IFLA WLIC 2013
SINGAPORE

**Submitted on:** May 31, 2013

# The image is the record – Similarity-based image search for visual materials in digitized library collections: The approach of the Bavarian State Library

**Klaus Ceynowa**

Deputy Director General, Bavarian State Library, Munich, Germany.
E-mail address: ceynowa@bsb-muenchen.de



**Abstract:**

*With 96,000 manuscripts, 20,000 incunabula, and more than 160,000 rare books of the 16th century, the Bavarian State Library is one of the major institutions for written cultural heritage worldwide. Huge parts of these unique collections are digitized and freely available for research and study. They contain vast amounts of images of immense value to researchers and students in arts and humanities. To make these treasures more visible, the Bavarian State Library, in cooperation with the Fraunhofer Heinrich-Hertz-Institute, has developed a similarity based image retrieval tool that operates directly on the level of the digital content. The colours and edges of the images themselves are used as metadata to explore related and similar images in a rich content-repository of rare books and manuscripts.*

**Keywords:** Similarity Search, Image Search, Visual Library Materials, Creative Search Capabilities, Automatic Metadata Generation

The Bavarian State Library in Munich (www.bsb-muenchen.de), founded in 1558, is the central state and repository library of Bavaria and one of the world's most important universal and research libraries. Its collections currently comprise almost 10 million books and 59,700 current journal subscriptions, mostly in electronic form, as well as 1.2 million e-books. With 96,000 manuscripts it ranks among the five largest manuscript libraries of the world, its collection of 20,000 incunabula is the largest worldwide, and with around 160,000 printed works of the 16th century, the Bavarian State Library is Germany's leading library regarding this time segment.

The Munich Digitization Center (MDZ) was founded at the Bavarian State Library already in 1997 (www.digital-collections.de) with the support of the German Research Foundation as the central academic funding agency in Germany. Today the Munich Digitization Center is a national competence centre for innovative digitisation technologies and Germany's leading

institution of mass digitisation of written cultural material, among other things through the consistent use of scan robotics and 3D scanning technology. Currently the Bavarian State Library already offers around 940,000 digitised books from its collections for free-of-charge use, among them many unique works. This is the largest digital data collection held by any German cultural institution. At the start of 2007 the Bavarian State Library was the first continental European library to enter into a public-private partnership with Google, initiating the digitisation of its complete copyright-free holdings from the 17th to the 19th century, comprising a total of over 1 million volumes.

In particular the manuscripts, incunabula and early prints of the library contain a broad variety of visual materials in the form of images, illuminations, xylographs, graphics, drawings and emblems. For a large part of the humanities and cultural studies disciplines, these images are crucial, frequently even more important than the texts themselves, in which they are embedded. Usually the individual image materials are not indexed and catalogued item by item, however. For scholarly research they consequently remain largely "hidden" even when the works in which they are contained are freely available in the Internet. In the light of the large number of such visual materials – in fact we are speaking about mass data – their intellectual indexing and cataloguing will not be possible in the future either.

To make these valuable visual collections searchable, retrievable and usable in the digital space nonetheless, the Bavarian State Library, together with its technology partner, the Fraunhofer Heinrich-Hertz Institute in Berlin (www.hhi.fraunhofer.de), has developed further a retrieval software that had originally been developed by the HHI for the detection of copy-right infringements of images, to become a web-based tool for similarly-based image search for manuscripts, incunabula and historical book collections.

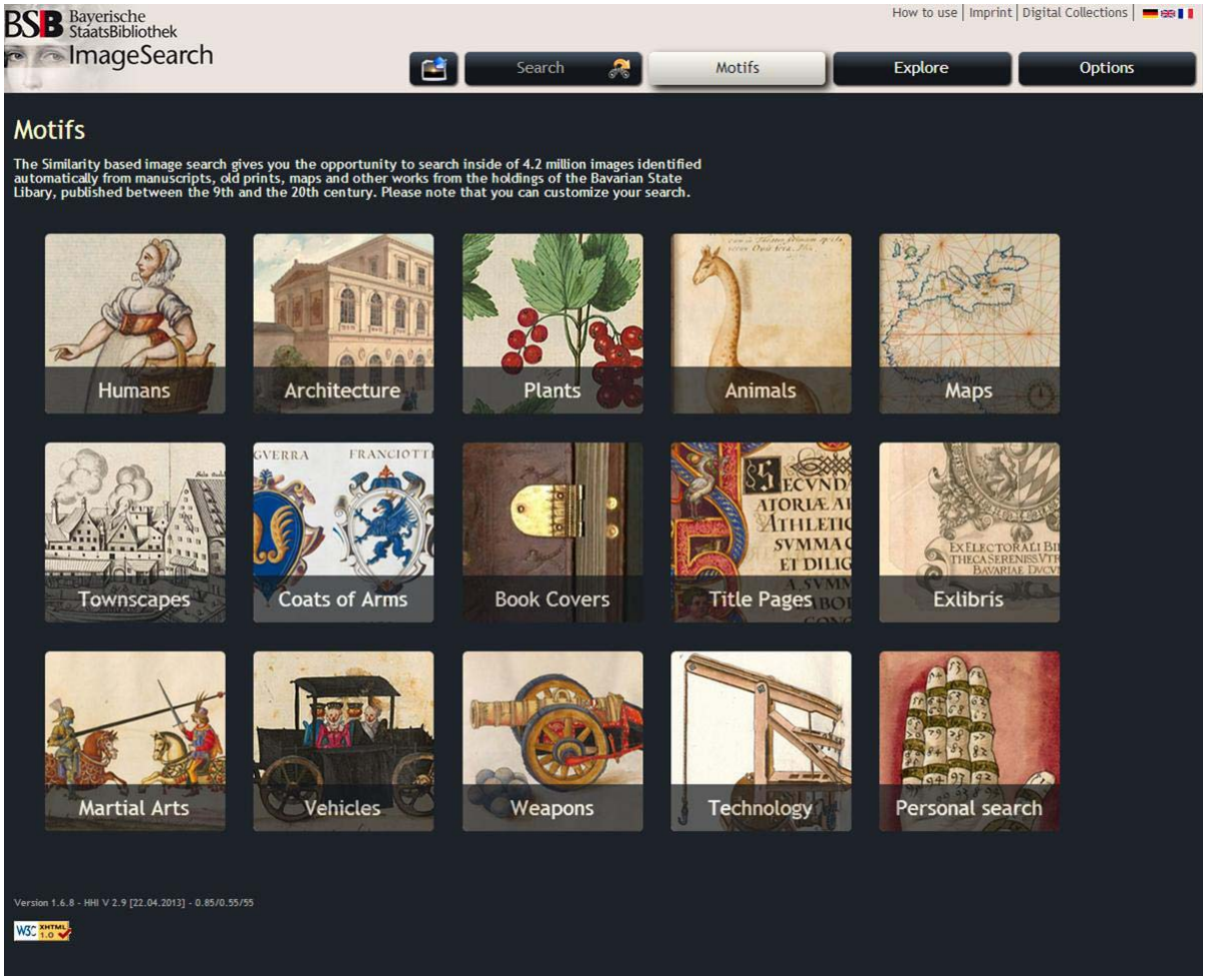


**Image 1: Homepage of the similarity-based image search**

With the new technology it is now possible for the user to search for visually similar images within a body of digital works on the basis of a selected image from a digitised book of the Bavarian State Library. Currently the searchable collection comprises around 73,000 digitised books from the 6th to the 16th century of the Bavarian State Library. Newly digitised works are automatically indexed and added to the database at certain intervals. The database currently consist of no fewer than 9.5 million digitised book pages, of which around 2.5 million pages bear pictorial representations. In total the searchable data body comprises around 4 million image elements.

The similarity-based search within this collection takes place in a fully automated manner, without any intellectual indexing or structuring. In the search exclusively visual features of the images are used, concretely the colour and edge information (structure) of an image. The search thus completely dispenses with conventional catalogue data such as bibliographic information items or subject headings. Simplified, this means: The image is its own record. The automated similarity-based image search can thus simultaneously be seen as a paradigm of non-text based access to the cultural heritage: the retrieval and segmentation of images in large text bodies and the targeted location of the images via the similarity-based search takes place without classic "metadata", exclusively on the basis of the characteristics of the searched object itself.

The software first analyses the book pages for image information items and subsequently extracts the colour and edge features and their distribution in the image. These features are then entered for each individual image in a descriptor file of maximally 96 KB, which is then stored in a database in a compressed state. Due to the small size of the descriptor, high-performance processing of the visual materials for search purposes is possible even in case of large sample sizes. The research and comparison procedures themselves take place exclusively on the basis of these descriptor files.

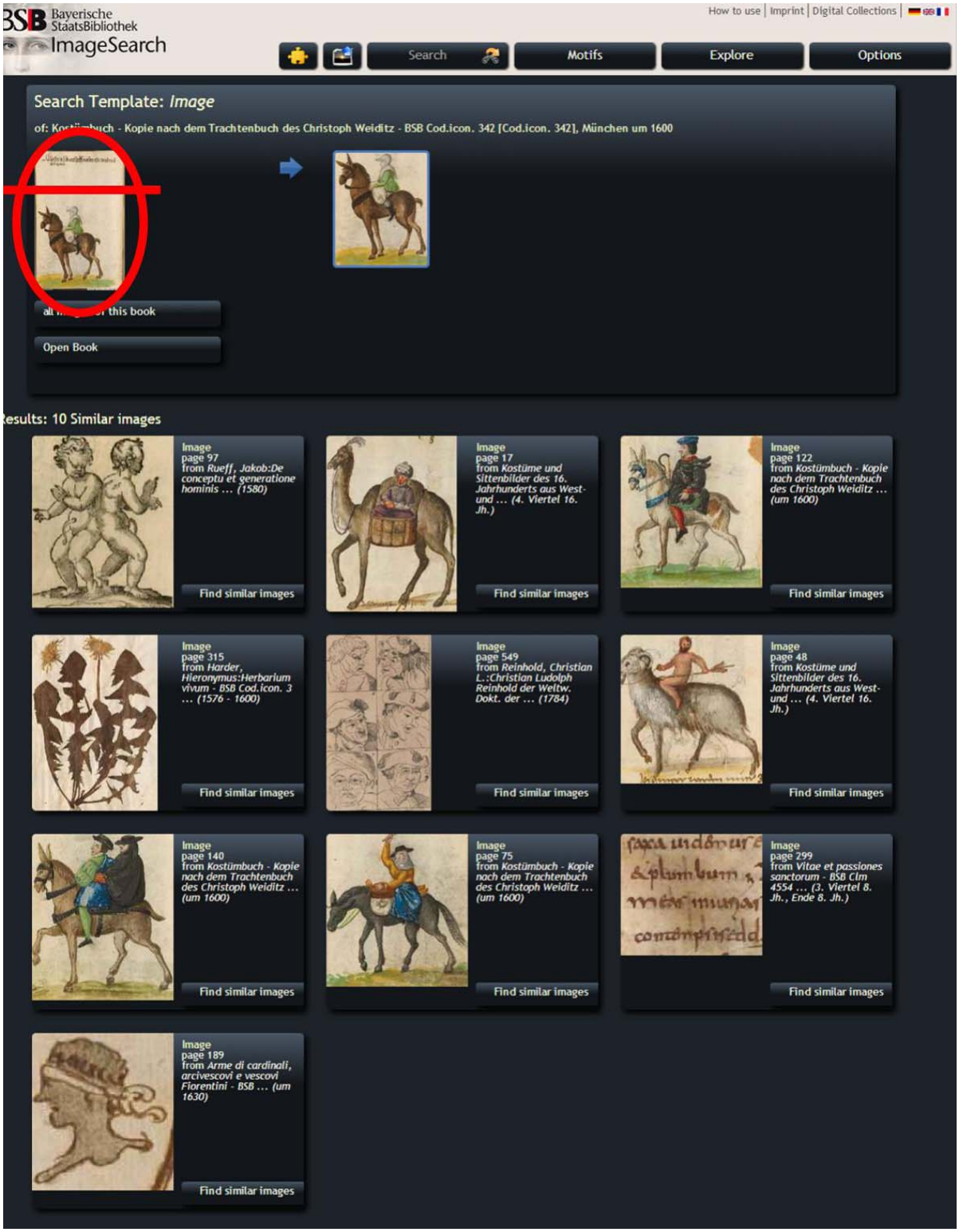

**Image 2: Automatic separation of picture and text**

The search can proceed from a known image in a digitised work that is of current interest to the user or it can be started by freely browsing a multiplicity of categories grouping thematically related works. As categories that can be selected directly from the homepage of the application there are currently available: man, architecture, plants, animals, maps, coats of arms, city views, book covers, title sheets, exlibris, art of warfare, vehicles, weapons, technology and astronomy. Upon picking a category, a user-modifiable random selection of thematically relevant images from the works allocated to the category in question is shown to the user.

The search process itself then leads to images that are similar to the respectively selected image. The works containing the searched image material can be browsed completely by means of a viewer for digital copies, and can be used as basis for further searches.

The number of hits in each case can be limited or expanded by the combined manipulation of the parameters of the colour and edge comparison. This is done by operating a slider that can be freely moved between the "poles" of the search for edge similarities and for colour similarities and that correspondingly adjusts the focus of the similarity search. At a value of 0.5 for example, a search is carried out for similarities in colour features and edge features in equal measure. Further, a slider is available by means of which the desired degree of similarity as such can be varied. Here it applies that the higher the pre-set value is, the greater is the similarity. A value of 1.0 means congruency, i.e. only such images are displayed which were detected as consistent regarding the distribution of colours and edges.

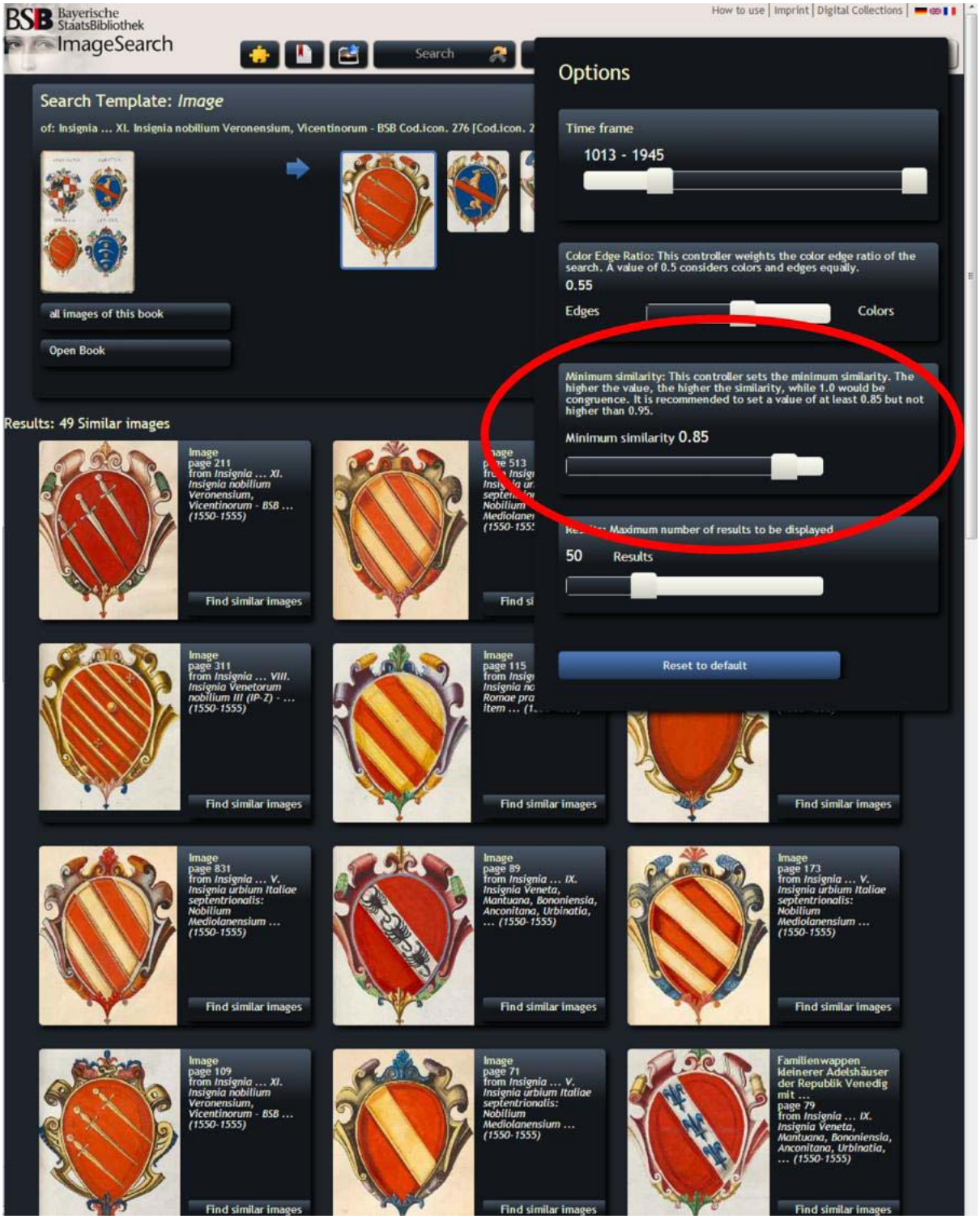


**Image 3: Variation of the degree of similarity**

The variability of the degree of similarity represents a highly innovative feature of similarity-based image search, since it is likely to blur the familiar dichotomy of machine computation on the one hand and human creativeness on the other hand. The search for 100% similarity leads only to identical images, thus for example to the use of exactly the same motif by an artist in several manuscripts. For an expert, this is normally nothing to write home about.

However, when the pre-set degree of similarity is reduced to around 95%, the results show for example surprising variations in motif or formal and stylistic consistencies in completely unrelated image contents. These are results that can raise completely new research questions. The option to vary the degree of similarity thus works as a type of "creativeness generator" by provoking unexpected questions about visual materials, with these questions in particular leading to new findings. This feature is all the more powerful as the user can also upload his own images pertaining to his current research to the database and compare them with the images of the digital collection of the Bavarian State Library. Work with the similarity-based image search is consequently not limited to the existing database of the digitised historical book collections of the Bavarian State Library, but can be expanded and complemented by one's own material.

Moreover, the software makes it possible to separate images and text when they are both located on the same page. This is a particular challenge in medieval manuscripts, where there is frequently no clear separation between image and text, and where images and text are frequently arranged "randomly" on the page. To conquer this problem, the image recognition software uses the technology of black-and-white binarisation of the page information so as to detect the morphology, for example when a heraldic figure was drawn between text passages. Here, the coat of arms is recognized, due to its chromaticity, as a "black" area, in contrast to the black and white text, and is correspondingly recognized as a delimited element. Further, it is also possible to segment parts of an image (e.g. of one heraldic figure when several heraldic figures are placed on one page, or of one tendril element in a larger floral pattern) and to subsequently search for the respectively selected partial element.

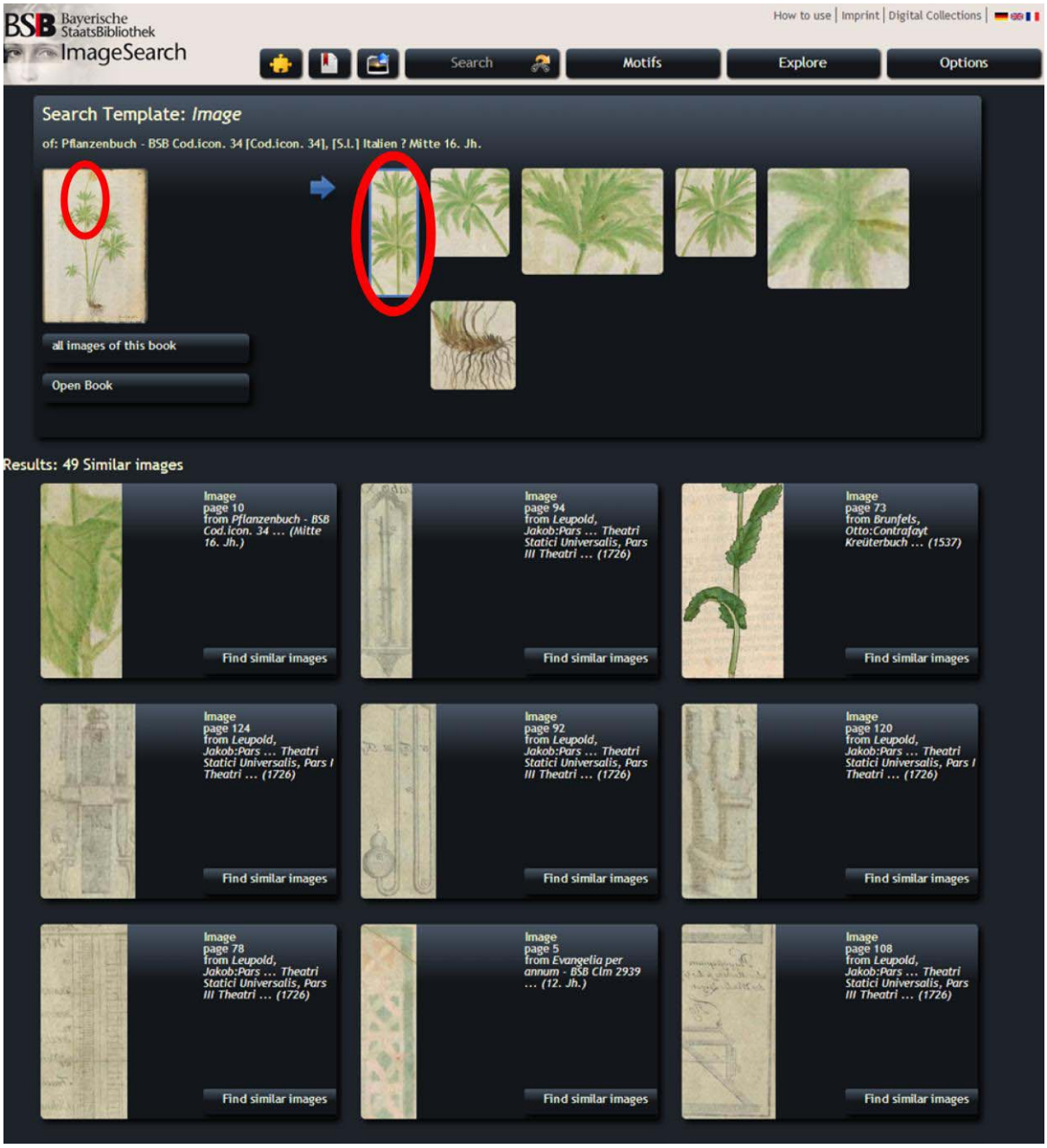


**Image 4: Retrieval in separated elements of pictures**

The layout of the image similarity search is designed in the fashion of so-called "responsive web design": The appearance and the operation of the application adapt flexibly to the different size and resolution of the display of the respective terminal (e.g. laptop, tablet PC or smartphone), thereby also making the application suitable for mobile use scenarios.

The similarity-based image search creates new possibilities for research in the field of humanities. On the one hand, it helps in the first place to find image material hidden in large quantities of digitised text bodies. On the other hand it offers the possibility to subsequently correlate this material with similar images in a multiplicity of ways. It should be taken into account that the search for similarities between images and image elements takes place in a

strictly automated fashion and is subject exclusively to formal characteristics of colour distribution and edge structures. Consequently, also such material is found that is formally, but not necessarily thematically similar to the reference image. The positive side of this is that in this fashion the similarity-based image search is in a position to offer multifaceted inspiration not only to art historians, but also to artists and designers, by making structural or formal consistencies visible in materials which are not thematically related to each other.

Beyond its direct purpose of use, the image search of the Bavarian State Library also offers new perspectives in regard of the capturing and provision of new mass data. The software operates directly on the level of the digital visual materials themselves: in the descriptor file produced per image there is captured solely information about the formal structure of the image itself. This means that no human-intellectual performance is required: the search is "free" of subject headings or keywords and classification systems. The digital object basically functions as its own metadatum: the image is the record. The exploration of the possibilities offered and also the boundaries set by this new access option of digital contents will be an important task of the future information-scientific research.

The similarity-based image search thus represents a further contribution to increasing awareness and improving accessibility of the unique written cultural treasures safeguarded by large universal libraries such as the Bavarian State Library. The similarity-based image search can be accessed on the Internet at http://bildsuche.digitale-sammlungen.de, via the cultural portal of the Free State of Bavaria at www.bavarikon.de or directly via the website of the Bavarian State Library (www. bsb-muenchen.de).