that the possibility of generating different indexes or treating all the MultiMatch information within one standalone index is currently under discussion between MultiMatch partners, so the following subsections may be modified in the near future. In any event, we consider that specifying such indexes (from an abstract point of view) is useful to offer a minimal description of the indexing structure needed to reach a more comprehensive, understandable definition of the search engine.

Textual Indexes

The retrieval capabilities described above require at least the following indexes:

- **Full Text Resource Index.** The full text associated with each resource (web pages, proprietary contents, images, audio and video) will be indexed with state of the art information retrieval tools and techniques. Depending on how cross-language retrieval is performed (query translation, document translation or a combination of both), this index may be augmented with

translations or related words in other languages. The index might also be augmented with texts in hyperlinks pointing to the resource.

- **Classification.** Each resource will be classified to enable the *classified results* functionality described in Section 2.2. For instance, a web page containing a review of a book must be classified as *artwork review*.
- **Automatically extracted Metadata.** In the example above, besides classifying a page as *artwork review* the system must extract the title of the artwork, the type of site (cultural heritage site, educational site, tourism site), the language of the page, and possibly (unless it is done dynamically) a brief keyword-based summary of the content in the four languages of the prototype.
- **Text mining indexes for authors and works of art.** Each author and artwork identified by the system will be rendered as a complex cultural object which will contain the following information:

    1. **Metadata**. Probably extracted from structured encyclopaedic repositories such as the Wikipedia or proprietary annotated contents. This might include images (portraits in the case of authors, photographs in the case of works of art) and short definitions.
    2. **A signature with the terms/ ontology concepts most closely associated with the author/artwork** Extracted from the web pages where it appears. This signature can be used to build the *tag cloud*.
    3. **An index of the author/artwork related web pages**.

    Using this information, MultiMatch will be able to discover relationships which could be seen as a semantic network of the cultural heritage space. For instance, MultiMatch should discover the following relationships between different authors: a) association scores between two creators, b) terms/things in common between two creators, and; c) creations and keywords most closely associated with a creator. Similar relations apply to the works of art space and to the cultural heritage sites space.

Transcribed Speech Index

Spoken audio content and spoken data streams accompanying visual content will be represented either by the transcribed output of automatic speech recognition systems or using manual transcriptions accompanying the supplied content, (e.g. the closed-caption transcriptions that are manually prepared for hearing-impaired users of video content).

The textual transcriptions will be subjected to the same information retrieval content indexing methods as standard text. In addition, specialised methods to alleviate the negative effects of recognition or manual transcriptions on retrieval will be explored.

Image and Video Index

Video content will be subjected to automated analysis to identify distinct points of shot changes associated with a change in camera position. Segmented shots will then be further processed to automatically identify one or more individual keyframes which are deemed to be most representative of the event depicted in the shot. Keyframes can be used for retrieval or to represent the shot in the presentation of video content in the user interface.

Images and video keyframes will be processed to extract standard low-level visual features such as colour distributions, textures and identifiable edges in the content. Extraction of higher-level visual features, such as objects, may be explored depending on their suitability for the content entered into MultiMatch archives.

Novel methods to be explored for visual indexing within MultiMatch include techniques for cross-model fusion to associate textual annotations with visual features which will enable images to search indexed based on assigned verbal metadata labels. Also, methods to learn associations between low-

level visual features such as colour histograms or textures and higher-level labels such as artists could be considered. Such knowledge would enable the MultiMatch visual analysis component to hypothesise the likely painter of an unknown image entered into the system, possibly cross-modally with associated textual data.

1.4 Interface Functionalities

MultiMatch will have at least two interfaces:

- **A web-based user interface.** The user interface will be highly configurable. 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. For instance, a user of the default search interface (see Figure 3) might choose to discard the "news results" box but add some of the advanced search fields (e.g. target languages, date range) in his own default search interface. There will be default interface configurations for education, CH and tourism search profiles.
- **An API (Application Programming Interface) providing web services for third-party applications and mashups.** All the main search functionalities of MultiMatch will be available as independent web services, which can then be embedded in other web applications, in order to ease the creation of additional CH-related web applications using MultiMatch's capabilities.

2 Functional Specifications for the First Prototype

On the basis of the main features described above, in this section we describe the functional specifications required to develop the first prototype for the MultiMatch system.

The main goals of this first prototype are focused on providing the MultiMatch system with a basic set of search features. Search results will be drawn from the specific cultural heritage data indexed for the project.

The following sections are divided as follows. First, we will introduce the crawling and indexing functionalities required for data acquisition, structuring and storing. Secondly, we will focus on the main search functionalities that the first prototype will offer. We will then present the retrieval functionalities expected and, finally, we will list the interface functionalities for MultiMatch.

2.1 Crawling and Indexing Functionalities

The first MultiMatch prototype will face the crawling and indexing tasks using the following document sources:

- **200 cultural heritage institutions and their web sites** obtained from a handcrafted white list crawl (focused on *museums*). It is critical to crawl and index *10.000 web pages from the United Kingdom and Ireland* at least. Also, it will be desirable to reach 10.000 web pages from each of the following countries: Spain, Italy and Holland.
- **Wikipedia contents and associated metadata.** For the MultiMatch project, we will only consider those Wikipedia items related with Wikipedia **artist** and **museums** categories. However, we should note that Wikipedia categories are noisy and include mistakes due to human factors; therefore, the site's understanding on cultural heritage must be taken in the very broadest sense. Contents coverage for different languages will be different but it is critical to index at least *10.000 Wikipedia items in four languages*. As an optional task, Wikipedia images should be indexed.
- **Proprietary contents provided by Alinari, Sound and Vision and Biblioteca Virtual Miguel de Cervantes partners**. Text, image and video contents from cultural heritage partners will be also indexed (at least a part of the total planned shared contents).

Coverage for all the languages supported by MultiMatch (i.e. English, Spanish, Italian and Dutch) will be provided by the first prototype and also the following file types:

- **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, MPEG4, MPEG 7.** For video documents.

The content indexing protocols applied in this first prototype will be the following:

- **For general text documents and transcribed spoken data:** All free text and uncontrolled language metadata contents should be pre-processed. Language dependent techniques will be applied for stop words removal, stemming and synonym matching.
- **For cultural heritage web sites (museum sites):**
  1. A manual CIDOC-compliant annotation on a subset of the 200 cultural heritage institutions will be made.
  2. Document indexing using text indexing techniques.
- **For Wikipedia contents:**
  1. Target Wikipedia contents will be indexed using text indexing techniques.
  2. They also will be parsed to generate metadata. We have to remark that the Wikipedia mark-up provides very general information about each item (mainly about the item within

Wikipedia structure and its main features such as title, contributor, category, etc.). In this way, for this first prototype it is intended only to index this type of information.

- **For spoken audio indexing:**
  1. 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.
  2. Audio will be transcribed (i.e. both audio and sound tracks of video) into an XML format that includes the words spoken and a time index of when they are spoken. These time-aligned transcripts can then be subjected to pre-processing techniques similar to those used for text. For the first prototype it is planned to include around 20 hours of transcribed audio in the MultiMatch collection.

- **For image indexing:** All the images crawled, extracted from image libraries and also video keyframes will be indexed using standard low-level visual feature extraction (i.e. colour, histogram, texture) and encoded with agreed metadata markup. Textual descriptions of such images will be also indexed to allow text retrieval.

- **For video indexing:** All video in MultiMatch will be indexed using shot boundary and automatic keyframe detection.

Functional specification tables with reference numbers can be found in the appendix of this document.

2.2 Search Functionalities

The first MultiMatch prototype should have the following searching functionalities:

- **Levels of search interaction:**

  1. **Default level.** MultiMatch will provide users with a default combined search on text, image and audio-video search services. The first prototype will retrieve for a single query all the text documents, still images and audio-video, which will be presented separately to the user. It is not planned to provide search results with full embedded metadata (which means that search results will not be rendered as cultural objects) but, when possible, the first prototype will make use of CIDOC manually annotated markup to improve the information stored in search results. Issues related with metadata-based search (such as creators and creations) will be addressed in the second prototype.
  2. **Specialized level.** As the default level will make use of a simple approach to standalone text, image and audio-video retrieval, it is also planned to provide MultiMatch with these search facilities as standalone search services. Image and video retrieval will also provide the user with text query-based and image query-based retrieval facilities, although the latter will only be provided for contents indexed within the MultiMatch collection. Finally, the initial manually annotated categories will be used by the first prototype to provide a simple browsing functionality of MultiMatch contents.

  The first prototype will provide the following search features for both searching levels:
  - Relevance feedback using search results.
  - Image relevance feedback for still image and keyframe search results.
  - Boolean search. Typical of these boolean operators would be “AND”/ “OR” searches.
  - Visual search using low-level features (e.g. colour histograms). The first prototype will only apply this search functionality within the indexed document collection, not for external images.

- **Search modes:**
    1. **Default mode.** MultiMatch will implement a default search mode that will provide the user with the capacity to enter a free text query (as a basic search function). MultiMatch will perform monolingual retrieval using the user's native language as a default mode.
    2. **Advanced mode.** MultiMatch will allow the users to customize their searches to retrieve information in different languages. At this point, it is planned that the first prototype will perform translingual searches in the four selected project languages. The results in different languages will be presented to the user as separate document sets. Filters depend on the information extracted from indexed data; for this first system it is planned to include filtering only for file type and for file size.

Multilingualism will be enabled at this first stage by translating the query from one language to another using a simple combination of machine translation and, optionally, also commercial multilingual/bilingual dictionaries.

2.3 Retrieval Functionalities

The first MultiMatch prototype will perform the retrieval process using the following document sources:

- **For default search:** MultiMatch will retrieve information from the following sources using its corresponding search services:

    1. Cultural heritage web sites indexed (text search).
    2. Wikipedia items indexed (text search).
    3. Biblioteca Virtual Miguel de Cervantes contents indexed (text search).
    4. Still images indexed from web pages and/or Wikipedia (image search)
    5. Alinari still images repository (image search)
    6. Sound and Vision videos (video and audio search)

- **For specialized search:** MultiMatch will retrieve information from the following sources:

    1. **For image search:** MultiMatch will retrieve images from the Alinari still images repository, images indexed from web pages and/or Wikipedia.
    2. **For audio-video search:** MultiMatch will retrieve videos from the Sound and Vision repository.

The first prototype will also implement the following features for default and specialized searches:

- **For default search**: MultiMatch will provide the user with the following lists of items:

    1. A list of documents, ranked according to some relevance criteria, and retrieved from the set of text sources indexed in MultiMatch. When possible, the retrieved information will be combined with metadata to provide the user with more comprehensive information about the documents. Text documents will be summarized using query-biased snippet generation techniques which display the text in its original language or a summary translation at least in English. Search results in different languages will not be combined in the first prototype. They will be retrieved and ranked as separate document sets.
    2. A list of still images, ranked according to some relevance criteria, and retrieved from the set of image sources indexed in MultiMatch. When possible, the retrieved information will be combined with metadata to provide the user with more comprehensive information about the images.

3. A list of videos, ranked according to some relevance criteria, and retrieved from the set of video sources indexed in MultiMatch. When possible, the retrieved information will be combined with metadata to provide the user with more comprehensive information about the videos.

- **For specialized search:** MultiMatch will allow the user to perform standalone image and video searches providing a list of the following items:

    1. The image search service will provide the user with a image list, ranked according to some relevance criteria, which will give access to image thumbnails, original images, sources of the images (i.e. web pages where they are embedded, video files, image databases).
    2. The video search service will provide the user with a video list, ranked according to some relevance criteria, and retrieved using text retrieval techniques based on spoken audio transcriptions.

2.4 Interface Functionalities

The MultiMatch first prototype will provide the following interfaces:

- **Graphical Interface.** MultiMatch will implement an initial web based client interface to support the following activities: a default search, a specialized search for images and video, and a browsing facility.
- **Application Programming Interface.** MultiMatch will provide third-party users with an initial API to access default and specialized search level services. This API will be also used by the MultiMatch web interface to perform user searches and to retrieve results.

The main functionalities for the graphical interfaces will be the following:

- **User access privileges.** The user will use the MultiMatch system as anonymous or registered user. Registration implies:

    1. The user will receive a user account protected by username and password.
    2. The user account will give access to the user search history and language preferences.
    3. The use of client stored data will enable the configuration of the MultiMatch interface according to the user's preferred languages.

- **Browsing facilities.** The user will be able to navigate and visualize the following MultiMatch contents:

    1. The words most frequently appearing in the collection (e.g. tag cloud facility).
    2. A collection overview (e.g. a display or collage of items randomly chosen from the collection)
    3. The user will be able to browse the collection, when possible, by categories, based on pre-existing metadata information (this facility will be only available for specific contents indexed in the first prototype).

- **Language facilities.** The user will have access to the following language facilities at query time:

    1. The user will be able to retrieve documents in the language of the query as well as others (CLIR).
    2. The user's query will be translated into the other languages of the project and suitable matches will be identified.

3. The user will be able to view possible translations (in the case of an ambiguous term) and select the preferred version.
4. The user will be able to view results in languages besides his/her preferred language. The presentation mode can be based on preference (e.g. the user can elect to display a summary of the document both in the original language and a translated version of this summary, if necessary.)

- **Presentation and organization facilities.** The following shows the main presentation functionalities related with the user interface:

1. Results will be displayed in an arrangement that is deemed to be preferred by users (e.g., in a grid, slideshow format, etc.)
2. The user will be able to control or re-arrange the results display
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)
4. The user will be able to sort or re-organise the results in a simple way (e.g. by size/file type)
5. The user will be able to make an indication of an item's relevance using positive or negative relevance feedback.
6. 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.
7. The user will be able to group or cluster the results based on simple criteria (e.g. colour)
8. 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.
9. The registered user will be able to perform manual annotation of results.

3 Conclusions

This document has introduced the set of functional specifications needed to implement the first MultiMatch prototype using a general view of the final system as a baseline. First we have described an overview of what the MultiMatch system is expected to do in the future, although this description should be taken as a provisional list of functionalities (subject to change in the future) needed for a better understanding of the first prototype functionalities, which are introduced in the third section.

First prototype functionalities are divided in four subsections which cover all the issues related with crawling and indexing, search, retrieval and interface functionalities presented as tables in the Annex of this document.

The main goal of this first prototype is to implement the base MultiMatch system, which will be used and extended in the second prototype. This means that the first prototype will include a set of basic functionalities oriented to retrieve simple text, image and video resources, using a wide set of pre-selected cultural heritage sources instead of the whole web. This first prototype will allow the user to perform searches using a default search interaction level (to retrieve information in any type of media) and also to perform specialized searches to retrieve information in a specific media (images or video). Finally, it is also planned to include a browsing functionality to allow the free navigation of indexed MultiMatch contents.

4 References

[1] MultiMatch Project. Deliverable 1.2: Analysis of User Requirements.

5 Acknowledgements

This deliverable is the result of intense and in-depth discussions with colleagues of the MultiMatch consortium. The authors gratefully acknowledge the suggestions and input received from all partners and, in particular, from the internal reviewers of Alinari and WIND.

Annexes

First Prototype Specification Tables

Crawling and Indexing Functionalities

Functional Specification 1.1. Languages Supported

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

Functional Specification 1.2. Document Types

Reference No.	Document Type	Document Format(s)
1.2.1	Audio	MIME types audio/mpeg, audio/x-wav
1.2.2	Still Image	bmp, jpg, gif, tiff, png
1.2.3	Text	MIME types: text/plain, text/html, text/xml
1.2.4	Video	mpeg1, mpeg4, mpeg7

Functional Specification 1.3. Document Sources

Reference No.	Document Source	Description
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
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
1.3.3	Alinari collection	Image documents with metadata.
1.3.4	Biblioteca Virtual Miguel de Cervantes	Text documents with metadata.
1.3.5	Sound and Vision videos	Audiovisual contents with metadata.

Functional Specification 1.4. Indexing protocols

Reference No.	Document Set	Description
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.
1.4.2	Subset of indexed documents from 200 CH institutions	1.4.2.1. Document indexing using text indexing techniques
		1.4.2.2. A manual CIDOC-compliant annotation on a subset of the 200 cultural heritage institutions will be made.
1.4.3	Wikipedia contents	1.4.3.1. Document indexing using text indexing techniques.
		1.4.3.2. Items parsing to generate simple metadata.
1.4.4.	Biblioteca Virtual Miguel de Cervantes	1.4.4.1. Document indexing using text indexing techniques.
		1.4.4.2. Items parsing to generate metadata.
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.
1.4.5	Allinari Collection	1.4.5.1 Image indexing using image indexing techniques.
		1.4.5.2 Metadata contents indexing using text indexing techniques.
		1.4.5.3. Image parsing to generate metadata.
1.4.6	Sound and Vision videos	1.4.6.1. Video indexing using shot boundary and automatic keyframe detection.
		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 still image indexing.

Search Functionalities

Functional Specification 2.1. Search levels

Reference No.	Search Level	Description
2.1.1.	Default	2.1.1.1 MultiMatch will provide a combined search facility for free text, image and audiovisual retrieval.
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.

Functional Specification 2.2. Search modes

Reference No.	Search Mode	Description
2.2.1.	Default	2.2.1.1 MultiMatch will perform monolingual retrieval using the user's native language as a default mode.
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

Functional Specification 2.3. Search facilities

Reference No.	Search Facility	Description
2.3.1.	Relevance feedback	2.3.1.1. Text relevance feedback
		2.3.1.2. Image relevance feedback
2.3.2.	Boolean search	2.3.2.1 Typical of these boolean operators would be “AND”/ “OR” searches
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).

Retrieval Functionalities

Functional Specification 3.1. Retrieved sources

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

Functional Specification 3.2. Retrieved results

Reference No.	Search Level	Description
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.
		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.
		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.
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
		3.2.2.2. The video search service will provide the user with a video list, ranked according to some relevance criteria,

Interface Functionalities

Functional Specification 4.1. Interfaces provided

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

Functional Specification 4.2. Accessing functionalities

Reference No.	Accessing Level	Description
4.2.1.	Anonymous	Perform searches on any of the defined search levels but without advanced features such as search history.
4.2.2.	Registered	4.2.2.1. Search history
		4.2.2.2. Manual setting of language preferences.
		4.2.2.3. Automatic setting of user's preferred languages

Functional Specification 4.3. Browsing functionalities

Reference No.	Functionality
4.3.1.	Visualization of the words most frequently appearing in the collection (e.g. tag cloud facility).
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)
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).

Functional Specification 4.4. Language facilities

Reference No.	Functionality
4.4.1.	The user will be able to retrieve documents in the language of the query as well as others (CLIR)
4.4.2.	The user's query will be translated into the other languages of the project and suitable matches will be identified
4.4.3.	The user will be able to view possible translations (in the case of an ambiguous term) and select the preferred version
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.)

Functional Specification 4.5. Presentation and organization facilities

Reference No.	Functionality
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.)
4.5.2.	The user will be able to control or re-arrange the results display
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)
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)
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.
4.5.6.	The user will be able to group or cluster the results based on simple criteria (e.g. colour)
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.
4.5.8.	The registered user will be able to perform manual annotation of results.

Images

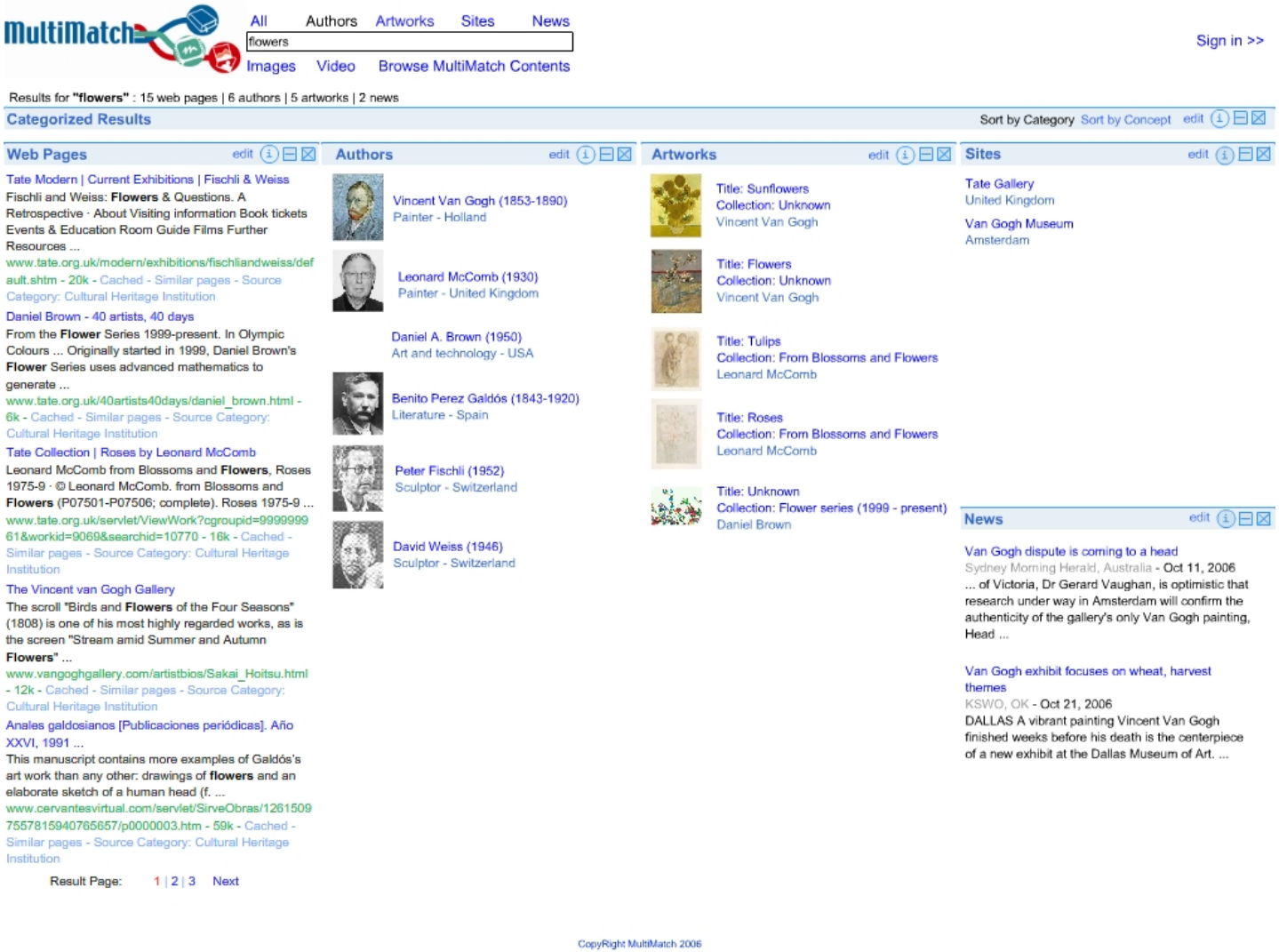

Figure 3: Default search functionality.

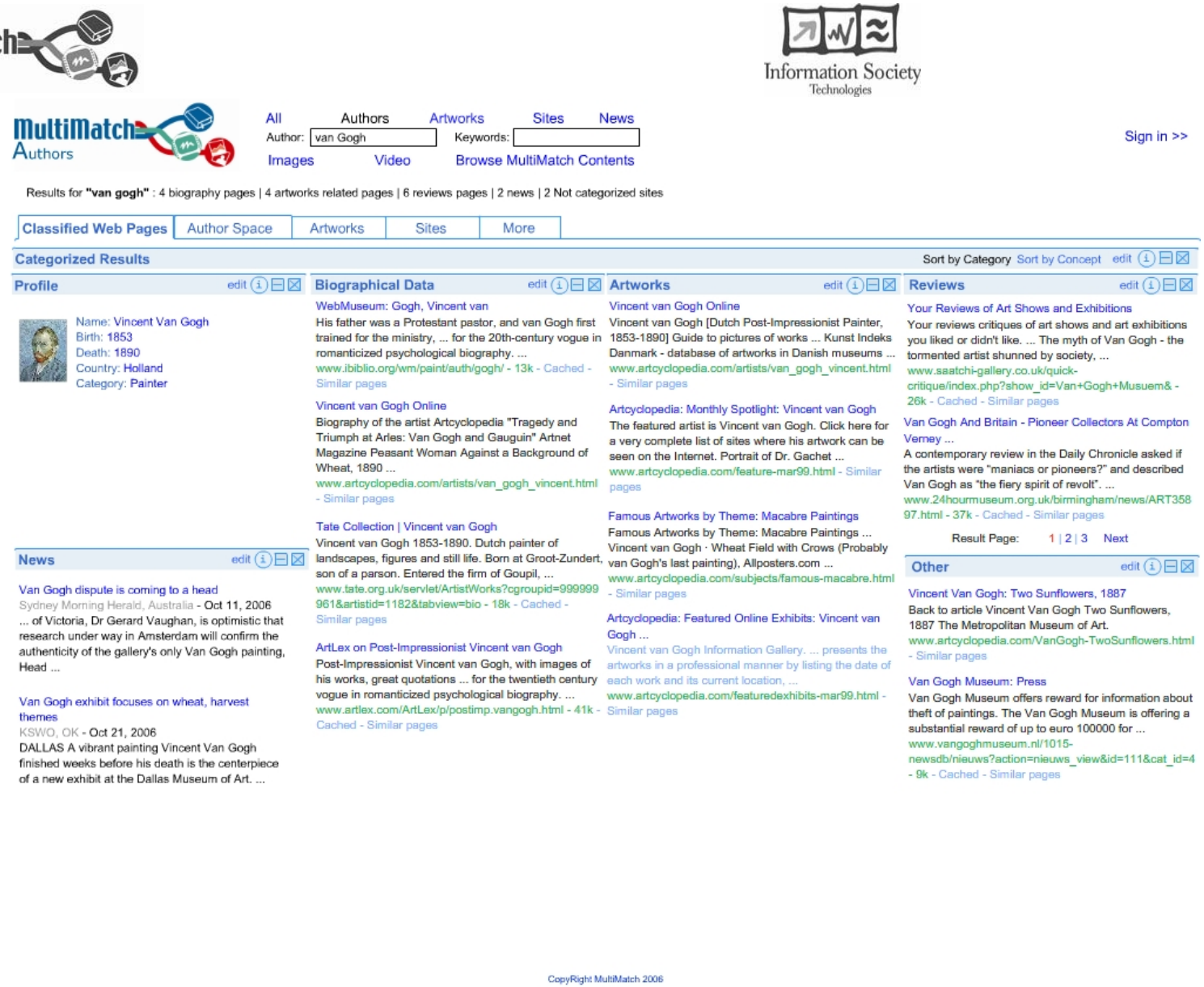

Figure 4: Author Search (i): Classified Web Page results

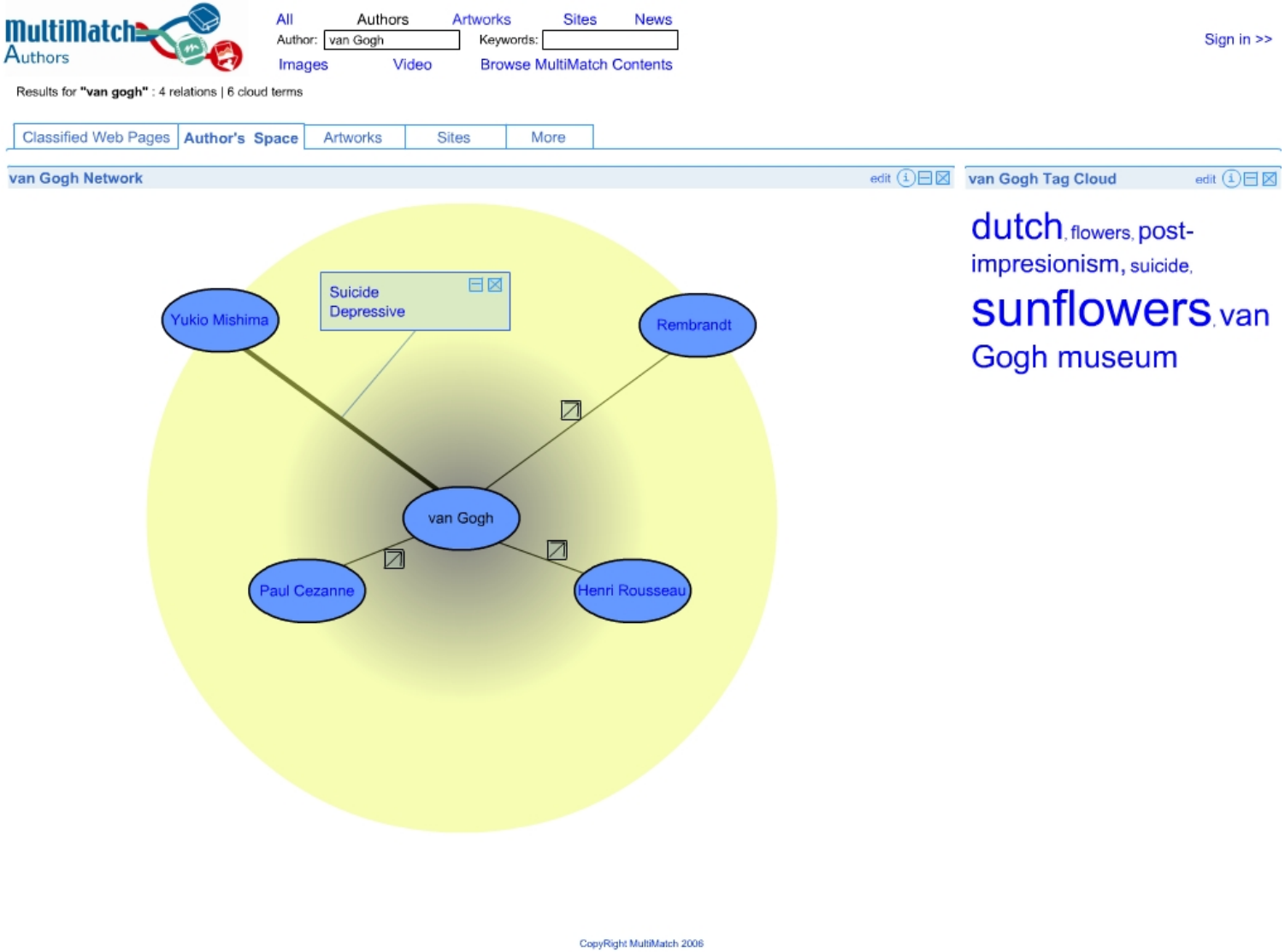

Figure 5: Author Search (ii): Author's network and tag cloud.

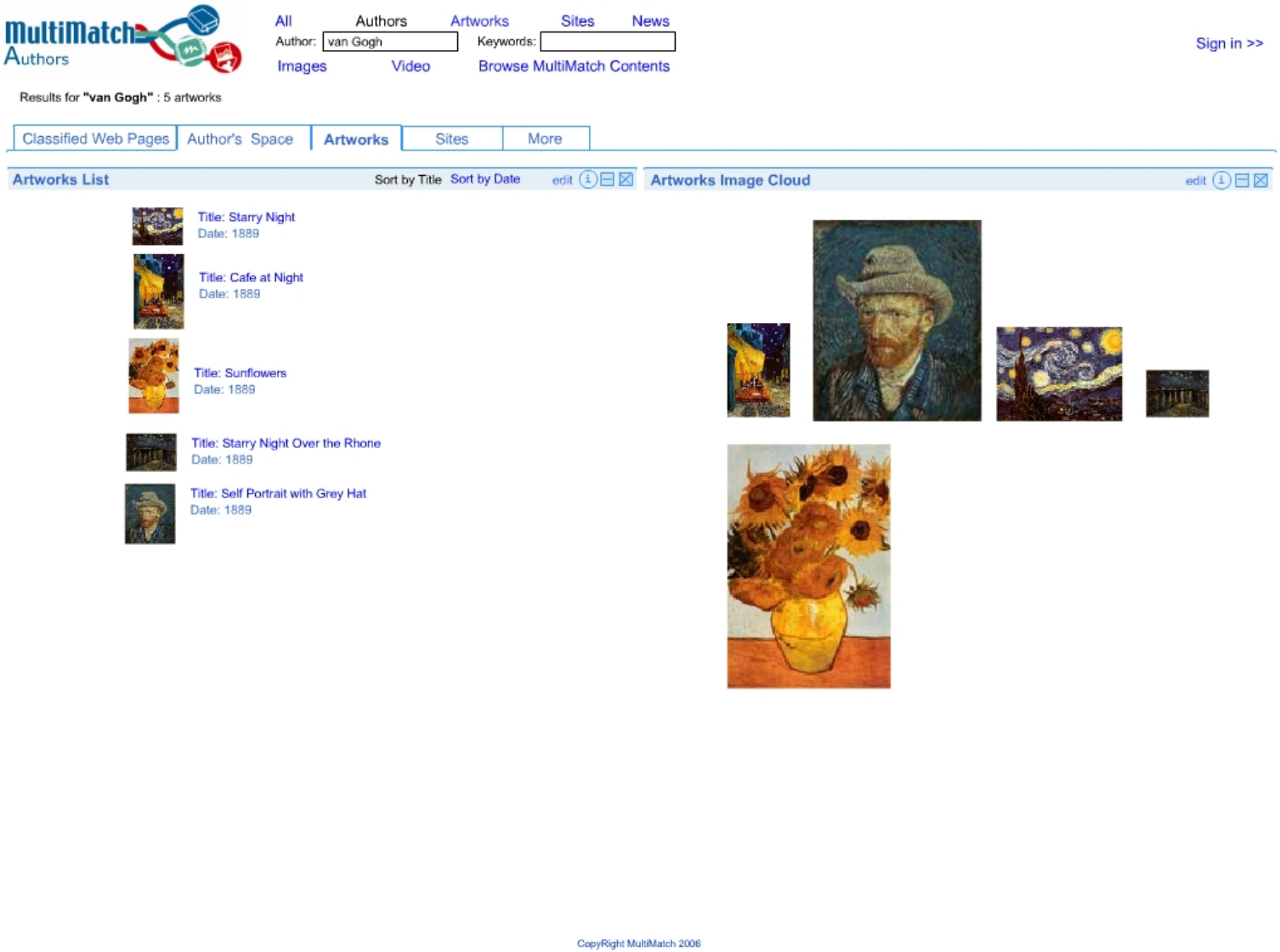

Figure 6: Author Search: Works of art associated with an author

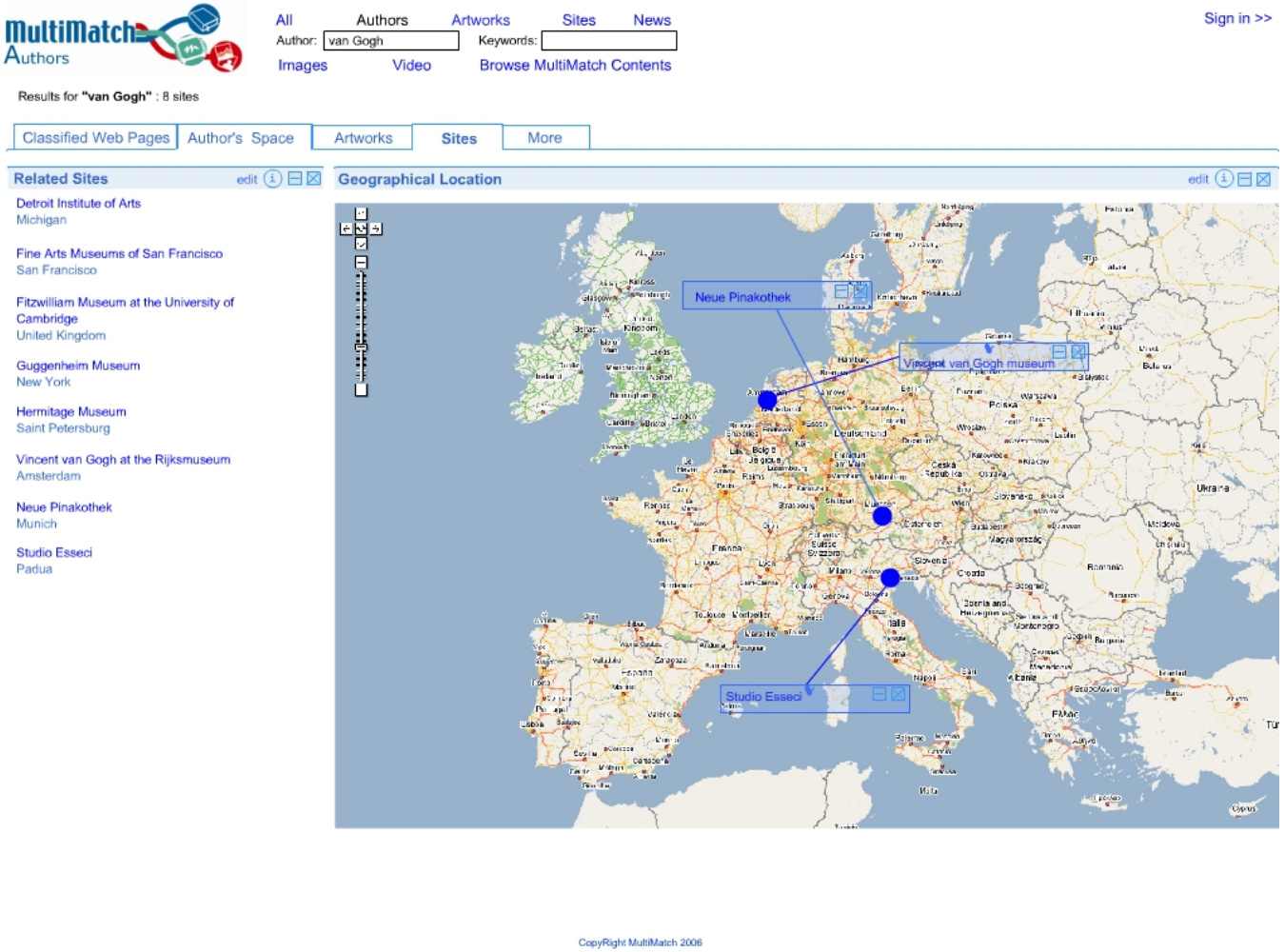

Figure 7. Author search: Related CH sites

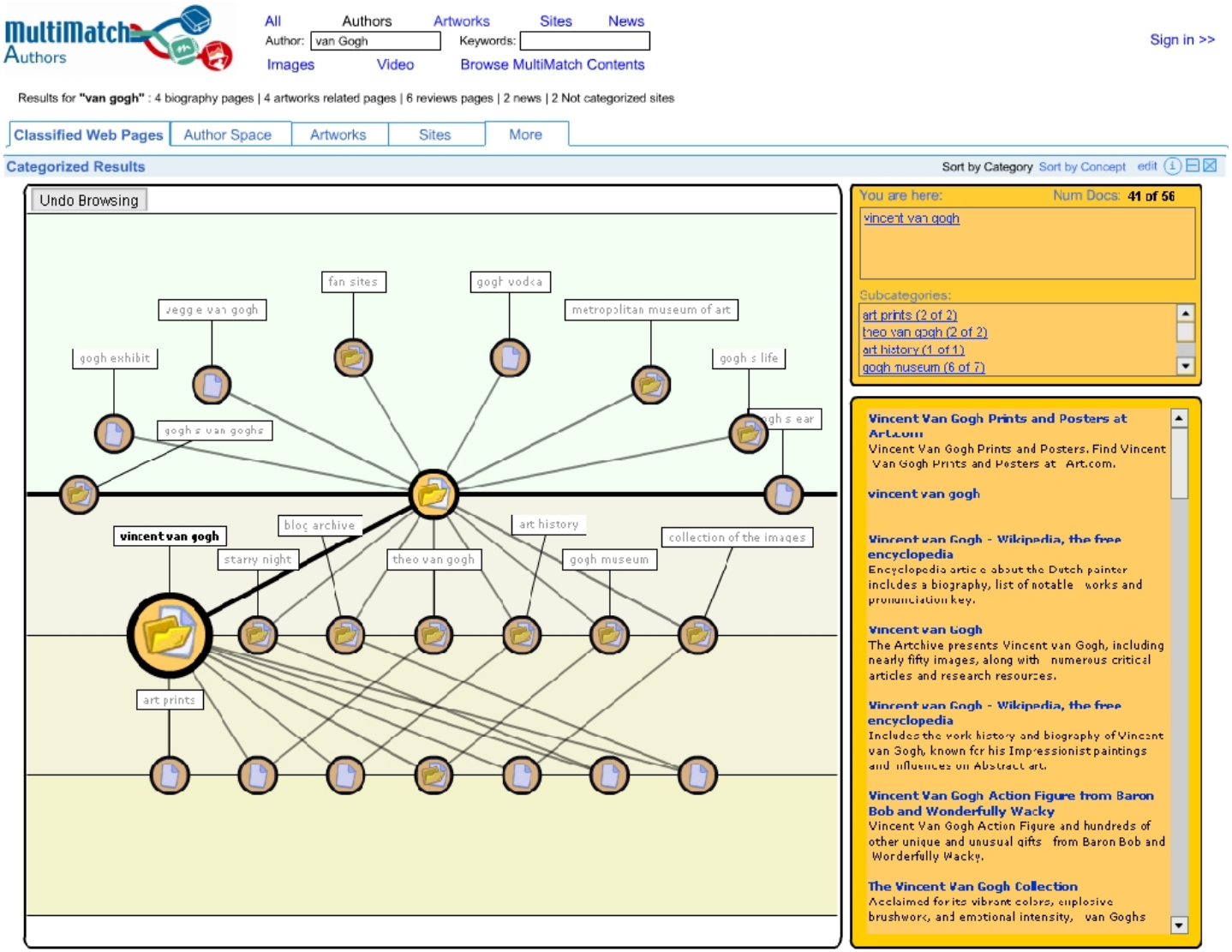

Figure 8: Author search (iii): clustered results

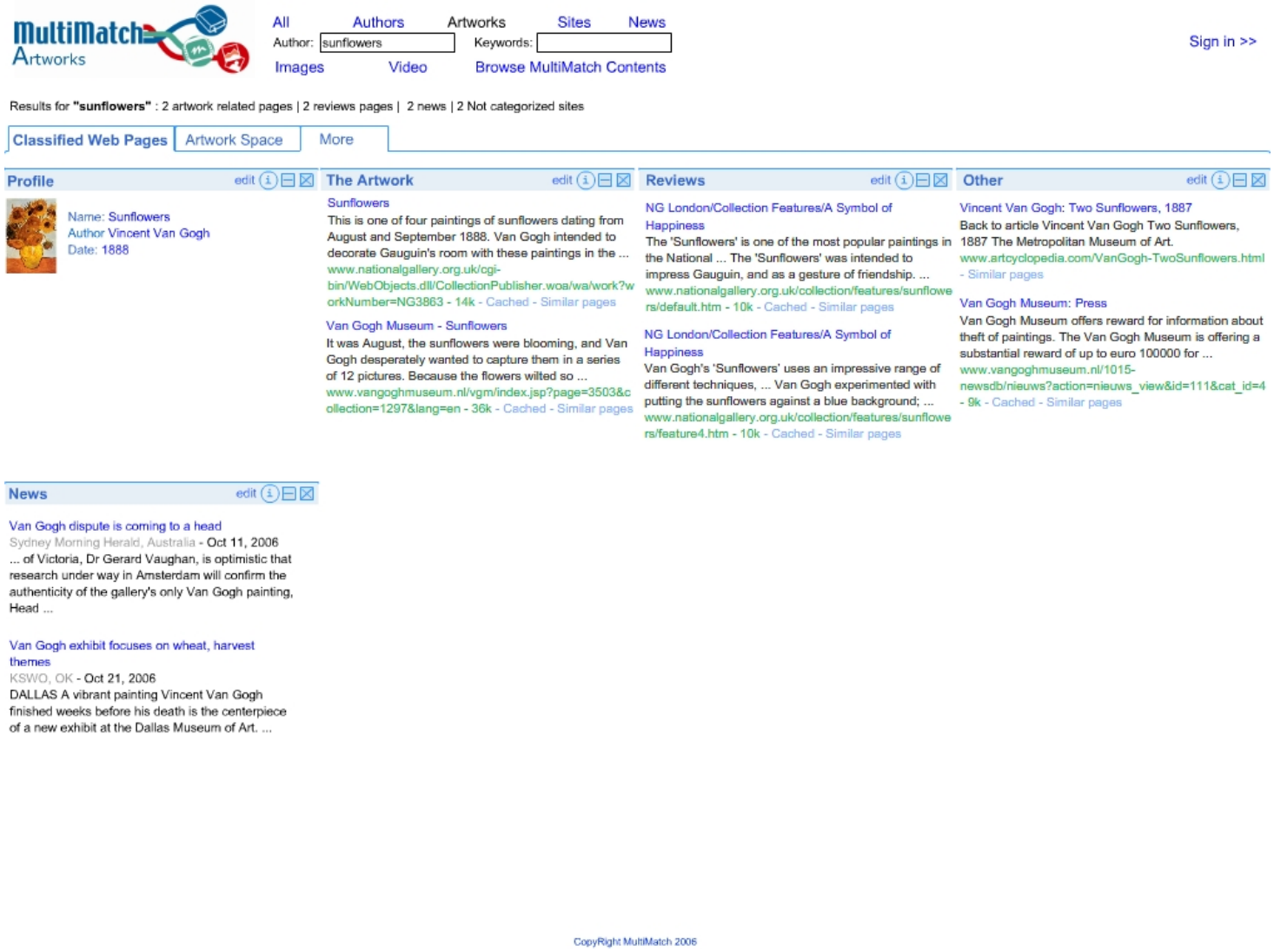

Figure 9: Artwork search (i): web page results

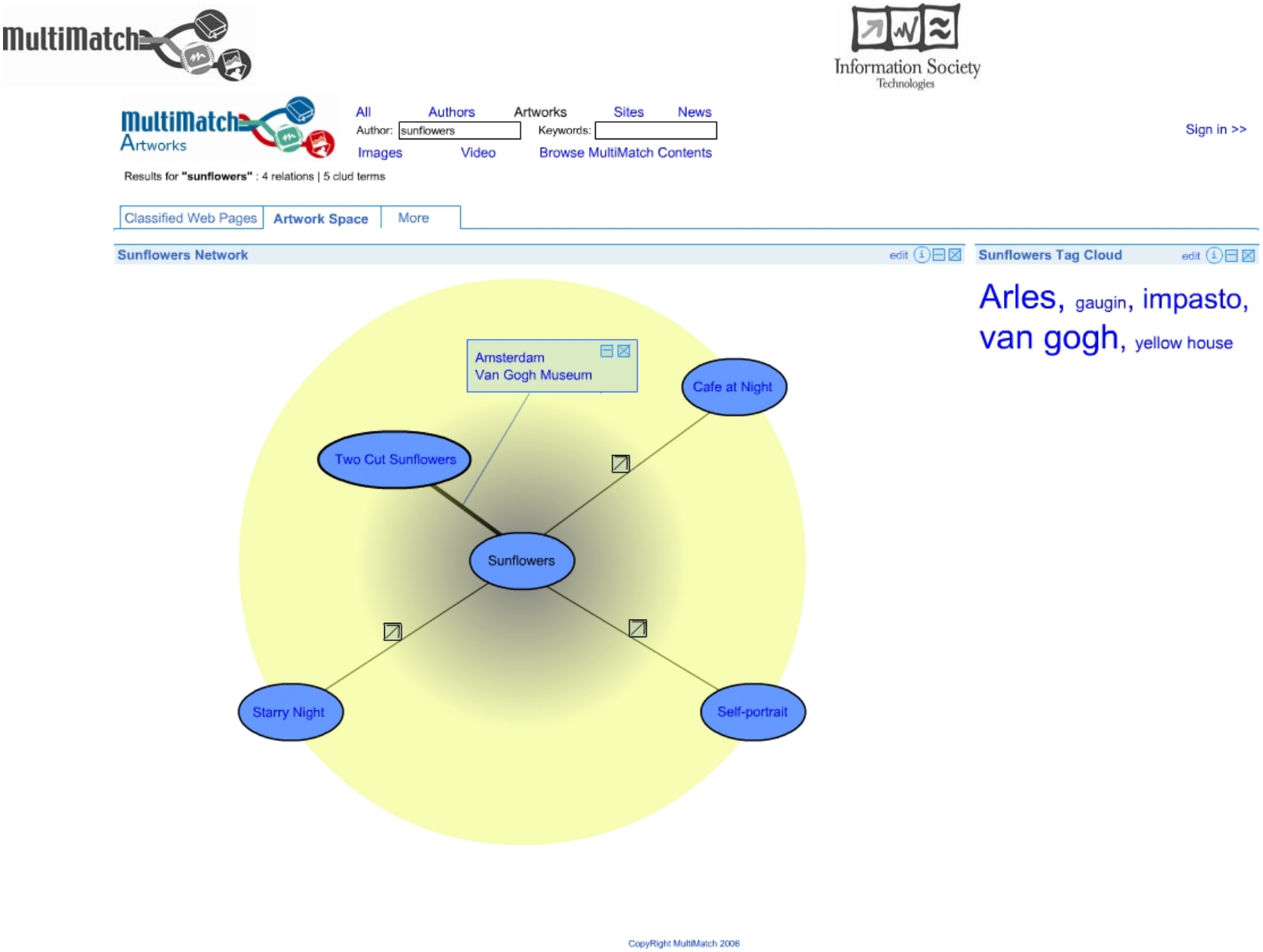

Figure 10: Artwork search (ii): artwork space and artwork tag cloud

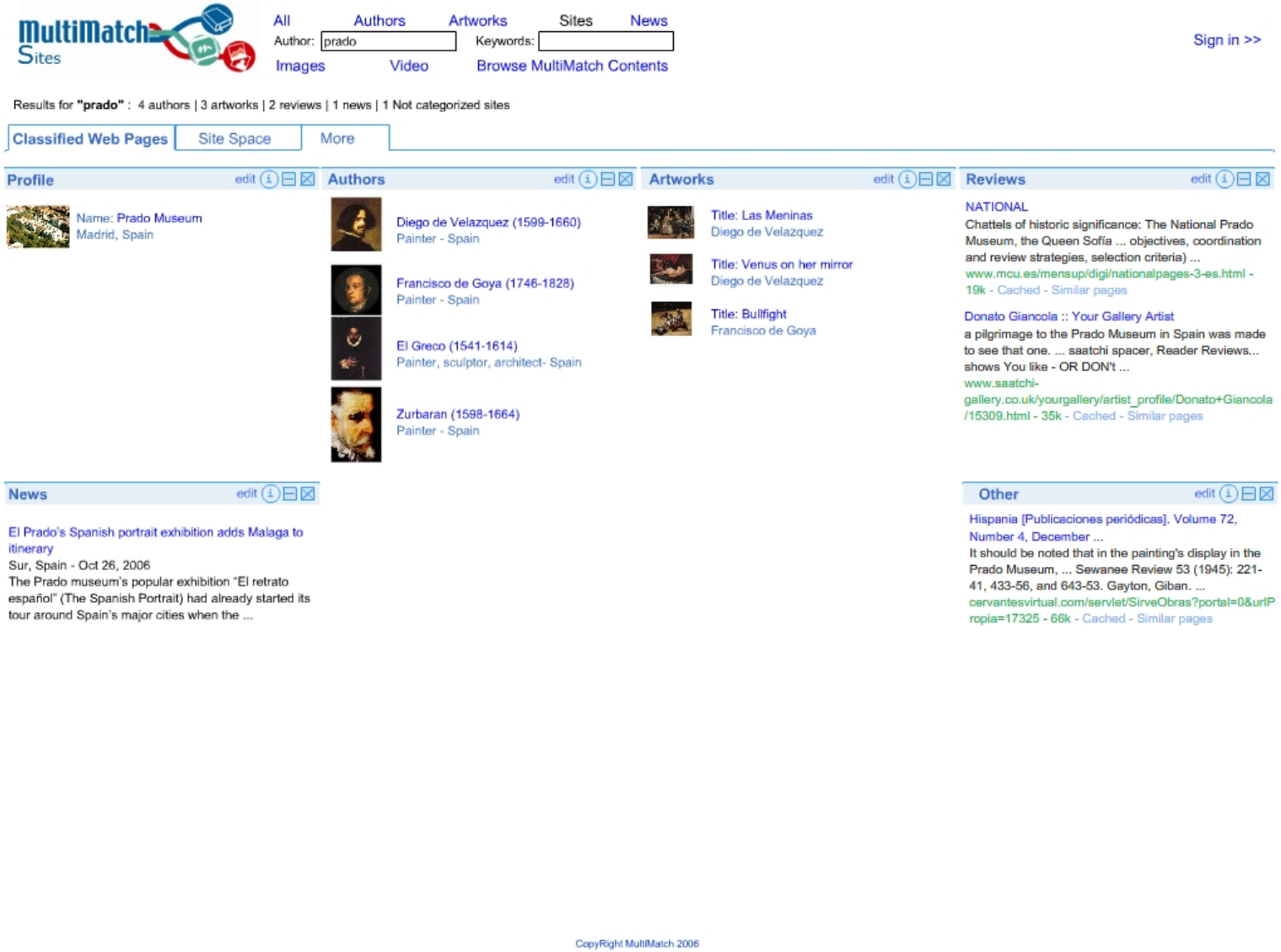

Figure 11: Site search (i)

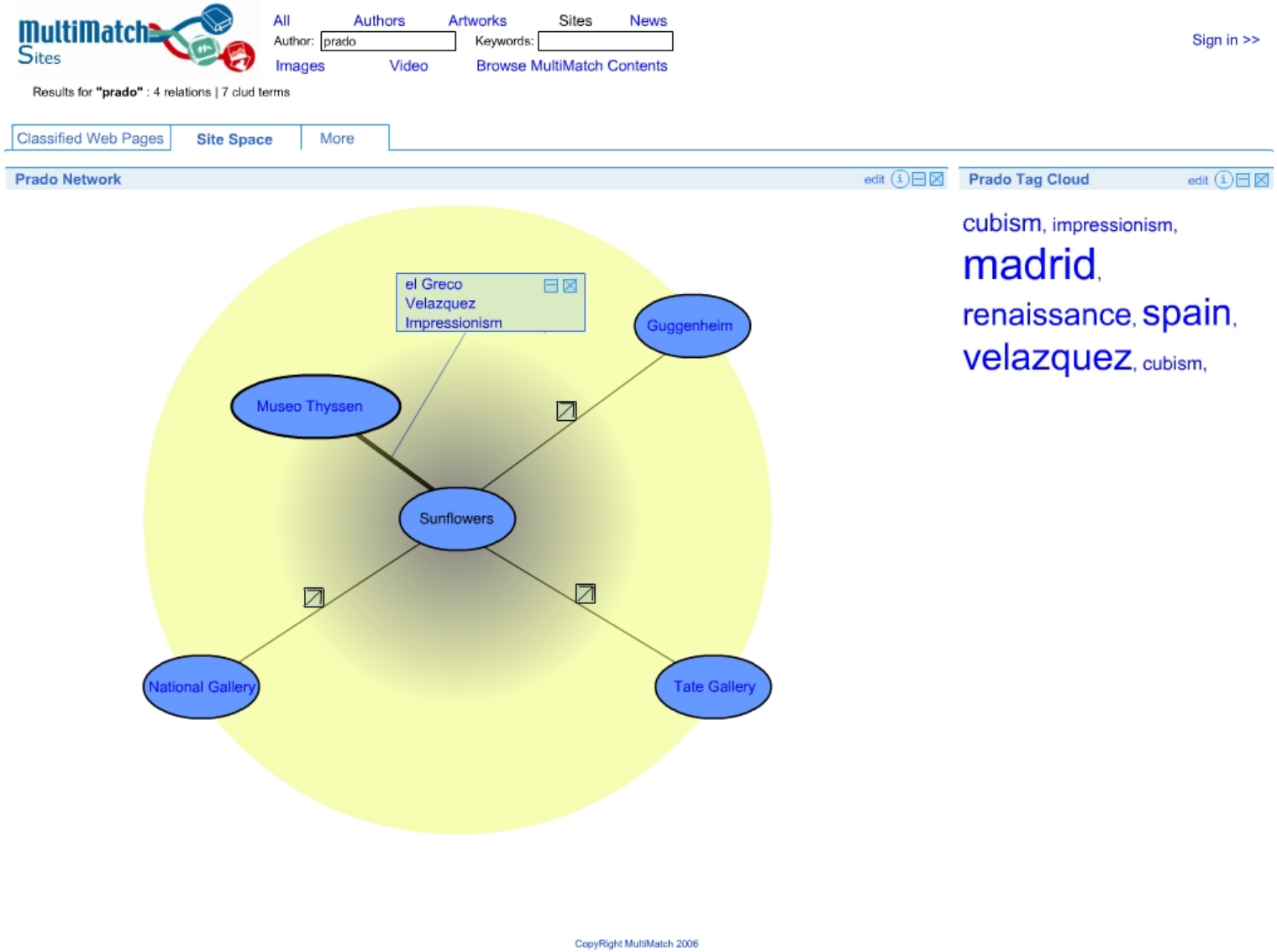

Figure 12: Site search (ii)

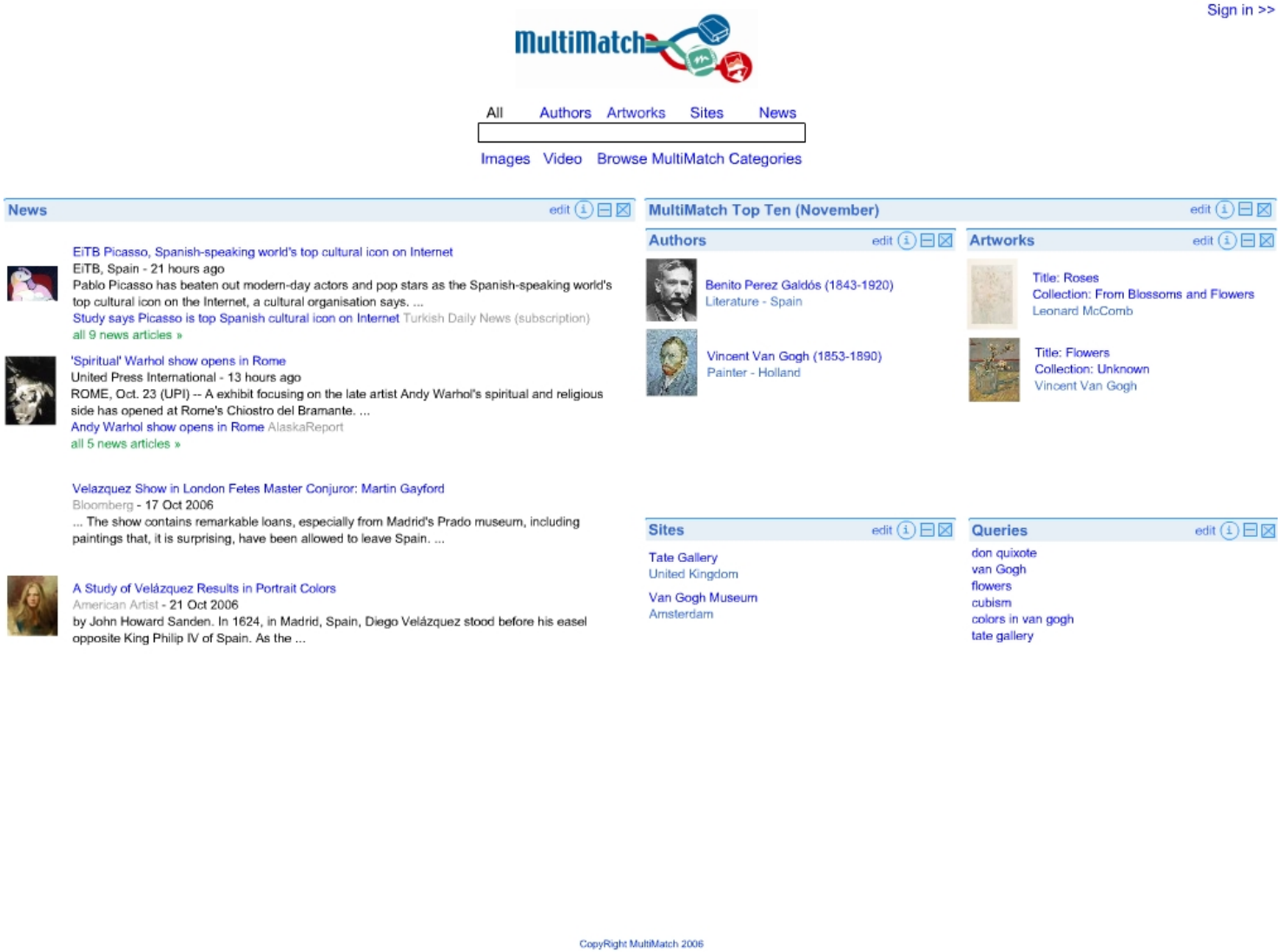

Figure 13: MultiMatch entry page with pointers to explore contents: news and popularity rankings

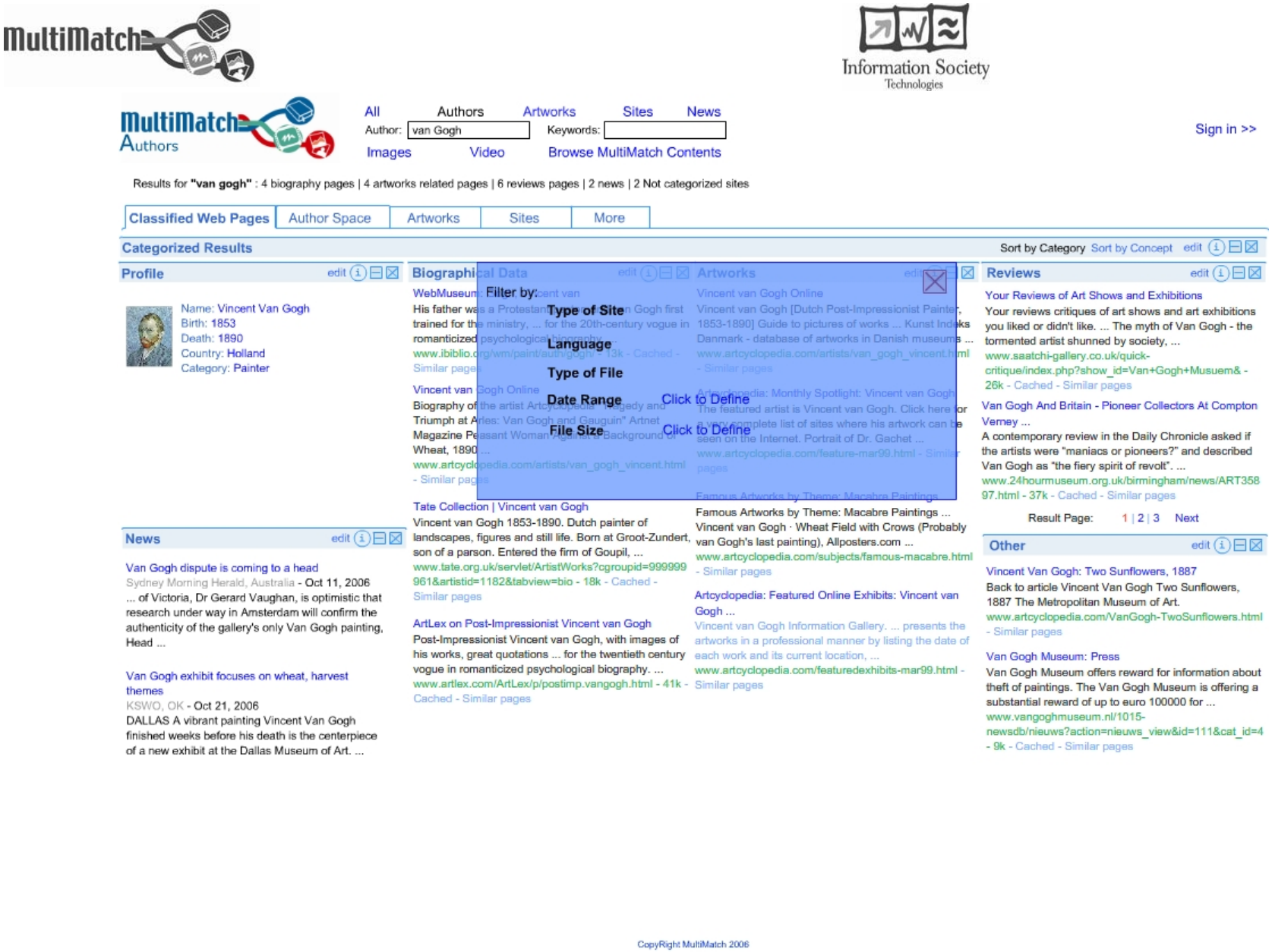

Figure 14: Search preferences

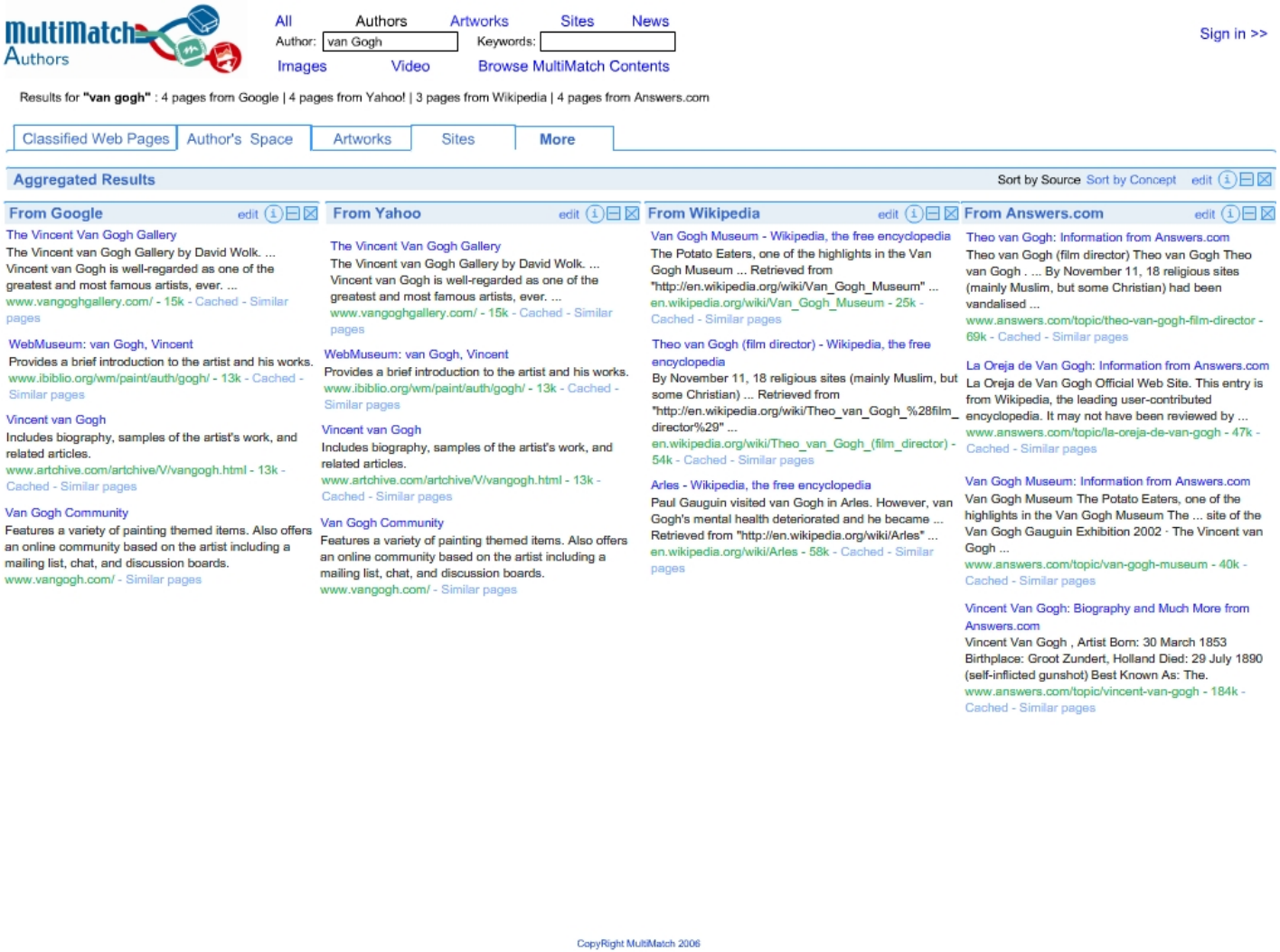

Figure 15: Meta-search results