Heritage Iconographic Content Structuring: from Automatic Linking to Visual Validation

Emile Blettery, Valérie Gouet-Brunet

Heritage Iconographic Content Structuring: from Automatic Linking to Visual Validation

EMILE BLETTERY, LASTIG, Univ. Gustave Eiffel, ENSG, IGN; Ville de Paris, DAC, DHAAP, France
VALÉRIE GOUET-BRUNET, LASTIG, Univ. Gustave Eiffel, ENSG, IGN, France

This article presents a global framework dedicated to the structuring of iconographic heritage collections. To alleviate the poor interlinking both between collections and contents, a first step of automatic linking exploiting content-based image retrieval approaches is evaluated and adapted to the visual variability of such heritage contents. To ensure understanding and analysis of the contents in a structured fashion, a 3D immersive web platform is also introduced alongside visual-based analysis tools. Finally, by exploiting both automatic linking and manual interventions in the visualization platform, an iterative, semi-automatic structuring pipeline is proposed to solve difficult cases missed by automatic structuring, and then improve structuring optimally. Here, we demonstrate the potential of the proposal on the geographic iconographic heritage of Paris, with a dataset of 10k images belonging to several institutions, thus poorly connected nor organized globally.

CCS Concepts: • **Information systems → Top-k retrieval in databases**; **Image search**; • **Human-centered computing → Interactive systems and tools**; **Visualization**; • **Applied computing → Arts and humanities**.

Additional Key Words and Phrases: Image retrieval, Re-ranking, Data linking, Graph visualization, Geographical iconographic heritage

ACM Reference Format:
Emile Blettery and Valérie Gouet-Brunet. 2024. Heritage Iconographic Content Structuring: from Automatic Linking to Visual Validation. *ACM J. Comput. Cult. Herit.* 37, 4, Article 111 (August 2024), 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Improving the structuring of iconographic collections is key for many applications requiring their management at large scale, in order to better access them and more generally analyze them. These contents can be structured in a variety of ways, from indexing them using metadata or spatializing them on maps, up to mass visualization for a better overall overview. Usually performed manually, this structuring can be done more or less automatically, with several methods exploiting metadata, visual similarities between contents or spatialization. This structuring is often performed at collection or institution scale, thus adapted to the specificities of each collection (*e.g.* with dedicated metadata standards) but not suited to structure contents between collections. Indeed, collections are usually organized in silo, each having its own structuring paradigm. This is for example the case with archive institutions that operate at the national level but also at many regional levels in a lot of countries.

Such in silo structuring proves detrimental to the optimal exploitation of this novel, growing and rich digital(ized) source, that would benefit from cross-fertilization. Indeed, querying each collection independently using its own structuring has its limits and interlinking contents in an interoperable or uniformed fashion appears paramount to exploit the full potential of such contents at a larger scale. Because of the variability in metadata

Authors' addresses: Emile Blettery, LASTIG, Univ. Gustave Eiffel, ENSG, IGN; Ville de Paris, DAC, DHAAP, France, emile.blettery@ign.fr; Valérie Gouet-Brunet, LASTIG, Univ. Gustave Eiffel, ENSG, IGN, France, valerie.gouet@ign.fr.

ACM 1556-4673/2024/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

contents and standards from one collection to another, we have chosen to integrate, in the structuring process, approaches directly relying on the content itself, *i.e.* on content-based image indexing and retrieval (CBIR).

Tackling heritage iconographic contents through such structuring tools is not a solved problem either, because their specificity, as detailed in **Section 2**, resides also in their high visual heterogeneity which remains challenging for CBIR. With the final objective of content interlinking within and between collections, the first contribution of this paper is to evaluate and improve existing automatic CBIR approaches when confronted to the visual specificities of such contents. **Section 3** thus first reviews and evaluates existing image descriptors and re-ranking approaches. Three novel methods designed to better handle the impact of the visual diversity of heritage content are also presented.

As automatic structuring of such contents with visual-based approaches proves to be challenging, browsing through the collection in a structured fashion appears essential to evaluate the performance of the automatic linking, particularly in the context of collection interlinking where the visual variability between contents is high. Thus, a graph-based representation of the dataset's structure is introduced and leveraged in **Section 4** within a 3D, spatialized, web-based visualization platform. The structure is forged by several types of links between the contents and several visual paradigms of visualization, that allow the user to observe the collections and their structure visually, as well as easily evaluate the performance of the automatic linking process.

Finally, leveraging both the automatic linking approaches and the interactive visualization platform, an iterative, semi-automatic structuring process is proposed in **Section 5**. It jointly uses the best of both worlds, on one side the large scale linking methods and on the other the specialized expert knowledge. This combination allows for an iterative semi-automatic approach of graph-based structuring with the objective of increasingly consolidating the structure.

Though potentially applicable to any iconographic heritage, we have chosen to demonstrate the relevance of this synergy for the domain of *geographic* heritage contents, and more precisely to the specific task of instance retrieval where different visual representations of the same place are linked. We apply our work on several collections of the geographic iconographic Parisian heritage, which fits well such a structuring exercise because hosted over numerous French institutions and thus indexed with different standards, which makes them poorly connected and organized globally, whereas they share iconographic contents documenting common geographical areas.

2 ICONOGRAPHIC HERITAGE CONTENT

Iconographic heritage consists of all past visual representations of a society's way of life, culture, buildings, artwork, technological innovations and so on. The objects and scene depicted can be multiple: from common objects like vases to mundane scenes of life via monuments but also portraits for instance. Those contents can be found in many different public or private GLAM institutions (Galleries, Libraries, Archives and Museums), at the city, regional or national level, including for instance mapping agencies. To better understand the challenges of analyzing and indexing *by content* iconographic heritage collections, we first revisit here some general characteristics of these contents. Among the high diversity of contents, this work focuses on geographic iconographic heritage, that is depictions of immovable places that suffer more or less changes through time while remaining in the same place. Then secondly, this section presents the dataset we have designed to evaluate the performance of automatic indexing approaches.

2.1 General characteristics

Automatic structuring approaches for heritage image collections suffer mainly from two main aspects, described in the following section: the heterogeneity of the collections and their specific organization and structuring.

High visual diversity. The extremely diverse aspect of the data is first due to the multiple visual aspects of these contents. From paintings to photographs via historical maps or drawings, the representations are multiple in terms of visual depiction, as illustrated in Figure 1. Even for a single acquisition source, *e.g.* photographs, the content variability can be great, in terms of quality, resolution, colorimetry, point of view or scene modification through time. The increasing digitization process only furthers this diversity as many new contents are progressively made available, as shown by online galleries such as the Europeana Collections[1]. Thus, iconographic heritage contents must be apprehended as a nebulous constellation of objects potentially hard to link together automatically with off-the-shelf approaches.

Heterogeneous metadata. This nebulous aspect is increased by the heterogeneity of the metadata associated to the contents. Metadata can be very different depending on multiple factors like the collection considered, the type of information given, the curator of the collection, and also the purpose of the collection at its inception. Amongst GLAMs, different data models and associated ontologies are used to organize the data. Museums mostly rely on *Cidoc Conceptual Reference Model*[2] (Cidoc-CRM) whereas archives rather exploit *Records in Context - Conceptual Model*[3] (RiC-CM). Furthermore, the metadata is associated to a specific time and could represent a situation that evolved through time. An easy example to understand the heterogeneity and the variability of metadata is the location information potentially associated with a content. It can come as a single sentence for an address or as 2D/3D point coordinates, up to 6D location with mapping agencies [42]. The location can have had meaning at the time but now corresponds to a renamed or destroyed street. This example shows the potential heterogeneity within the metadata, which reinforces the global diversity of those contents and complexifies their structuring.

Structuring strategies. To structure the collections, several attempts have been initiated by curators, with the objective on enriching and linking the contents. First, ontologies exploiting the metadatas available and tailored to the considered contents were designed. Amongst the data models presented before, several ontologies have been defined, depending on the vocabulary required by each collection (*e.g.* RiC-O[4], PROV-O[5] or CRMsci[6]). By exploiting them, more or less sophisticated links between contents can be created, for instance using the solid RDF data structure [15] in a graph-based linking scheme. However, linking contents organized using different ontologies requires to produce a equivalency table between vocabularies, which is still a complex task today, in perpetual evolution as soon as a new area of analysis is considered. To bypass those hard constraints, other graph-based databases such as Neo4j[7] exploit the paradigm of Labeled Property Graphs [48] and are more flexible but more prone to errors, with a lesser quality linking. Finally, to emancipate the structuring from any pre-designed specific vocabularies and concepts, approaches based on folksonomy exploit the opportunity to make data available to a large audience, involving non-specialists to index the contents based on simple tags, linking contents in the process. Although powerful in terms of scale, those approaches lead to a low-level linking and a higher number of errors thus involving a posteriori multiple expert verifications.

All those structuring efforts have helped enrich the collections but lead to drawbacks in terms of inter-collection linking. From very specific but unconnected ontologies to low-level, uncertain tagging, none of the solutions prove to be a solid base for interlinking contents between collections. This aspect make iconographic contents ideal and challenging objects to apply visual content-based approaches. Independent from metadata or existing indexing structures, those approaches could allow to link contents and bridge gaps between isolated collections.

[1] https://www.europeana.eu/
[2] https://www.cidoc-crm.org/
[3] https://www.ica.org/
[4] https://www.ica.org/en/records-in-contexts-ontology
[5] https://www.w3.org/TR/prov-o/
[6] https://cidoc-crm.org/crmsci/
[7] https://neo4j.com/

2.2 The considered dataset

Iconographic heritage can represent cultural aspects (from monuments to mundane places and scenes of life), but also natural or geographical landscapes, depicting scenes at different times in the past. As a support to evaluate our entire work on iconographic heritage content structuring, a dataset representative of the previously listed specificities was designed; it is presented below and is later called $DB_{query+dist}$. The dataset assembled consists of more or less recent heritage content depicting outdoor areas of Paris between 1915 and 2015 from a mostly ground-level perspective. The collections belong to eight providers, coming from Parisian institutions, the French mapping agency and the Computer Vision community:

- the Department of Architectural History of the City of Paris,
- the COARC, a service of the Department of Architectural History specialized in religious buildings,
- the mobile mapping 2015 Stereopolis dataset from the French mapping agency [42],
- the Planet's Archives - Paris of the Albert Kahn Museum,
- the Cité de l'Architecture et du Patrimoine,
- the Médiathèque du Patrimoine et de la Photographie,
- the Commission for the Old Paris,
- the Paris6K public benchmark [45].

Fig. 1. Examples of images from dataset $DB_{query+dist}$

In total, we assembled a dataset of 1,637 images of which an example is shown in Figure 1, divided into 31 classes depicting regular buildings, renowned monuments (*e.g.* the Panthéon), churches (*e.g.* the Saint-Sulpice church), and remarkable buildings (*e.g.* the Lavirotte building). Beyond the fact that the buildings depicted are various, the dataset's complexity resides in the high variability in terms of representation scales, from full streets, places, fountains, etc. to facades via tiny architectural details, a typical aspect of heritage collections which were mostly acquired on the fly, focusing on multiple aspects of the scene. To further challenge image retrieval in the experiments, we added 8,197 various images as distractors (from the Department of Architectural History of the city of Paris), which leads to a total of **9,834 images** in the dataset.

Due to the large time period of acquisition and the multitude of providers, $DB_{query+dist}$ displays a large number of specific challenges for image retrieval:

- different techniques of acquisition, colors, etc.
- different resolutions, levels of details, artisticity, points of view, etc,
- collection specificities increasing the above differences,
- changes in the depicted scenes due to the evolution of Paris throughout the century.

In addition to these images, several metadata may be available sometimes, such as an acquisition date or a location. The latter may be of various types, from an address manually provided (it is the case with some images, *e.g.* those from the Dept. of Architectural History of the City of Paris) up to a precise 6D pose of the camera (with the mobile mapping system Stereopolis).

3 AUTOMATIC LINKING

Automatically linking contents can be performed in multiple ways, more or less suited to the specificities of iconographic heritage; Figure 2 illustrates the whole pipeline of content-based image retrieval. A large part of the work presented in this section was published in [7]; this section summarizes those contributions and also extends them. It first details the evaluations performed to evaluate state-of-the-art image descriptors, in order to find the most-suited one for the iconographic heritage collections considered (Section 3.1). Once this first step of the CBIR pipeline is evaluated, the second step of the pipeline is investigated: Section 3.2 first explores state-of-the-art re-ranking approaches and presents three re-ranking approaches more suited to handle visual heterogeneity of the contents. Finally, Section 3.3 introduces a novel discussion on the importance of provider entropy for re-ranking, in the specific context of collection interlinking.

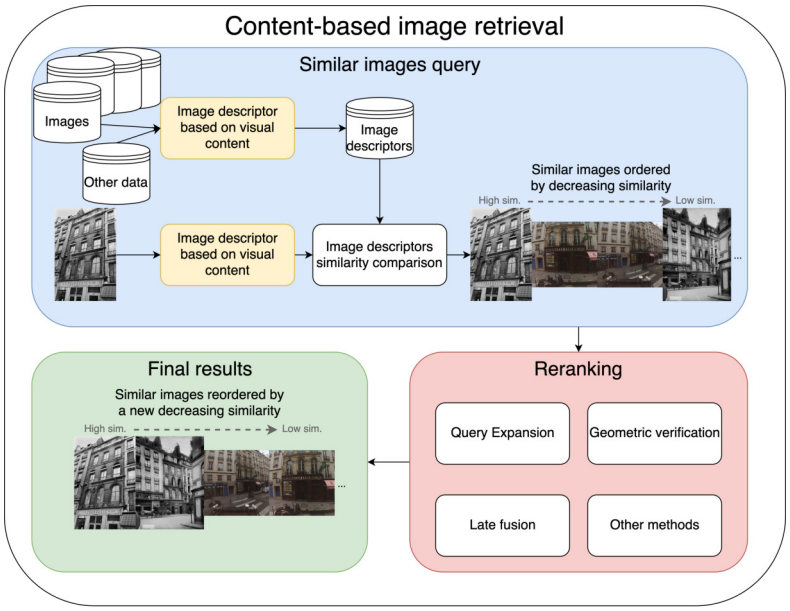


Fig. 2. Content-based image retrieval general pipeline

3.1 Content-based image retrieval

Most common and efficient image descriptors suited for CBIR are now learned from large scale datasets using different models. The methods are numerous and have successively become state-of-the-art [13]; we revisit them briefly below.

3.1.1 State of the art. Most networks use backbones initially designed for classification tasks and adapted for image retrieval; one can mention VGG [54], ResNet [26] or ResNest [68]. The features extracted from those

backbones are then exploited to create the final image descriptor. It can lead to global features (that describe the image content globally) after a pooling operation, as in SPoC [3], GeM [47] or more recently CVNet [30]. Local features, describing salient areas in the image, can also be extracted and then aggregated to be compared with local features from another image. To select the most meaningful features, attention mechanisms are used, such as in DeLF [39] or How [60]. The aggregation step afterwards for comparison between images can be done in various ways as with the visual-bag-of-words paradigm. We specifically mention ASMK [59] which is efficient with local features, itself first associated with How features. Several works also combine global and local features to leverage the specific performances of both approaches: one can mention DELG [12] and CVNet [30] which use local features in a second, trained, re-ranking step and DOLG [67] which fuses both types of features into a single descriptor. Finally, new approaches leverage vision transformers as backbones for feature extraction, like [64] which uses a pyramidal approach to create a backbone for multiple vision tasks or [18] which leverages a cross-shaped attention mechanism on arbitrary input resolution for several downstream tasks.

3.1.2 Evaluation framework. For this study, four state-of-the-art descriptors are evaluated: DELG [12], R101-GeM [26, 47], CVNet-Global [30], How+ASMK [59, 60]. All methods are deep detectors and descriptors. The first three produce a global feature per image and are trained on Google Landmarks Dataset v2 (GLDv2) [65], whereas the last one, trained on SfM120k [46], produces local features and then aggregates them using ASMK [59] for comparison between image descriptors. For all experiments, we have chosen to not retrain the involved networks, first because it does not exist a training dataset dedicated to iconographic heritage contents and second because the regular evolving and growing of these heritage collections would require frequent retraining to maintain optimal descriptors. To quantify the performance of the descriptors on $DB_{query+dist}$, we use the mean Average Precision score (mAP), as designed by [47][8]. All experiments are run on a Tesla-V100 GPU with 16 GB RAM and 10 CPU cores. Table 1 presents the results of this evaluation, on the dataset including or not the distractors.

Table 1. Score of tested image descriptors on dataset $DB_{query+dist}$.

mAP	Dataset w/o distractors	Dataset
DELG [12]	53.2	-
R101 - GeM [26, 47]	57.9	38.5
CVNet-global [30]	**67.3**	37.1
How + ASMK [59, 60]	55.1	**41.0**

The first evaluation is performed on the dataset without distractors. At this step, CVNet outperforms all other descriptors by a large margin. Local descriptor How + ASMK (How-A) reaches only third place. When working with the whole dataset $DB_{query+dist}$, distractors with more visual heterogeneity but also similar features in multiple instances (frequent in our dataset with architectural photographs) are added in the retrieval process. This drastically changes the order of performance of the descriptors. Indeed, How-A becomes the best performing one, while CVNet gets to third place, though with less difference in performance. The reason for such a change most probably lies in the type of descriptor. Indeed, as a local descriptor, How-A tends to be more discriminative when visual elements are similar in parts of the images or at different resolutions, compared to global descriptors (CVNet or R101-GeM) which tend to encapsulate a global representation of the content. This aspect explains the bigger robustness to distractors exhibited by How-A. The local aspect of the descriptor also proves interesting for cross-provider retrieval, local features being more easily retrieved across different visual aspects. In our context

[8]https://github.com/filipradenovic/cnnimageretrieval-pytorch

of collection interlinking, this is paramount, as further discussed in Section 3.3 when focusing on cross-collection retrieval.

3.2 Re-ranking strategies

After having retrieved the images most similar to the query, via the nearest neighbor descriptors, a common second step in a complete pipeline of CBIR is a re-ranking one, which consists in reordering these responses based on another similarity score. Multiple approaches have been designed, this section successively quickly presents and evaluates the most representative ones in the state of the art (Section 3.2.1), and also proposes new approaches more suited to the visual heterogeneity of iconographic heritage: Section 3.2.2 exploits geometry to extend the scene encoded by the query image, in 3D or in 2D and Section 3.2.3 leverages structuring information extracted from metadata to weight the visual similarity.

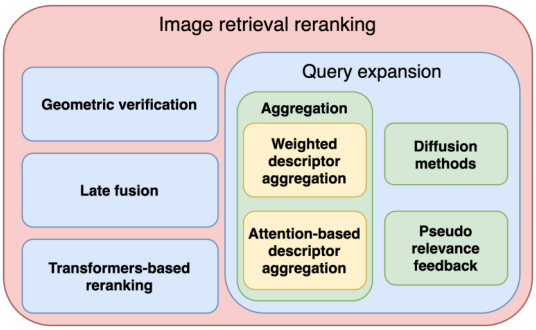


Fig. 3. Re-ranking methods paradigms

3.2.1 Re-ranking state of the art. As presented in Figure 3, re-ranking methods can be divided into several categories. Late fusion exploits multiple features or similarity distances to fuse the similarity lists; [63] for instance exploits multiple distances between the features to obtain several retrieval lists to fuse.

After having retrieved image candidates, a very common re-ranking step is the geometric verification one, as in the retrieval pipelines of [12, 30, 67]. When considering rigid objects in the scene, the principle is to match local features between images, estimate the geometric transformation (affine, homography) existing between them from those matches and re-rank the images based on the number of matches that fit this transformation. The local features exploited can be derived from the retrieval step, such as in DELF [39] but also be those used in other tasks like Structure-from-Motion (such as SuperPoint [17]). Matching the features can be performed with several matchers, we mention SuperGlue [50] and LightGlue [33] which prove to be the most robust to high visual heterogeneity. Finally, the estimation of the transformation is mostly performed within a RANSAC (RANdom SAmple Consensus) loop which fits the data to a model while being robust to outliers (incorrect matches). RANSAC [20] is a proven method and multiple adaptations have been proposed like in [5, 12, 37].

Another family regroups query expansion methods, which take advantage of contextual information from the first retrieved images list. It can be by aggregating the features of the query and its most similar images to increase the meaningfulness of the query descriptor in order to improve the retrieval results. Multiple adaptations have been proposed, such as changing the aggregation weighting scheme (Average-QE [3], α-QE [47], etc.). More recent methods [25, 70] use an attention mechanism to select images and their weight in the aggregation process. In the setting of pseudo-relevance feedback [31, 32], it is assumed that the first retrieved images depict the same

scene while further retrieved ones are incorrect. Based on this prior, the query's descriptor is modified to be closer to those of the supposedly correct images and further from those of the *a priori* incorrect ones.

A specific kind of approach within query expansion re-ranking relies on diffusion, which propagates the similarity through the k-NN graph of similar images. Such solutions have achieved state-of-the-art performance on many benchmarks [4, 16, 28, 41, 53]. In the article, we will focus on a representative method of this category, called GNN-R afterwards, from [69].

Finally, learning methods of re-ranking have been proposed to exploit the recent paradigm of transformers: re-ranking transformers [58] is a network that predicts the similarity of an image pair directly, provided their global and local features. Meanwhile, inspired by query expansion approaches, [40] proposes a transformer-based network that aggregates affinity features among the first results to enrich the representations of the images with some contextual information. [71] exploits transformers first for global image description and retrieval and then for re-ranking. It exploits feature correlation rather than geometric verification.

We have evaluated on our dataset the most representative and efficient approaches of each category (without retraining models). Table 2 provides the order of magnitude of the improvement obtained facing simple retrieval, in terms of mAP score.

Table 2. Modification of mAP score by adding a re-ranking step after retrieval

Approach	Order of magnitude
Weighted descriptor aggregation [14, 47]	+ 0.1
Pseudo relevance feedback [31]	< + 0.5
Transformers-based: CSA [40], RRT [58]	- 10
Geometric Verification :	
RANSAC [17, 33, 50]	+ 0.5-1.5
CV-Net Rerank [30]	- 2
Diffusion [53, 69]	+ 16

Approaches exploiting the visual descriptors to compute a new descriptor for the query [14, 47] or modify it [31] do not perform well, due to high visual variability within the heritage images which conduces to variable descriptors. Trained approaches [30, 40, 58] are detrimental, a retraining step is probably needed to reveal their potential. We observe that a classical RANSAC [17, 33, 50] remains pertinent to some extent, while diffusion approaches [53, 69] are extremely beneficial. Unlike other re-ranking approaches using only a small neighborhood of the query belonging to the first retrieved images, the diffusion ones exploit the similarities of a larger neighborhood involving indirect neighbors. This bypasses the high heterogeneity of the data and re-ranks in a more global fashion, thus leading to a large improvement in terms of mAP score.

The scenario we are evaluating in our work focuses on linking contents depicting the same, specific places. To this end, three novel re-ranking approaches were proposed, better suited to both this scenario and the visual heterogeneity of the contents. The three proposed re-ranking approaches, published in [7] and revisited in Sections 3.2.2 and 3.2.3, leverage information specific to each query to increase the discriminative aspect of the retrieval between specific objects.

3.2.2 Geometric query expansion. To further investigate re-ranking dedicated to iconographic heritage, we have studied several strategies to adapt to the visual specificities of the contents, by focusing on the potential large variation in viewpoints and levels of detail through the collections, which may impair the performance of classical

pairwise geometric verification. To overcome this, within a geometric query expansion setting, two methods have been designed, namely R3D and R2D. The core of these approaches were published in [7] and are revisited here.

R3D approach first reconstructs a 3D point cloud of the scene using Structure-from-Motion algorithms (in our case Colmap [51, 52], with default parameters) based on keypoint matches extracted with [17] and matched with either [50] or [33] (same as with RANSAC). For each query, the first ten retrieved images are exploited to reconstruct a scene. If it succeeds, that is if the scene is reconstructed with at least two images including the query, the first k retrieved images are then repositioned in the scene with 2D-3D registration. The images are then re-ranked based on their geometric adequation to the scene, mainly by examining their number of matches with the scene. This scene encodes a more global geometry than the single query image, thus correct images different from the query can still be linked to the scene and thus to the query [7].

As a full 3D reconstruction is a costly process, an approximate 2D version of R3D was proposed, with the objective to effectively mimic its main features: **R2D approach** extends the geometric significance of the query image representation by integrating the most significant features from the first ten retrieved images. To this end, triplets of images including the query and two out of the ten images are created and keypoints are matched all around. The keypoints of the query image are then divided between solid points (matched in a loop pattern in the triplet of images), unsolid ones (matched only with one image). Furthermore, keypoints from the other images are reprojected in the query, provided there are enough solid matches to estimate an homography between the query and the retrieved images. Once the extended set of keypoints of the query image is defined, geometric verification is performed in a classical fashion between the query and the first k images. The similarity score is based on the number of inliers, weighted by their type (solid, unsolid, reprojected). In this way, no costly reconstruction is performed, while a more global scene geometry is encoded in the query image's description [7].

Both approaches R3D and R2D were evaluated in [7] but we present and discuss a deeper evaluation of their performance here in Section 3.2.4 and Table 3, with different feature matchers.

3.2.3 Spatial location weighting. As previously shown with R3D and R2D, exploiting structural (geometrical) information improves the re-ranking performance. To continue to evaluate other tracks exploiting the specificity of the manipulated data, we chose to be interested in their metadata, starting from the observation that in practice, some metadata are present at least partially, in some of the collections. Like image retrieval, such data may provide useful links between images that can be simply but efficiently combined with visual similarity.

An easily structuring information is the position of the images. The nature of the information is varying, from 6D poses from mobile mapping (Stereopolis) to 2D points from geocoding textual addresses. Furthermore, the quality of the information can greatly vary from certain GPS acquired data to manually added addresses subject to human error. In the specific case of iconographic heritage, the time gap can also lead to modification in the environment (*e.g.* streets renamed or created).

Thus, to each image I is given a confidence coefficient c_I based on its location's quality, when available. Based on the available locations in the dataset, we build a spatial proximity score $s^s_{I,J}$ between images I and J, based on the spatial Euclidean distance $d_{I,J}$ between I and J. This score allows to weight their visual similarity, namely score $s^v_{I,J}$, obtained from initial retrieval. This weighting allows to give more importance to similar images which correspond to close locations, under the scenario that we want to bring closer images showing the same areas. The final weighted similarity score $S_{I,J}$ between I and J is defined as:

$$S_{I,J} = s^v_{I,J} \times {s^s_{I,J}}^{c^s_{I,J}}, \tag{1}$$

where $c^s_{I,J}$ is a confidence score based on both location's confidence scores:

$$c^s_{I,J} = \begin{cases} \frac{1}{c_I \times c_J} & \text{if } s^s_{I,J} < 1 \\ c_I \times c_J & \text{otherwise} \end{cases} \tag{2}$$

If not both images have a location information, score $c^s_{I,J}$ is of 0, meaning that $S_{I,J} = s^v_{I,J}$.

This approach is presented in details in [7] and also evaluated in Section 3.2.4 and Table 3. The confidence score c_I is 1 for Stereopolis locations, as it is precisely acquired as part of the mobile mapping, while for geocoded addresses, we set it below (0.9 for queries, 0.8 for distractors), because they are considered slightly less reliable due to massive human intervention.

3.2.4 Evaluation and combination of re-ranking methods. We evaluate the methods presented previously but also their combination, when relevant. Combining multiple approaches should be performed in such a way that each re-ranking approach benefits from the previous one(s). For instance, R3D will benefit from a previous RANSAC step which ranks closer geometrically coherent images. Furthermore, as the diffusion-based re-ranking leverages the graph of k-nearest neighbors for most similar images, it gains from the best ranking list for each query, that is with other re-ranking methods being performed before diffusion. To summarize, pairwise approaches RANSAC or location weighting should be performed before R3D or R2D which leverage the first ten images and diffusion should be performed as a final step, exploiting all previous re-ranking in a global fashion. Indeed, a first step of geometric/spatial re-ranking is performed and then the diffusion process is applied multiple times.

All experiments are run on a Tesla-V100 GPU with 16 GB RAM and 10 CPU cores. For geometry-based approaches, two matchers are used, indicated by the suffixes -SG if SuperGlue [50] is used and -LG if it is LightGlue [33]. For the location weighting method, different locations are used: either only Stereopolis data (Sp), all queries'data (No Dist) or all data available including distractors (All). The diffusion step used is from [69] and is called GNN-R afterwards. Table 3 presents the results obtained, in terms of both mAP and mean retrieval time.

Table 3. mAP scores and mean retrieval time for multiple combinations of re-ranking steps. Indicated in color are the first, second and third best results for each column, and in **bold** the best score overall.

Descriptor + Re-ranking step(s)	Diffusion after previous re-ranking				Mean time
	No GNN-R	GNN-R × 1	GNN-R × 2	GNN-R × 3	
How-A	41.0	57.2	59.3	57.0	
How-A + RANSAC-SG	41.5	57.2	59.3	57.0	+120s
How-A + RANSAC-LG	41.9	61.2	65.5	63.3	+100s
How-A + R3D-SG	44.4	61.9	64.2	61.9	+220s
How-A + R3D-LG	43.2	61.1	63.2	60.7	+210s
How-A + R2D-SG	36.2	59.6	62.9	60.5	+150s
How-A + R2D-LG	41.9	61.0	64.4	62.1	+140s
How-A + location weighting (Sp)	42.0	58.9	61.8	59.5	+1/30s
How-A + location weighting (No dist)	40.5	57.8	61.1	59.0	+1/30s
How-A + location weighting (All)	42.5	60.2	63.1	61.8	+1/30s
How-A + RANSAC-SG +R3D-SG	44.9	62.9	**65.8**	63.3	+340s
How-A + RANSAC-LG +R3D-LG	43.0	61.8	64.1	61.9	+300s
How-A + RANSAC-SG +R2D-SG	36.9	60.1	63.0	60.5	+270s
How-A + RANSAC-LG +R2D-LG	41.7	61.2	64.3	62.2	+240s
How-A + location weighting (Sp) + R3D-SG	44.7	62.4	64.9	62.4	+220s
How-A + location weighting (Sp) + R2D-LG	41.9	61.1	64.7	62.1	+140s

First, the proposed methods are all but one performing as well or better than a classical RANSAC. In the case of R2D-SG, the matcher used is too lax to perform a correct approximation of the scene. Second, R3D is better

than R2D, which is consistent with R2D being an approximation of R3D. Third, exploiting the available location information appears relevant, especially when using the locations of the distractor images (further discriminating similar yet incorrect images). Although not performing as well as the R3D approach, they are much less costly in computation time.

When it comes to combining approaches, combining diffusion steps is always significantly beneficial up to two times (GNN-R × 2), but decreases the performance with a third one. Combining a classical RANSAC before the R3D step and then two steps of diffusion leads to the highest performance (65.8%), with an increase of 15% compared to simple retrieval and 6.5% against retrieval and two diffusion steps.

To summarize, for maximum re-ranking performance with our dataset, two steps of diffusion are mandatory. Before diffusion, if time is not a concern, a first step of RANSAC-SG followed by R3D-SG before diffusion is best. However, if time is a concern, simply combining a first step of RANSAC-LG and two steps of diffusion is more than three times faster for only a 0.3% performance drop. This high performance of a single RANSAC-LG step as prelude to diffusion is proof that the mAP before diffusion is not the only parameter ensuring efficient diffusion. This is discussed in the following Section 3.3.

3.3 Provider entropy for optimal diffusion

Through the previous experiments, another aspect, essential to automatic content-based interlinking *between* collections, is revealed by the varying performance of the diffusion re-ranking step facing initial retrieval. Indeed, we observe that the initial retrieval/re-ranking performance (in terms of mAP score) is not a sufficient indication as to how well the diffusion process will perform. An example in Table 3 is R2D-SG : while less efficient, by 4.8%, than simple retrieval with How-A (36.2% vs. 41%) prior to diffusion, two steps of diffusion bring its mAP score 3.6% higher than simple retrieval with two diffusion steps (62.9% vs. 59.3%). On the contrary, RANSAC-SG improves the mAP score by 0.5% before diffusion, while after diffusion, its score is similar to How-A with diffusion.

This observation is interesting, knowing that image retrieval pipelines are used to chaining approaches together, based on the best performance step by step. Through the collections considered here, an intuition explaining this behavior can be found with the performance of retrieval in terms of *cross-collection* retrieval. Indeed, retrieval performs best within a single collection, where contents are more visually similar (due to similar acquisition protocols for instance). Thus, if the first retrieved images are of the same provider (with a limited variability of the descriptors), the diffusion tends to operate within a single collection, limiting its impact. However, if multiple providers are present in the first retrieved images (with a higher descriptor variability), the diffusion will perform with a larger pool of images across collections, leading to a better global performance.

Table 4. Evaluation of the impact of diffusion depending on provider's entropy

Descriptor +	Entropy	mAP before	Diffusion after previous re-ranking		
Re-ranking step	@20	GNN-R	GNN-R × 1	GNN-R × 2	GNN-R × 3
How-A	36.4	41.0	57.2	59.3	57.0
How-A **(Max entropy)**	61.4	41.0	66.8	69.9	67.3
How-A + RANSAC-SG	38.0	41.5	57.2	59.3	57.0
How-A + RANSAC-SG **(Max entropy)**	61.5	41.5	67.2	70.1	67.7
How-A + R3D-SG	41.5	44.4	61.9	64.2	61.9
How-A + R3D-SG **(Max entropy)**	61.6	44.4	71.4	73.7	70.4
How-A + R2D-SG	42.4	36.4	59.6	62.9	60.5
How-A + R2D-SG **(Max entropy)**	57.7	36.4	69.3	72.1	69.3

To confirm this hypothesis, the entropy *between providers* was artificially maximized within the first 20 images retrieved: without modifying the mAP score (by considering responses from the same class), a larger number of providers were injected amongst the first 20 retrieved images. Table 4 summarizes the results of this experiment for some of the re-ranking combinations.

This Table shows that independently of the mAP score, maximizing the entropy (lines with "Max entropy") leads to a mAP improvement at each step of diffusion ranging from 8.5% to 10.8%, regardless of the previous re-ranking. This confirms that more than a great mAP score after simple retrieval, in the context of heterogeneous contents interlinking, a combination of high mAP and high description variability in the responses (here provider entropy) is the most-desirable outcome for multiplying the benefits of diffusion through collections.

4 A PLATFORM FOR VISUALIZATION AND VALIDATION

Automatic linking of contents usually results in lists of similar contents, in decreasing order of similarity. Although this type of result can be evaluated using scores like the mAP presented before in Section 3, it prevents the full appreciation of the global structuring of the dataset. Associated to the automatic structuring approaches and their score-based evaluation, understanding the structuring of the dataset gains to be performed using a visualization of the created structure, which can exploit the spatialization in the case of the contents we consider. The idea is not new: since the 1980s and the so-called "spatial turn", spatialization and visualization of data has become paramount to comprehend data in its context and endogeneously enriches its understanding (by an expert as well as a novice). From simple browsing to global analysis of the data by discovering patterns via simple validation of the structuring of the dataset, the spatialization and visualization of heritage iconographic contents proves beneficial for content interlinking.

This section first introduces existing visualization platforms for iconographic contents in Section 4.1. Section 4.2 introduces a graph-based representation of the dataset in a structured fashion, that is exploited in Section 4.3 which presents the 3D environment we have designed, most-suited for this work on interlinking iconographic collections. Finally, visual-based analysis tools dedicated to the evaluation of the structured dataset are introduced in Section 4.4.

4.1 State of the art on iconographic content visualization

With the ever-increasing digitization of heritage iconographic content, structuring and visualizing them has become an objective of multiple GLAMs, research projects but also dedicated social network platforms. Because of their various objectives ranging from simple display to interactive visualization on the web, the structuring paradigms are numerous. [66] deeply reviews structuring and visualization paradigms for cultural heritage collections in all forms. Such variety in the visualization has also been investigated in the VIKUS research project [44]. In this work however, we focus on iconographic collections.

Single modality structuring. A first structuring paradigm exploits a single information type. Thus, some platforms simply organize them using metadata to structure and query contents, that are visualized without much structure between each other. One can mention *Gallica* [23], platform of the French National Library or the *BaseMemoire* [21] of the French Culture Ministry. The Oronce Fine platform [62] more specifically organizes contents in a graph based on metadata. Other platforms only exploits visual similarities to structure contents and visualize them in the space of the visual descriptors. *PixPlot* [19] is an example of those platforms. From another perspective, contents may also simply be organized based on their 2D location. That is the case with the *Remonter le temps* platform [22] where the French Mapping Agency displays its aerial photographs from the last century. The *iART* platform [56] unifies visual and textual similarity into one feature space exploited for both querying and visualizing. Contents can be displayed in a ranked or a clustered fashion to identify recurrent patterns throughout the collections.

Metadata and 2D spatialization for structuring. Other platforms then combine informations for structuring, querying and interacting with the contents. First, some use metadata and a 2D location information. Thus, contents can be queried or displayed using tags or simply browsing on a map. *HistoryPin* [27] and *Navilium* [38] use this in a social network setting while the Albert Kahn museum [1] uses it to promote its collections.

Structuring with metadata and 3D visualization. Platforms also use metadata and a specifically created 3D visualization. The project *HistKL* [35] proposes a 4D visualization platform using a specifically designed 3D model of Dresden for visualizing 6D-located images (via the process from [36]) that can be queried based on time period, metadata and position in the scene. Finally, other platforms aim at combining metadata and global 3D positioning. *SmapShot* [6] and now *Images of Switzerland Online* [57] propose to query contents based on metadata, concepts or visual similarity but also visualize them on a map and in their 3D context (meaning that the user can not fully move around in a 3D platform). On the contrary, the *ALEGORIA* project [24] proposes a platform for querying image contents based on metadata and visual similarity and then visualizing them in a fully open 3D platform where contents can be localized, and visualized jointly with other localized images and other contextual data (see [9]). Finally, within the Virtual City project [34, 49] focused on 4D modeling of urban data, [29] works on visualizing multimedia contents to enrich the visualization and understanding of urban dynamics; note that these solutions have been proposed with the aim of interoperability, reproducibility and sustainability.

Although combining several information brings analysis potential to the platform, the objective of existing platforms is mostly to visualize the contents in a structured fashion rather than to analyze the quality of the global structuring itself, meant for spatialized iconographic contents and their various degrees of similarity. The following sections develop a graph-based structured representation of the dataset that is then exploited by a visualization platform that we believe is more suited for such analysis purposes, allowing for a simple visualization environment while offering analysis tools for experts to evaluate the automatic structuring approaches on the dataset they know.

4.2 A graph-based representation of the dataset's structure

As previously shown in Section 3, the diffusion approach exploited for re-ranking optimally apprehends the dataset and the similarities between images as a graph. Apprehending the structured dataset as a graph proves beneficial in several aspects. First, a graph can jointly represent several similarity links, which proves essential in our work where several similarity links can be exploited to connect images within the collections, relying either on visual similarity or spatial location and proximity. Second, the logic inherent to a graph induces a structured visualization that can be leveraged for understanding the dataset's structuring.

We have chosen to go deeper in such organization, for both structuring and visualization purposes, by considering the dataset's structure as a graph-based representation. More formally, we exploit three kinds of links between items:

- **Visual similarity links**, which connect two images similar in terms of visual content. Such links do not exploit the metadata and can be processed automatically with approaches described in Section 3;
- **Spatial similarity links**, which automatically connect two images according to their spatial proximity in the environment, if available. Other criteria could have been chosen (*e.g.* semantic similarity) but the geolocation criterion is a useful one in several domains, *e.g.* the geographical heritage we consider in this work;
- **Location links**, connecting a location to any image at this location. Two images with matching location information coming from a different source (*e.g.* GPS or geocoded address) can be connected to the same location.

These links lead to a multi-edge graph-based representation of the collections' structure, where the nodes are image contents (and metadata) or a location. To ease the global understanding and visualization of the structure,

we exploit the global similarity score $S_{I,J}$ defined in Equation 1, which encodes the various possible links (visual or spatial).

Once the graph representation is created, a 3D visualization platform can serve as interface to visualize this structure, with the ambition of providing interaction tools to the user in order to improve the understanding of the dataset and its structuring. The developed visualization paradigm is presented next in Section 4.3.

4.3 GraphXR as visualization platform

To visualize the structured dataset in a graph-like fashion, we have selected GraphXR[9], a web-based visualization platform dedicated to graph data. We use a free version of the platform that suits our needs in terms of volume of visualization. Combined to a Neo4j graph database, we have exploited it in order to make possible the visualization of the image collections and their link-based structure; see the example of Figure 4. Our proposal allows for visualizing multiple links of different kinds but also thumbnails of the images directly within the platform, ensuring a smooth browsing. It also permits spatialization by pinning localized nodes to a map in the 3D environment. Furthermore, as a web-based platform, the visualization could be performed from anywhere by anyone, as long as the data is made available on a server running the Neo4j database and with the images available via the web. For the current implementation and experiments, we stuck to a local solution.

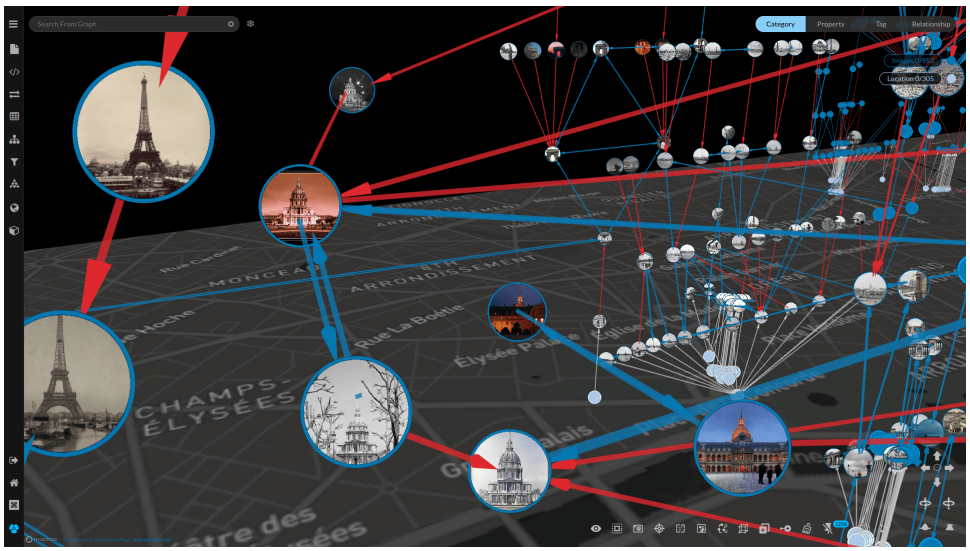


Fig. 4. The 3D graph-based visualization platform

The entire visualization platform, provides a 3D visualization area for the graph and / or map but also multiple tools and visualization solutions, further detailed in Appendix A and visualizable in a video on page 2 of this website [8].

To summarize, the platform proposed offers the user the possibility to perform multiple actions of which we list the most useful below:

- Choose the visualization paradigm (map visualization, which similarities to display, which information to display on the nodes, etc.),

[9]https://www.kineviz.com/graphxr (free version)

- Exploit graph algorithms to automatically compute new information on the graph that can modify the visualization,
- Remove an incorrect link,
- Create a link of any similarity (visual, spatial or expert),
- Add information on a link or a node,
- Update the graph database.

Section 4.4 below goes deeper in the presentation of these functionalities for advanced analysis of the collections structure. The global environment developed also allows to create macros to automate several processes in order to speed up the correction process and go smoothly from one visualization to another in a smooth fashion, giving all users, even beginners, the whole range of potential actions.

4.4 Visual-based analysis tools

Going further than a simple visualization tool, several visual representations and algorithms can be leveraged to visually evaluate the quality of the dataset's structuring. We present them in the following.

4.4.1 Visual representations. Several representation choices were made and may appear in the different illustrations of the platform. The nodes are distinguished by their visual representation. The location nodes are always pinned to the map and in light blue color. Differently, the image nodes are never directly pinned to the map and three potential visualizations are possible:

- In dark blue, the nodes simply indicate their type (that is Images);
- Node properties can be displayed in various colors (scaled or not), for instance the community of the node or its betweenness coefficient (see Sections 4.4.2 and 4.4.3);
- The thumbnail of the image that the node represents can also be represented in the node, allowing for a quick check of the scene depicted. A link to visualize the image in full is also available in the node's information panel.

The links in turn are distinguished by color, detailed next:

- **Salmon pink** ones represent the **visual** similarity between two images;
- **Green** links means a **spatial** similarity between images;
- For **global** similarity links, to further display information on the structuring, we differentiate them in three categories:
  - **Blue** similarity links represent **global, strong, reciprocal** similarity links;
  - **Red** ones represent **global, strong, single-sided** similarities;
  - **Purple** links express a **global, low, single-sided** similarity between images.

Exploiting such representations alongside graph algorithms provides visual clues to visually evaluate the structuring as detailed in the following examples (Sections 4.4.2, 4.4.3 and 4.4.4), more dynamically illustrated in pages 3 to 5 of this website [8].

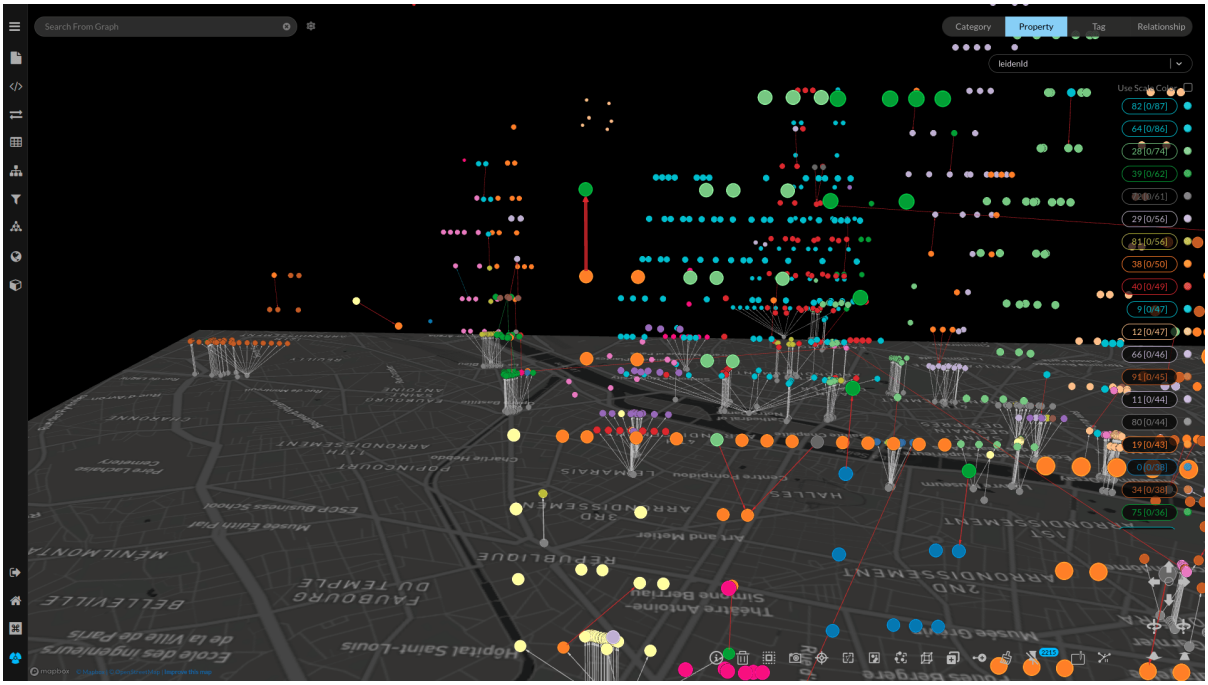
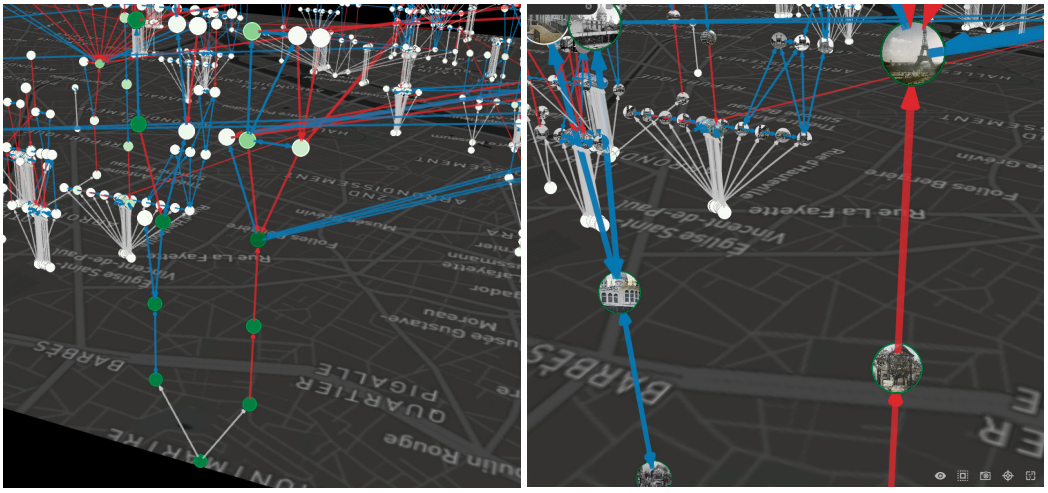


Fig. 5. Cross-community links visualization. The node's color represents its community. The visualized links are only cross-community ones (potentially incorrect).

4.4.2 Cross-community links. Within graphs, communities can be exhibited based on the different links between nodes. In our case, as we structure a dataset composed of multiple classes, an ideal structure representing a classification in terms of depicted object would be a single community for each class and no links between communities. This "ideal" case never presents itself due to the difficulties automatic methods have with such data; however, exploiting communities can still prove useful. In our platform, communities can be exhibited in the graphs by exploiting similarity links using for example the Louvain algorithm [10] or its improvement, the Leiden algorithm [61]. Visualizing cross-community links like in Figure 5 allows the user to identify which content is often mistakenly linked. This could help either to fine-tune automatic structuring algorithms to deal with such cases but also focus a manual verification step on such problematic cases.

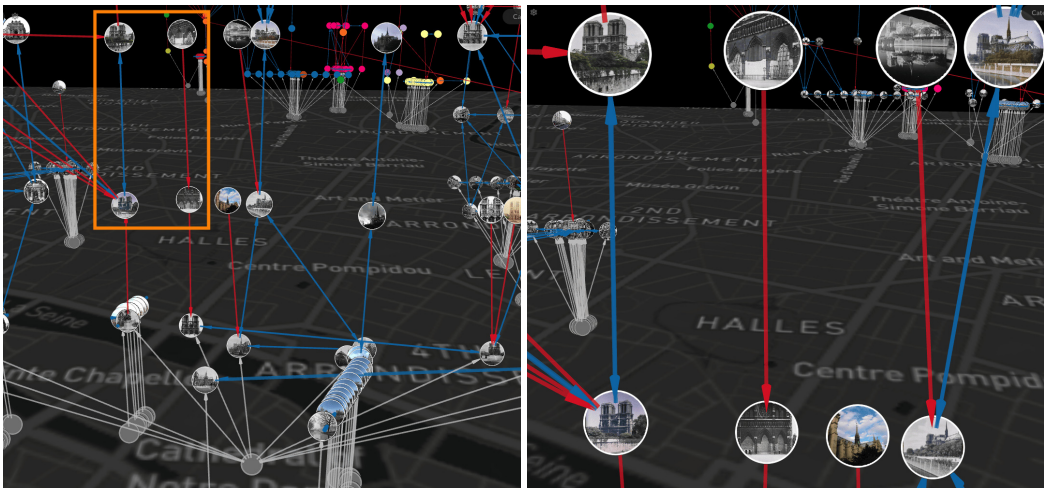
4.4.3 Highly central nodes. Exploiting once again the fact that in an ideal setting, no link should exist between communities, problematic links create bridges between several communities (in this case groups of similar nodes). To identify them, the betweenness coefficient is quite efficient as it estimates in how many shortest paths between any two nodes a specific node is. Thus, a high betweenness score reveals a node linking multiple groups of nodes, thus having potentially incorrect links, as shown in Figure 6. In the platform, visualizing for each node its betweenness coefficient, computed with Brandes' algorithm [11], helps identify cases where automatic linking fails for manual correction of the structuring for instance.



(a) Identification of highly central nodes (dark green) (b) Visual check of the visual similarity links

Fig. 6. Highly central nodes visualization. The node's color represents its betweenness coefficient. A high coefficient indicates potentially incorrect links, as illustrated with the thumbnails in (b).

4.4.4 Spatialized tree representation. Creating trees based on similarities, with location nodes as roots, displays naturally together nodes which are spatially close. Indeed, the GraphXR graph representation can organize nodes in a tree fashion, considering all the similarity links to place the nodes in the scene. This display setting endogenously regroups together non-located nodes not linked to each other but linked to nodes that are spatially close.



(a) Spatially close images in the visualization (b) Visual check of what they depict

Fig. 7. Spatialized tree representation visualization. Close images in the tree-based representation indeed depict the same object, even if no similarity exists between them.

Figure 7 illustrates this. Thus, not linked but potentially similar nodes are identified. In this example, it turns out that the object depicted is the same but without visual similarity between the images. This type of visualization brings out the structure and the inner dynamics of the dataset, enriching the browsing through the entire collection.

Visualizing the dataset in a structured manner helps with its understanding while focusing on potential faults in the structure with dedicated visual clues helps evaluate it. Within an immersive platform where the user can manually intervene on the structure, these clues can be seen as support for manual corrections dedicated to improving the structuring. This leads to a semi-automatic structuring framework detailed in Section 5.

5 MANUAL CORRECTION FOR SEMI-AUTOMATIC STRUCTURING

Visualizing the dataset in a global and structured fashion helps in understanding and analyzing it. Furthermore, the specific visual analysis tools we have proposed in Section 4.4 allows any user to identify faults in the structure, either incorrect links or missing ones. As the user can simply browse through the data, the platform allows for manual intervention on the structure of the graph. Therefore, the user can employ its expertise to correct mistakes and add new structuring information. Those corrections can then be diffused throughout the whole dataset for an even greater impact. This section first introduces in Section 5.1 the different improvements on the structure that can be performed either manually in the platform or automatically due to the graph-based representation. Section 5.2 then develops and evaluates the iterative semi-automatic process which proves promising in alleviating the drawbacks in structuring brought by automatic approaches.

5.1 Improvements of the structuring

In the setting of a graph-based representation of the structured dataset, the structuring can be improved in two ways. First, as detailed in Section 5.1.1 with a expert manually modifying the structure (deleting or adding information). Second, by automatically propagating (and adding) information throughout the graph, as shown in Section 5.1.2 with the location information.

5.1.1 Manual interventions. Visualizing the dataset as a structured graph with visual clues to evaluate the structure first proves to be an excellent opportunity for the expert to intervene on said structure. Indeed, a simple example is the visualization of cross-community links. As the expert can simply visualize them to evaluate the structure, they could also delete them if they are wrong in order to improve the structure. Thus, experts can remove links (of different nature), add new links based on their knowledge but also add information to the nodes and links to enrich the global dataset. Examples of these actions are detailed in Appendix B.

Those interventions on the graph are then exploited to update similarities between images, which led us to extend the global similarity score S_{IJ} of Equation 1 by integrating this new type of similarity:

$$S_{I,J} = \mathcal{N}\left(s^v_{I,J} \times {s^s_{I,J}}^{c_{I,J}{}^s} + {s^e_{I,J}}^{c^e}\right), \tag{3}$$

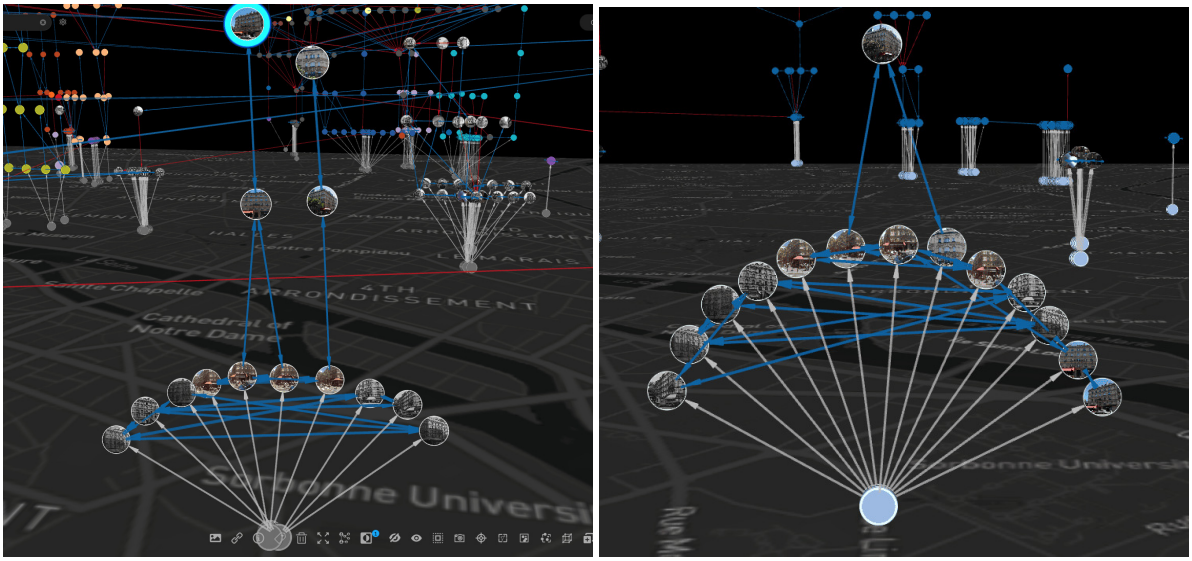
Here $s^e_{I,J}$ is the expert similarity score that reflects the expert's opinion on the similarity between images (independently of $s^v_{I,J}$ and $s^s_{I,J}$) and c^e the confidence associated to the expert's knowledge of the data. Finally, we normalize $S_{I,J}$ in $[0, 1]$ using $\mathcal{N}$, a min-max normalization over S. One can notice that should no expert nor spatial similarity be present, $S_{I,J}$ represents simply the visual similarity, and should simply no expert similarity be present, $S_{I,J}$ encodes the visual similarity weighted by the spatial similarity as defined in Section 3.

Furthermore, if a similarity link between images I and J is deleted through the visualization platform, J is then ranked last in the ranking list of I and the images initially ranked after J are moved closer to I by 1. And vice-versa for I in the ranking list of J. Additionally, $S_{I,J}$ is set to 0, leading all "aggregated" similarity links (visual, spatial or potentially wrong expert ones) to be deleted, with their respective similarities set to 0 too.

An example of an expert similarity link can be as follows: between two pictures taken with a large time gap, the same building depicted twice may have undergone massive renovation (visual similarity $s^v_{I,J}$ is low), or its address may have changed due to the renaming of the street (spatial similarity $s^s_{I,J}$ may be incorrectly low), but it is still the same building. In this case, the link brought by expert can ensure that the two pictures stay linked together, transcending differences in visual aspect and associated metadata. Those links are displayed in **yellow** in the 3D visualization platform.

5.1.2 Location propagation. Linking contents gives the possibility to propagate information through such a graph-based structure. Once enough similarity links are established in the graph-based representation, we continue to exploit spatial information by propagating localization: indeed, some images can be localized from the locations available with their first similar images retrieved. One can easily imagine that propagating this information will increase the impact of structuring through spatial similarity (Section 3.2.3).

There exist many techniques for estimating a location from several candidate locations [43, 55]; here we simply choose to average the 2D position of the first candidate locations which are spatially coherent together (*i.e.* sufficiently clustered spatially). This process can be repeated multiple times until the requirements for propagation are not met anymore, based on some criteria such as the number of linked located images, on the confidence over the visual similarity or the location, etc. To such a new location is associated a confidence score c_I (see Equation 2), based on the spatial coherence of the locations exploited for the propagation and their associated confidence scores.



(a) Detection of a cluster of highly connected localized and non-localized images

(b) Propagation of the locations to non-localized images

Fig. 8. Example of the location propagation process

A visual example of location propagation is displayed in Figure 8: at first (a), in the cluster of similar images, 10 are located (directly linked to a location node pinned on the map) and 4 are not (they are only linked to the spatialized ones). Leveraging the similarity links between them (in blue), the locations are propagated, resulting (b) in 13 located and 1 non-located images.

5.2 Iterative semi-automatic structuring process

As automatic linking can lead to a visualization on which manual modification can be performed, the structuring process can thus be considered as a two-step iterative and semi-automatic process, presented and evaluated below.

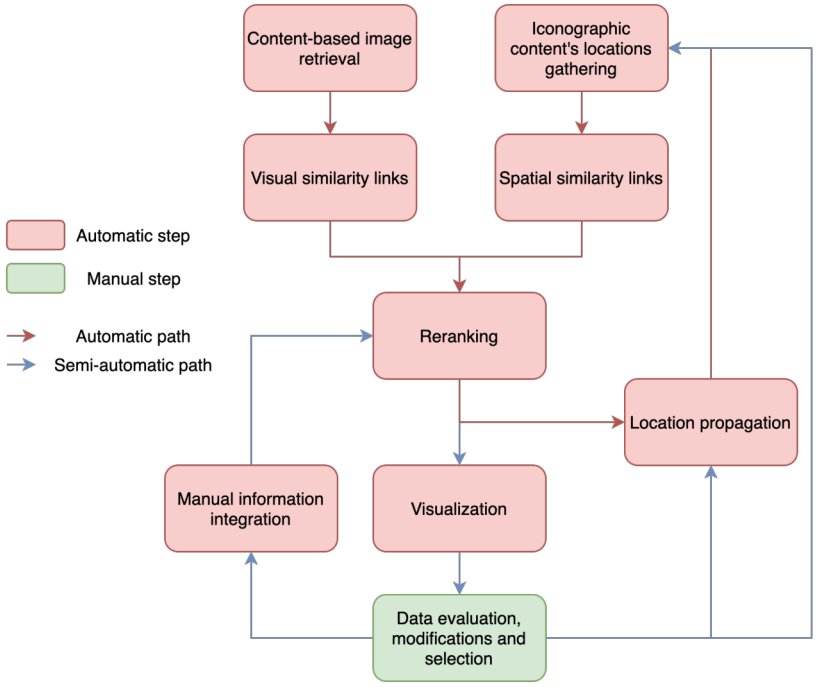


Fig. 9. Overview of the semi-automatic structuring process

5.2.1 Semi-automatic iterative process. It is decomposed in two steps, as follows:

(1) Automatic building of visual and spatial similarity links; the techniques employed are the best ones of Section 3 and of Section 5.1.2 for location propagation. These links feed the graph-based representation of the global structure at large scale.
(2) Manual assessment and improvement of some complex configurations automatically highlighted with visual clues in the 3D visualization environment (Section 4). The graph-based representation of the collection's structure is updated with these new inputs. Then repeat step (1) if necessary.

In this semi-automatic process, summarized with Figure 9, an expert first visualizes the results of the automatic process, and corrects mistakes using the visual guides proposed. The corrected retrieval results are then fed to the diffusion-based re-ranking process or the location propagation process that run offline. The new results can again be visualized for the expert to add new corrections and so on. This approach exploits the best of both worlds, on the one hand automatic processes which handle a large number of data at once and on the other hand an expert knowledge that ensures strong results but is way more time-consuming.

5.2.2 Whole process evaluation. To evaluate the performance of the proposed process, each possible improvement is performed, each step building on top of the previous one. First a step of location diffusion is automatically done,

after a simple retrieval. Then, two sets of, in turn, deletion of incorrect links, creation of spatial similarity links and creation of expert similarity links are performed on two sets of links of different depth (potentially more or less correct). The expert similarities in our case are dedicated to further connect highly different contents representing the same object, that is linking them beyond visual dissimilarity or in the absence of spatial similarity.

The impact of the iterative process can be assessed quantitatively in terms of mAP score, as previously. Table 5 illustrates how this score evolves through iterations involving automatic linking (without and with a diffusion step) enriched with automatic and manual inputs.

Table 5. mAP scores evolution through iterations associating manual linking to simple automatic linking

Action #	Intervention type	Automation level	Number of added information	mAP before diffusion	mAP after diffusion
1	How-A + Location weighting (Sp)	Automatic	-	41.97	61.77
2	+ Location propagation	Automatic	85	42.32	62.20
	Interventions on the first 5 links				
3	+ Deletions (visual)	Manual	70	42.36	62.32
4	+ Creations (expert)	Manual	30	42.40	62.46
5	+ Creations (spatial)	Manual	33	42.43	62.58
	Interventions on the 5th to 10th links				
6	+ Deletions (visual)	Manual	78	42.44	62.59
7	+ Creations (expert)	Manual	26	42.48	63.83
8	+ Creations (spatial)	Manual	27	42.51	64.21

In this experiment, the number of image nodes is 1,637, the number of located images is 537, and the total number of links is around 7,000. The experiments rely on the three different sets of links, for which a series of automatic and manual operations are performed (actions #1 to #8 in Table 5):

- First, we start by weighting a simple retrieval with How-A, using Stereopolis locations ("Sp"), to remain coherent when using automatic location propagation and subsequent weighting with the new locations. The results of these automatic steps correspond to actions #1 and #2 of Table 5.
- The second part of our experiments applies on the structure provided by the first 5 links of the structure from the previous steps. Those links being the strongest, they have a high probability of being correct. Various manual interventions are performed, with results corresponding to actions #3 to #5.
- Third, we perform manual interventions on the structure created with the 5 to 10 first links of the previous manual structuring steps. Those links may thus be more uncertain than the first five, leading to a potentially noisier structure. The results of these interventions correspond to actions #6 to #8.

The amount of information added to automatic image retrieval (location propagation, targeted manual interventions on similarity links) is small, representing each time about 2% of the total information, while the mAP scores reveal that it notably improves the structuring, with a multiplied impact after the automatic diffusion of this new knowledge. The overall improvement of 2.44% of mAP score (from 61.77% to 64.21%) is significant and proves the impact of targeted structuring coupled with a diffusion process. For comparison with the best automatic approaches of retrieval (Table 3), here the overall mAP score reaches the level of the second-best approach combining a single step of re-ranking before diffusion (R3D-SG). Several other conclusions can also be drawn from those results.

First, with action #2 of Table 5, we observe an improvement of the mAP score from 61.77 (using starting locations from Stereopolis only) to 62.20 when computing 85 new locations by propagation. The improvement is quite substantial in terms of global structuring but also in terms of added information, here 85 more locations represent an increase of 16% in terms of localized images. Second, the manual interventions improve the mAP score in a very limited way (0.19% overall) when not coupled with diffusion. But they nourish favorably the diffusion process by bringing it up to a 2.01% improvement. Furthermore, while the first manual interventions on the first 5 links improved the structuring by 0.38%, working on the deeper links improves the mAP score by 1.63%. This shows that correcting the deeper and more uncertain structure has much more impact when diffused afterwards.

To go further in the experiments, Table 6 quantifies the impact of manual processing when building on an initial step of automatic RANSAC-SG + R3D-SG which represents the best solution in terms of mAP score. Once again, two groups of three types of manual interventions were performed, with a number of interventions of the same magnitude, the two groups being performed on links of different depth. Once again, the manual interventions prove relevant as they allow for a 1.1% mAP gain on top of the best automatic performance. Furthermore, the impact of diffusion to propagate the structuring is once more obvious. With a total mAP gain of 0.1% before diffusion, the diffusion multiplies it to reach 1.1%. Once again, improving the structuring, in a targeted fashion, as much as possible before diffusion shows great promise.

Table 6. mAP scores evolution through iterations associating manual linking to optimal automatic linking

Action #	Intervention type	Automation level	Number of added information	mAP before diffusion	mAP after diffusion
1	How-A + RANSAC-SG + R3D-SG	Automatic	-	44.86	65.79
	Interventions on the 5th to 10th links				
2	+ Deletions (visual)	Manual	73	44.87	65.85
3	+ Creations (expert)	Manual	35	44.93	66.25
4	+ Creations (spatial)	Manual	27	44.93	66.43
	Interventions on the 10th to 15th links				
5	+ Deletions (visual)	Manual	72	44.94	66.55
6	+ Creations (expert)	Manual	31	44.95	66.73
7	+ Creations (spatial)	Manual	28	44.97	66.89

Finally, Table 7 redisplays several main mAP results from the two previous Tables and adds insights in terms of complexity, through the number of images re-ranked and computation time. This comparison shows that the ratio between computation time and mAP score improvement is quite favorable for our semi-automatic process.

Indeed, 1 hour of automatic process on the whole dataset, such as RANSAC, does not allow to reach such improvement, as it allows for re-ranking only about 5 images per query, which is very low compared to the 135 we evaluated it in Table 3. Thus, re-ranking only 5 images improves the mAP score after the diffusion process of only 0.01%. Furthermore, due to the large computation time of more complex re-ranking approaches like the best performing one RANSAC-SG + R3D-SG, in 2 hours, their re-ranking could not be performed for all 1,637 query images in the dataset.

The semi-automatic process shows promise both as an alternative and as a suppletive to a longer and exhaustive automatic re-ranking process. Indeed, when building on the quick but less performant location weighting scheme,

Table 7. mAP score improvements through different strategies, facing complexity in terms of re-ranked images amount and computation time

Combination of retrieval and re-ranking approaches	k re-ranked images per query	Re-ranking computation time (hours)	mAP after diffusion
How-A	-	-	59.3
Automatic re-ranking approaches			
How-A + Location weighting (Sp)	135	1/60	61.8
How-A + RANSAC-SG + R3D-SG	135	150	65.8
How-A + RANSAC-LG	135	45	65.5
How-A + RANSAC-SG/RANSAC-LG	5	1.5	59.3
Iterative semi-automatic process building on an automatic location weighting step			
How-A + Location weighting (Sp)	135	1/60	61.8
+ Automatic location propagation	-	-	62.2
+ Manual interventions on the first 5 links	-	1	62.6
+ Manual interventions on the 5-10 links	-	1	64.2
Iterative semi-automatic process building on an automatic RANSAC-SG + R3D-SG step			
How-A + RANSAC-SG + R3D-SG	135	150	65.8
+ Manual interventions on the 5-10 links	-	1	66.4
+ Manual interventions on the 10-15 links	-	1	66.9

in two hours, the semi-automatic process reaches a mAP score of 64.2%, equivalent to that of a simple step of R3D-SG (second best when one single re-ranking step is used before diffusion). When building on the best performing re-ranking combination (RANSAC-SG + R3D-SG), two hours of semi-automatic process further increase the mAP score by a significant 1.1%, reaching a new highest score.

Furthermore, should a visualization platform be exploited, leveraging expert and targeted corrections in a semi-automatic framework proves to be an interesting solution. Indeed, some benefits inherent to the semi-automatic approach are evident. First of all, the corrections manually performed have a high probability of being correct, especially when dealing with the most complicated cases that the automatic approaches miss. Furthermore, as illustrated in Section 3.3, a high provider entropy is necessary for the diffusion to reach its full potential. Manual (and certain) linking oriented towards inter-collection linking could then further multiply the effects of the semi-automatic process and even more improve the structuring globally.

Scalability. Finally, although the proposed semi-automatic framework is promising, its suitability to handle ever growing collections should be discussed. In terms of automatic image retrieval and re-ranking steps, the computation cost may be high but can be optimized through parallelization on a cloud if necessary. More specifically, all costly re-ranking steps (RANSAC, R2D and R3D) can be considered as strong scaling applications [2]. They are independent at computation time and could be parallelized so that for N queries, N processors run simultaneously. The diffusion step essential to the semi-automatic process is also able to handle a very large amount of images without linearly increasing its computation cost, making the process able to handle a much larger collection. Currently, we do not consider this point as a main challenge because we apprehend it as an offline pre-processing step. On the other hand, the scalability of the visualization platform may be the bottleneck. Indeed, GraphXR recommends limiting the maximum number of nodes to 10,000 in a classical setup. From a

practical point of view, a too large number of links could render the visualization incomprehensible, which is why we recommend displaying only a subset of links, or concentrating on an area, *i.e.* a district, of the scene under study, at the same time providing smooth navigation under GraphXR.

6 CONCLUSION

In this article, we have introduced a complete structuring paradigm for interlinking iconographic heritage collections, here dedicated to geographic heritage collections where spatial location and similarities can be exploited. Faced with in silo structuring specific to image collections, we first focus on automatic linking approaches based on the visual content, and review several strategies to improve retrieval at best for these contents. They are all based on re-ranking and are applied on off-the-shelf image descriptors. Note that the latter are not retrained here, but could be when dedicated training datasets will be made available for image heritage contents. Once the limits of these automatic approaches had been reached, due to the high visual variability between multi-source heritage contents, we have chosen to investigate the visualization of the obtained structure, through a spatialized graph-based representation of the dataset and its links. This representation allows for an immersive and interactive visualization in a 3D web-based platform of the iconographic contents in a structured manner. Visualizing the automatically created structure provides the user with a global organized overview of the dataset and helps to identify errors but also discover patterns in the structure, thanks to dedicated visual clues that focus on the most challenging areas. When exploiting jointly automatic indexing and localization approaches with the visualization platform, manual interventions on complex situations can be performed and then diffused throughout the graph, thus propagating structuring information. This iterative, semi-automatic structuring process proves efficient, combining the best of both worlds, expert manual knowledge and large-scale automatic computation. We have shown with this pipeline that visualizing the dataset in a global and structured fashion allows to wisely inject targeted manual interventions. Adding those interventions within an automatic structuring process helps to go beyond the limits imposed by the specificities of iconographic heritage contents to purely automatic dataset-wide structuring.

Although demonstrated with visual and spatial similarities, both the representation model and the structuring pipeline are generic and could accommodate other types of information. For instance, structuring based on semantic information could be added: this information could be extracted from associated metadata or via automatic content-based classification. Once added to the structured dataset, semantic information could be propagated through the dataset, but it could also be leveraged to weigh on spatial or visual similarities and display new visual organization patterns in the dataset.

Our generic structuring model and semi-automatic structuring process could be applied to various collections, each with their own specificities, the primary objective being to add structure to the collections. Increasing the structuring of such collections could first benefit the collections themselves. Indeed, the novel structure can be leveraged for metadata consistency checks, but also as a conduit to propagate metadata information from one content to others, thus improving and enriching the collections. Such structure could also be leveraged to improve the querying process of such inconsistently structured and linked contents. Historians or sociologists for instance could explore the newly structured contents (visually or not), discovering novel data and potentially hidden patterns. Although our proposed, GraphXR-based platform proves to be an adequate solution for visualizing structured contents, exploiting such image collections could be useful in multiple fields, from digital tourism or education to urban planning (architectural refurbishing for instance) via land use monitoring. Exploiting iconographic collections in this various contexts could require specific visualization and interaction tools. Thus, more tailored and conventional search user interface may be needed to exploit the structuring from our proposed process in a more user-friendly and goal-oriented way.

ACKNOWLEDGMENTS

This work was financed by the City of Paris and the French ANRT through Cifre grant 2019/1841. It was carried out using HPC resources from GENCI-IDRIS (grant 2022-AD011013510R1).

REFERENCES

[1] Albert Kahn museum. 2016. *Albert Kahn museum's collections browsing platform.* https://opendata.hauts-de-seine.fr/explore/dataset/archives-de-la-planete/information/?disjunctive.operateur

[2] Gene M Amdahl. 1967. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference.* 483–485.

[3] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *International Conference on Computer Vision.* 1269–1277. https://doi.org/10.1109/ICCV.2015.150

[4] Song Bai, Peng Tang, Philip H S Torr, and Longin Jan Latecki. 2019. Re-ranking via metric fusion for object retrieval and person re-identification. In *Conference on Computer Vision and Pattern Recognition.* 740–749. https://doi.org/10.1109/CVPR.2019.00083

[5] Daniel Barath, Jiri Matas, and Jana Noskova. 2019. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10197–10205.

[6] Nicolas Blanc, Timothée Produit, and Jens Ingensand. 2018. A semi-automatic tool to georeference historical landscape images. *PeerJ* 6 (2018), 1–7. https://doi.org/10.7287/peerj.preprints.27204

[7] Emile Blettery and Valérie Gouet-Brunet. 2023. Re-ranking Image Retrieval in Challenging Geographical Iconographic Heritage Collections. In *Proceedings of the International Conference on Content-based Multimedia Indexing.* 1–7.

[8] Emile Blettery and Valérie Gouet-Brunet. 2024. *Platform demonstration and visual results.* https://www.umr-lastig.fr/emile-blettery/results.html

[9] Emile Blettery, Paul Lecat, Alexandre Devaux, Valérie Gouet-Brunet, Frédéric Saly-Giocanti, Mathieu Brédif, Laetitia Delavoipière, Sylvaine Conord, and Frédéric Moret. 2020. A spatio-temporal web application for the understanding of the formation of the parisian metropolis. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 6, 4/W1 (2020), 45–52. https://doi.org/10.5194/isprs-annals-VI-4-W1-2020-45-2020

[10] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[11] Ulrik Brandes and Christian Pich. 2007. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos* 17, 07 (2007), 2303–2318. https://doi.org/10.1142/S0218127407018403

[12] Bingyi Cao, André Araujo, and Jack Sim. 2020. Unifying Deep Local and Global Features for Image Search. In *European Conference on Computer Vision*, Vol. 12365. 726–743. https://doi.org/10.1007/978-3-030-58565-5_43

[13] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. 2022. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[14] Ondřej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. 2007. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In *International Conference on Computer Vision.* https://doi.org/10.1109/ICCV.2007.4408891

[15] Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J Carroll, and Brian McBride. 2014. RDF 1.1 concepts and abstract syntax. *W3C recommendation* 25, 02 (2014), 1–22.

[16] Agni Delvinioti, Hervé Jégou, Laurent Amsaleg, and Michael E Houle. 2014. Image retrieval with reciprocal and shared nearest neighbors. In *International Conference on Computer Vision Theory and Applications*, Vol. 2. 321–328. https://doi.org/10.5220/0004672303210328

[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Conference on Computer Vision and Pattern Recognition Workshops.* 224–236. https://doi.org/10.1109/CVPRW.2018.00060

[18] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Conference on Computer Vision and Pattern Recognition.* 12124–12134. https://doi.org/10.1109/CVPR52688.2022.01181

[19] Douglas Duhaime. 2017. PixPlot visualization platform. https://dhlab.yale.edu/projects/pixplot/

[20] Martin A Fischler and Robert C Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (1981), 381–395. https://doi.org/10.1145/358669.358692

[21] French Culture Ministry. 2019. *Base Mémoire, French Culture Ministry's platform for heritage content.* https://www.pop.culture.gouv.fr/

[22] French Mapping Agency. 2016. *Remonter le temps, French Mapping Agency platform for heritage data.* https://remonterletemps.ign.fr/

[23] French National Library. 2015. *Gallica, French National Library website.* https://gallica.bnf.fr/

[24] Florent Geniet, Valérie Gouet-Brunet, and Mathieu Brédif. 2022. ALEGORIA: Joint Multimodal Search and Spatial Navigation into the Geographic Iconographic Heritage. In *ACM International Conference on Multimedia.* 6982–6984. https://doi.org/10.1145/3503161.3547746

[25] Albert Gordo, Filip Radenovic, and Tamara Berg. 2020. Attention-based query expansion learning. In *European Conference on Computer Vision*, Vol. 12373. 172–188. https://doi.org/10.1007/978-3-030-58604-1_11
[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, Vol. 2016-Decem. 770–778. https://doi.org/10.1109/CVPR.2016.90
[27] HistoryPin. 2010. *HistoryPin collaborative platform.* https://www.historypin.org/en/
[28] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. 2017. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *Conference on Computer Vision and Pattern Recognition*. 2077–2086. https://doi.org/10.1109/CVPR.2017.105
[29] Vincent Jaillot, Valentin Rigolle, Sylvie Servigne, John Samuel, and Gilles Gesquière. 2021. Integrating multimedia documents and time-evolving 3D city models for web visualization and navigation. *Transactions in GIS* 25, 3 (2021), 1419–1438. https://doi.org/10.1111/TGIS.12734
[30] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. 2022. Correlation Verification for Image Retrieval. In *Conference on Computer Vision and Pattern Recognition*. 5374–5384. https://doi.org/10.1109/CVPR52688.2022.00530
[31] Wei-Chao Lin. 2019. Aggregation of Multiple Pseudo Relevance Feedbacks for Image Search Re-Ranking. *IEEE Access* 7 (2019), 147553–147559. https://doi.org/10.1109/ACCESS.2019.2942142
[32] Wei-Chao Lin. 2022. Block-based pseudo-relevance feedback for image retrieval. *Journal of Experimental and Theoretical Artificial Intelligence* 34, 5 (2022), 891–903. https://doi.org/10.1080/0952813X.2021.1938695
[33] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. 2023. LightGlue: Local Feature Matching at Light Speed. In *International Conference on Computer Vision*. https://arxiv.org/pdf/2306.13643.pdf
[34] Liris Laboratory Vcity Team. 2023. Virtual City Project. https://projet.liris.cnrs.fr/vcity/
[35] Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. 2019. A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and Vr/Ar. *Virtual Archaeology Review* 10, 21 (2019), 1–13. https://doi.org/10.4995/var.2019.11867
[36] Ferdinand Maiwald, Christoph Lehmann, and Taras Lazariv. 2021. Fully automated pose estimation of historical images in the context of 4D geographic information systems utilizing machine learning methods. *ISPRS International Journal of Geo-Information* 10, 11 (2021), 748. https://doi.org/10.3390/IJGI10110748
[37] Lionel Moisan, Pierre Moulon, and Pascal Monasse. 2016. Fundamental matrix of a stereo pair, with a contrario elimination of outliers. *Image Processing On Line* 6 (2016), 89–113.
[38] Navilium. 2016. *Navilium collaborative platform.* https://www.navilium.com/
[39] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-Scale Image Retrieval with Attentive Deep Local Features. In *International Conference on Computer Vision*, Vol. 2017-Octob. 3476–3485. https://doi.org/10.1109/ICCV.2017.374
[40] Jianbo Ouyang, Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Contextual similarity aggregation with self-attention for visual re-ranking. *Advances in Neural Information Processing Systems* 34, 3135–3148. https://proceedings.neurips.cc/paper/2021/hash/18d10dc6e666eab6de9215ae5b3d54df-Abstract.html
[41] Shanmin Pang, Jin Ma, Jianru Xue, Jihua Zhu, and Vicente Ordonez. 2019. Deep Feature Aggregation and Image Re-Ranking With Heat Diffusion for Image Retrieval. *Transactions on Multimedia* 21, 6 (2019), 1513–1523. https://doi.org/10.1109/TMM.2018.2876833
[42] Nicolas Paparoditis, Jean-Pierre Papelard, Bertrand Cannelle, Alexandre Devaux, Bahman Soheilian, Nicolas David, and Erwann Houzay. 2014. Stereopolis {II}: {A} multi-purpose and multi-sensor {3D} mobile mapping system for street visualisation and {3D} metrology. *Revue Française de Photogrammétrie et de Télédétection* 200 (apr 2014), 69–79. https://doi.org/10.52638/rfpt.2012.63
[43] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. 2020. Benchmarking image retrieval for visual localization. In *International Conference on 3D Vision*. 483–494. https://doi.org/10.1109/3DV50981.2020.00058
[44] Vikus Project. 2014-2017. *Vikus Project's interactive demos.* https://uclab.fh-potsdam.de/vikus/
[45] Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Conference on Computer Vision and Pattern Recognition*. 5706–5715. https://doi.org/10.1109/CVPR.2018.00598
[46] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*. 3–20. https://doi.org/10.1007/978-3-319-46448-0_1
[47] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *Transactions on Pattern Analysis and Machine Intelligence* 41, 7 (2019), 1655–1668. https://doi.org/10.1109/TPAMI.2018.2846566
[48] Marko A Rodriguez and Peter Neubauer. 2010. Constructions from dots and lines. *Bulletin of the American Society for Information Science and Technology* 36, 6 (2010), 35–41.
[49] John Samuel, Vincent Jaillot, Clément Colin, Diego Vinasco Alvarez, Eric Boix, Sylvie Servigne, and Gilles Gesquière. 2023. UD-SV: Urban data services and visualization framework for sharing multidisciplinary research. *Transactions in GIS* (2023). https://doi.org/10.1111/TGIS.13049
[50] Paul Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning feature matching with graph neural networks. In *Conference on Computer Vision and Pattern Recognition*. 4938–4947. https://doi.org/10.1109/CVPR42600.2020.

00499
[51] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition*. 4104–4113. https://doi.org/10.1109/CVPR.2016.445
[52] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision*, Vol. 9907. 501–518. https://doi.org/10.1007/978-3-319-46487-9_31
[53] Xi Shen, Yang Xiao, Hu Shell Xu, Othman Sbai, and Mathieu Aubry. 2021. Re-ranking for image retrieval and transductive few-shot classification. In *Advances on Neural Information Processing Systems*. 25932–25943. https://proceedings.neurips.cc/paper/2021/hash/d9fc0cdb67638d50f411432d0d41d0ba-Abstract.html
[54] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015), 1–14. https://arxiv.org/pdf/1409.1556.pdf%E3%80%82
[55] Yafei Song, Xiaowu Chen, Xiaogang Wang, Yu Zhang, and Jia Li. 2016. 6-DOF image localization from massive geo-tagged reference images. *Transactions on Multimedia* 18, 8 (2016), 1542–1554. https://doi.org/10.1109/TMM.2016.2568743
[56] Matthias Springstein, Stefanie Schneider, Javad Rahnama, Eyke Hüllermeier, Hubertus Kohle, and Ralph Ewerth. 2021. iART: A Search Engine for Art-Historical Images to Support Research in the Humanities. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2801–2803.
[57] Swiss Art Research Infrastructure. 2022. *Images of Switzerland Online*. https://www.timemachine.eu/images-of-switzerland-online/
[58] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. 2021. Instance-level image retrieval using reranking transformers. In *International Conference on Computer Vision*. 12105–12115. https://doi.org/10.1109/ICCV48922.2021.01189
[59] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. 2016. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision* 116 (2016), 247–261. https://doi.org/10.1007/S11263-015-0810-4
[60] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. 2020. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In *European Conference on Computer Vision*, Vol. 12346 LNCS. 460–477. https://doi.org/10.1007/978-3-030-58452-8_27
[61] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports (Nature)* 9, 1 (2019), 5233.
[62] Nicolas Verdier, Eric Mermet, and Carmen Brando. 2017. Oronce Fine platform. https://psigehess.hypotheses.org/oronce-fine
[63] Qi Wang, Weidong Min, Daojing He, Song Zou, Tiemei Huang, Yu Zhang, and Ruikang Liu. 2020. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. *Science China Information Sciences* 63 (2020), 1–12. https://doi.org/10.1007/S11432-019-2811-8
[64] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*. 568–578. https://doi.org/10.1109/ICCV48922.2021.00061
[65] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2 A large-scale benchmark for instance-level recognition and retrieval. In *Conference on Computer Vision and Pattern Recognition*. 2575–2584. https://doi.org/10.1109/CVPR42600.2020.00265
[66] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. 2018. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE transactions on visualization and computer graphics* 25, 6 (2018), 2311–2330.
[67] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. 2021. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *International Conference on Computer Vision*. 11772–11781. https://doi.org/10.1109/ICCV48922.2021.01156
[68] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. 2020. ResNeSt: Split-Attention Networks. In *Conference on Computer Vision and Pattern Recognition Workshops*. 2736–2746. https://doi.org/10.1109/CVPRW56347.2022.00309
[69] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. 2020. Understanding Image Retrieval Re-Ranking: A Graph Neural Network Perspective. (2020). https://arxiv.org/abs/2012.07620
[70] Xulu Zhang, Zhenqun Yang, Hao Tian, Qing Li, and Xiaoyong Wei. 2022. Indicative Image Retrieval: Turning Blackbox Learning into Grey. *arXiv preprint* (2022). https://arxiv.org/abs/2201.11898
[71] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. 2023. R^2Former: Unified *R*etrieval and *R*eranking Transformer for Place Recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

A VISUALIZATION PLATFORM DETAILS

Figure 10 gives an overview of the 3D visualization platform's interface. More screenshots and dynamic layout are visible on website [8].

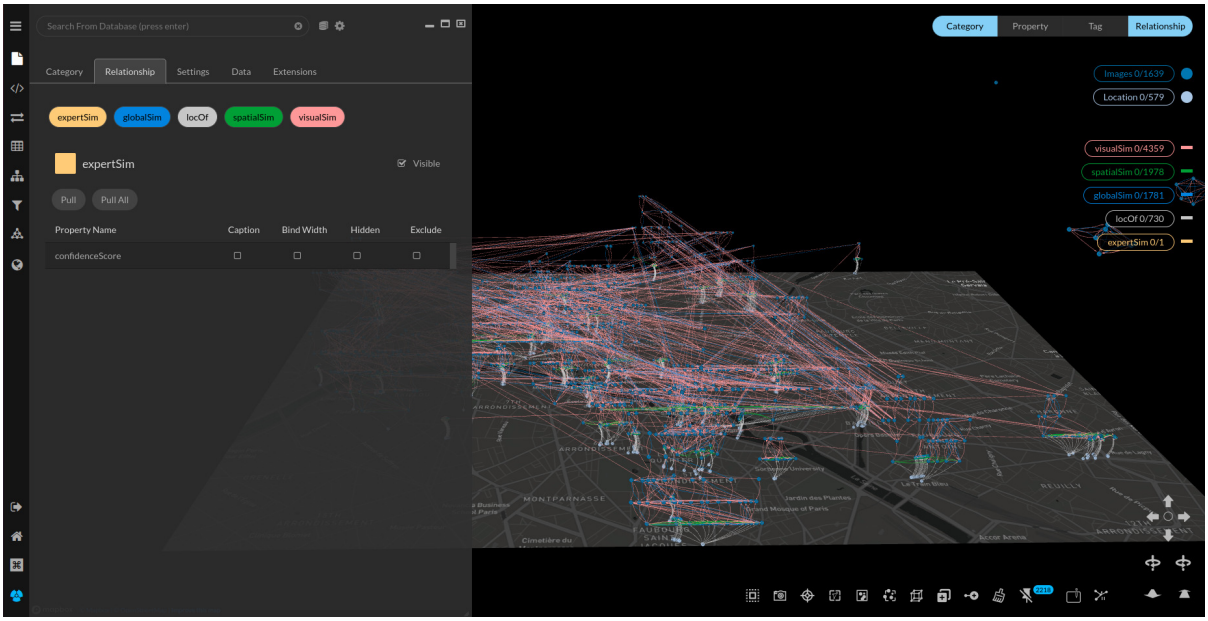


Fig. 10. Overview of the 3D visualization platform's interface

On the left, as illustrated in Figure 11(a), several menus are accessible. They allow multiple actions of which we list several below:

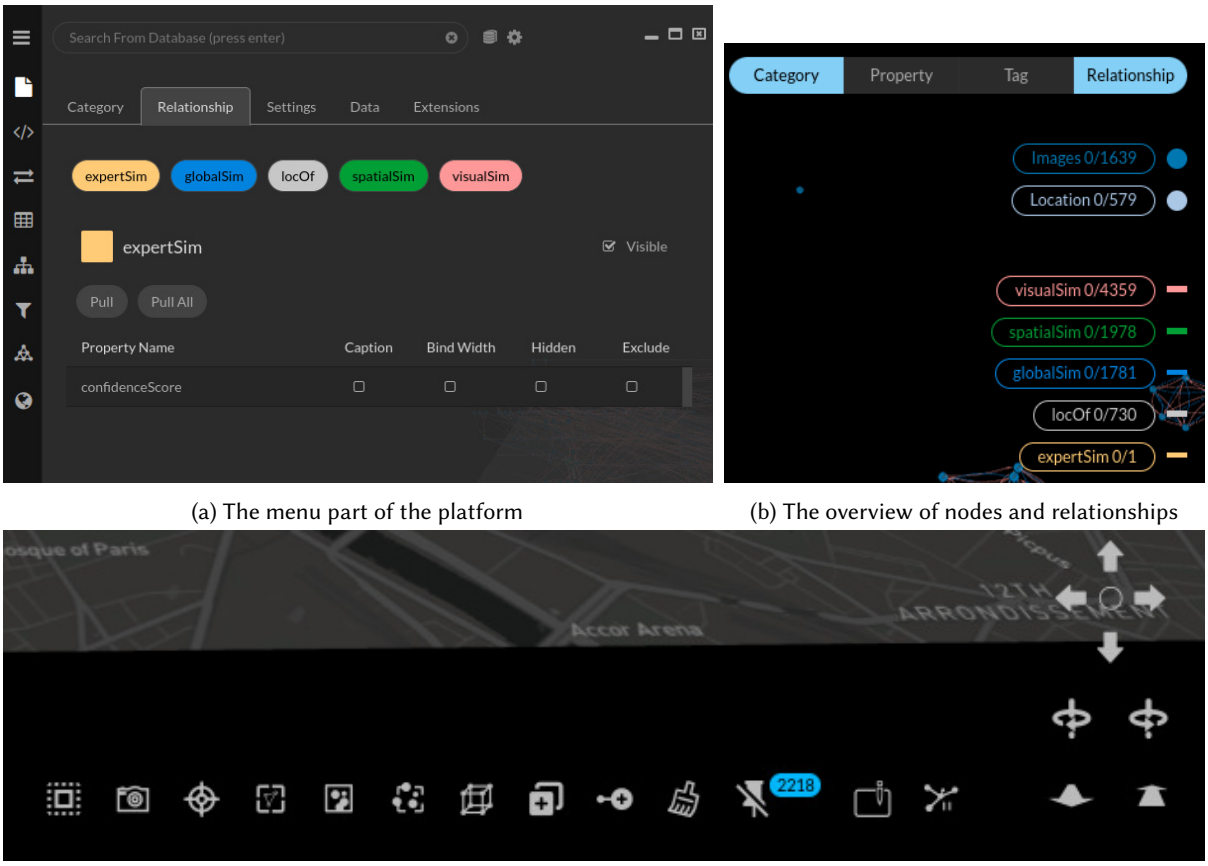
- loading data and querying both graph and database;
- apply transformations to the graph;
- compute graph algorithms;
- select from multiple layouts;
- filter the displayed data;
- add a map;
- exploit extensions of GraphXR.

On the top right, nodes and links are summarized. The user can select the nodes or links by category but also easily modify their representation, rather than going through the menu. Figure 11(b) illustrates this.

On the bottom right (see Figure 11(c)), several visualization tools are available, to select neighbors of a selected node, to invert the selection or hide the selected nodes for instance.

Finally, one of the most useful aspect of the platform to create a usable platform for any user is the possibility to create macros with the GraphXR specific tool, Grove. That is create a button that once clicked will perform successive actions in a specific order. This allows to load the data easily, but also perform algorithms computation or apply specific layouts to the graph. This functionality is illustrated in Figure 12 and proves essential to ensure two things. On the one hand, the user does not have to navigate through the different menus and remember the specific order in which to perform the actions. On the other hand, this ensures the fact that the platform can be

used by more-or-less tech-savvy users but also ensures consistency in the work performed. Indeed, with the same input data, the same macro will get to the same output, which is paramount if several users work jointly on the same data.



(a) The menu part of the platform

(b) The overview of nodes and relationships

(c) The tools of the platform

Fig. 11. Overview of the platform's interface

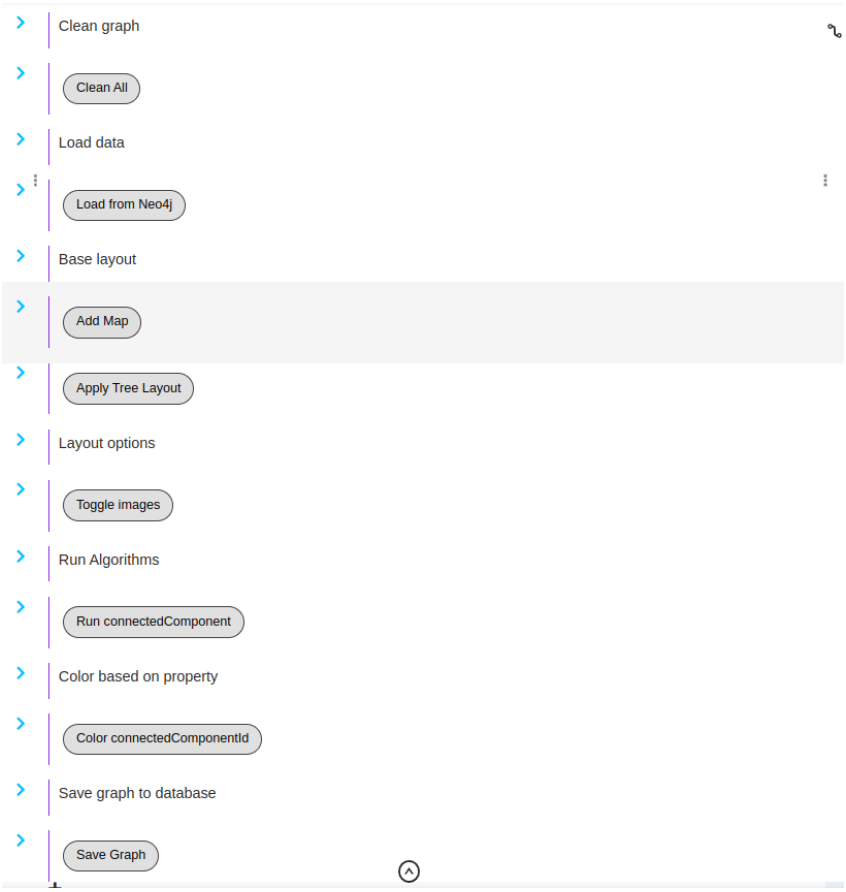


Fig. 12. Grove extension macro buttons examples

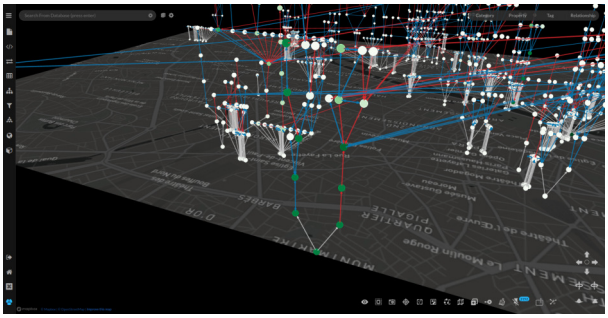
B MANUAL CORRECTIONS EXAMPLES

This section illustrates processes of manual corrections within the graph visualized in the 3D platform.

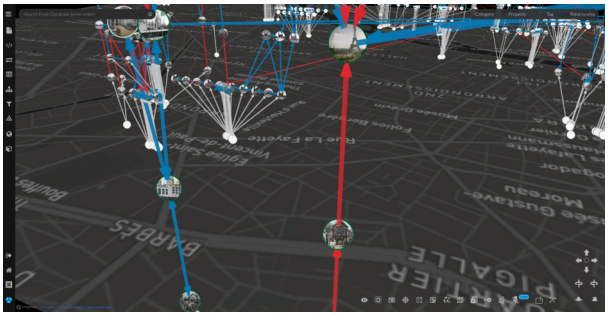
B.1 Highly central nodes edge clearing

In the platform, visualizing for each node its betweenness coefficient, computed with Brandes' algorithm [11], helps the user in identifying central nodes linked to multiple clusters/communities. Checking the edges of these central nodes is highly beneficial for global structuring of the dataset. Indeed, clearing the edges of those central nodes, that is deleting or strengthening links with spatial or expert similarities, ensures a clearer frontier between communities.

This is illustrated in Figure 13 where a chain of nodes seems to link two communities because their betweenness coefficient is high. Cleaning those links removes this high betweenness, indicating an improvement in the structuring. Highlighting nodes based on their centrality coefficient thus focuses the user's intervention on highly impacting evaluations.



(a) Identification of highly central nodes (dark green)



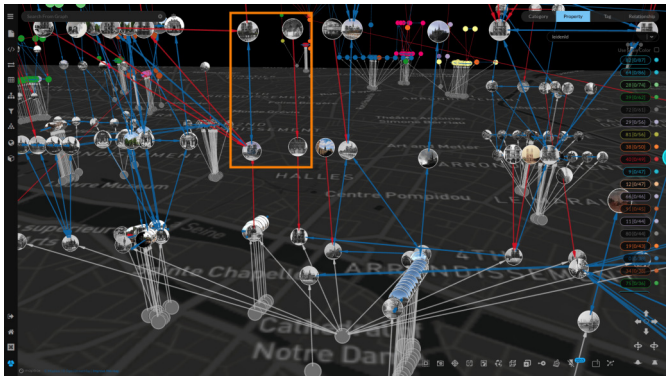
(b) Visual check of the visual similarity links

(c) Deletion of the incorrect link

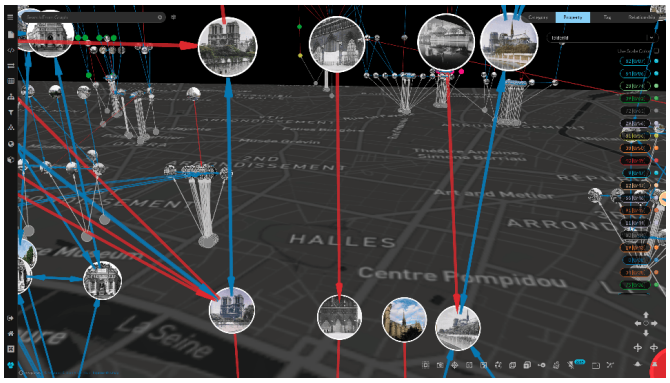
(d) Recomputation of the betweenness coefficient, showing that different communities are no longer wrongly connected

Fig. 13. Highly central node edge clearing

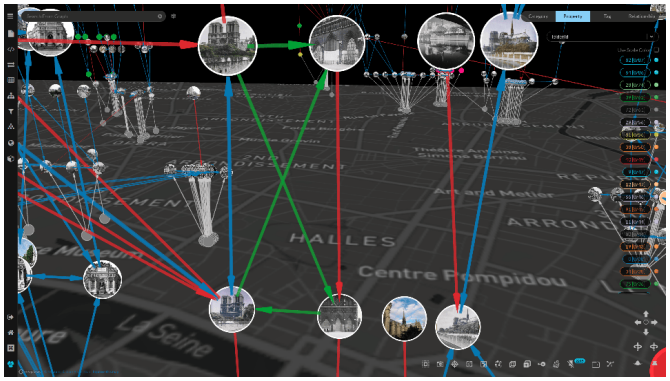
B.2 Spatialized tree representation for spatial linking



(a) Spatially close images in the visualization



(b) Visual check of their actual similarity



(c) Creation of spatial similarity links (green) for the following automatic diffusion process

Fig. 14. Spatial similarity links creation process aided by the tree representation

Exploiting a tree visualization easily allows users to densify the linking between the nodes as finding visually similar or spatially close images supposes to look only at nearby nodes. This is what Figure 14 represents. First, not-linked but close nodes are identified (in the yellow rectangle in (a)). Their thumbnails are then visualized (b) and it appears that they depict two parts of the same scene. Thus, there are no visual similarity links (at least not strong ones) but the user can create spatial similarity ones (in green in (c)). Actions #5 and #8 of Table 5 show that adding spatial similarities that are then used to improve global similarities and overall structuring via diffusion is as efficient as intervening on visual similarities, indicating that both interventions should be used jointly for best performance.

Furthermore, it also helps identifying clusters prime for location propagation as previously shown in Figure 8. That is clusters that can be selected in order to propagate automatically the location information of located images in the cluster to non-located images in the cluster, based on their visual similarity links.

Received 25 January 2024; revised 23 April 2024; accepted 20 May 2024