# Making Use of Pre-existing Street Art Object Metadata

Martin de la Iglesia

Braunschweig, Germany; E-Mail: martin.delaiglesia@gmail.com

**Abstract**

In graffiti and street art studies, we are currently facing a paradoxical situation: vast numbers of publications relevant to our field—some of them academic, most of them not; from journal papers to coffee-table books—are continuously being published, but even the scholarly-oriented among them typically provide only sparse data about individual graffiti pieces and street art objects. It is rare to find complete metadata records containing information about the artist, the precise location, measurements, and the date of completion. Efforts are being made by individual projects and researchers to gather comprehensive and structured metadata, but those efforts take time and yield only small amounts of data. While it is important that these efforts are continued, a different, complementary approach is proposed here that aims to 'quickly and dirtily' gather 'messy' data. The idea is to make use of work that has already been carried out instead of trying to describe the same artworks in better ways time and again. This requires us to learn how to deal with incomplete data from vastly different sources. Effectively, such an approach lowers the threshold for data sources to become useful for street art researchers. Almost anything can become a valuable resource, even amateur websites (including abandoned ones) and print publications about local and obscure street art. This paper demonstrates how to extract object metadata from street art websites and digitised printed books, and how to feed it into a database that can be a potential treasure trove of street art object data.

**Keywords**

data mining; data science; data wrangling; digitisation; metadata; non-academic publishing

## 1. Introduction

In the research of street art (including graffiti; 'street art' is used as an umbrella term in this text), it is a common problem that authors provide only incomplete or imprecise information regarding the artworks they write about (discussed in greater detail by de la Iglesia, 2015). In order to unambiguously identify any such object, the metadata provided would ideally include a photograph (or multiple photographs taken from different angles and at different points in time), the date of creation (or at least the date when the photograph was taken), the location (either as street address or geographic coordinates), the artist (including a machine-readable identifier such as an authority record URI, if available), the title (or all of the titles under which the work is known), a complete and precise transcription of any text present in the work, the technique/medium/genre (to distinguish e.g. stencil graffiti from style writing), the dimensions, and possibly other properties. One kind of attempt to respond to this need for metadata is to create graffiti databases in which the desired information is gathered and provided in a highly structured way; see, for instance, the projects INGRID, INDIGO, and Spraycity.at presented in this volume, or the author's website (de la Iglesia, 2007–2020).

However, as laudable as those project efforts may be, they suffer from a major shortcoming: considering the sheer number of artworks that have been created around the world, that are currently being created, and that in all likelihood will be created in the future, it is obvious that such efforts will never cover more than a small fraction of those artworks, given the laboriousness of the generation of sufficiently rich metadata records. Those databases are typical-


**Figure 1.** Detail of a screenshot of *Wooster Collective* previewing a post by Marc Schiller from 2005, featuring an artwork by Ragweed.

ly limited to a specific geographic area and/or period, while the vast majority of objects remain outside of their scope. Still, it is not as if all of those artworks were 'unknown', so to speak. Many of them do leave traces and are being covered by documentation efforts in a wider sense. If we take into account not only rich and structured but also incomplete and 'messy' metadata, we may find that a vast amount of graffiti-related information already exists: in the form of weblogs, photo websites and other Internet resources on the one hand, and coffee-table books, magazines and other printed matter, aimed at a wider audience beyond academia, on the other hand. Would it be possible to somehow tap into this vast amount of messy data and make it usable for research purposes at all? This is the central question that this contribution is trying to answer.

**2. Pre-existing Street Art Object Metadata**

2.1. Websites

Among the different kinds of resources on the World Wide Web, there are several with relevance to street art that come to mind: general-purpose pools of photographic images such as Flickr (which may have subcommunities dedicated to street art, e.g. "One World Street Art & Graffiti" with more than half a million of pictures uploaded since its foundation in 2009; https://www.flickr.com/groups/951083@N24), personal websites of individuals or groups dedicated to street art which often take the form of weblogs (such as *Wooster Collective*, see below), or posts on social media platforms such as Instagram which, however, are difficult to aggregate and extract data from. As an example of a street art website and how to make use of the data contained therein, let us now take a closer look at *Wooster Collective*.

Operated by Sara and Marc Schiller from New York, *Wooster Collective* (Schiller & Schiller, 2003–2018) had been for some time one of the definitive news sources about the global street art scene. The earliest retrievable post is from 2003, and after 2016 the posting activity has become so infrequent that this weblog can be considered inactive. *Wooster Collective* ran posts on various aspects related to street art, such as book releases, film releases, artist portraits, interviews, and exhibition openings, but there were also a number of posts that simply documented a new artwork that was deemed notable for some reason, either spotted by the *Wooster Collective* authors themselves or submitted by someone else. This is the category of posts that is of interest here. An example of such a post is pictured in Figure 1.

More precisely, what we see here is not yet the actual blog post but rather a preview as it appears together with other blog post previews on the front page or a search results page. Already at this point, we can distinguish four pieces of relevant metadata: an image representing the artwork in question, a date on which the blog post was published ("October 31, 2005"), a location mentioned in the title ("Montreal"), and an artist name given in the text body of the post ("Ragweed"). No categories or tags have been assigned to this post, as we can see from the empty "Posted in:" field.

When we click "Continue Reading" or the title to view the entire blog post, we find that it does not yield more information than the preview (Figure 2). We see raw HTML code—the language in which websites are written, which is normally rendered by the web browser to display websites and not itself visible to the user—that references two image files, but due to an error in the HTML code, the web browser displays the HTML snippet instead of the images. The first one of those images is identical to the preview image we have already seen, and we can also see that the entire text body of the blog post consists of "Artist: Ragweed". The only additional information we get in the full post view is the URL of the second image, but in this case, it shows a different stencil graffiti piece by Ragweed in Montreal, and as it is difficult to automatically determine whether multiple images in one blog post show the same or different artworks, it is safer for us to only extract the preview image and ignore any others, treating the blog post as if it was about a single artwork only.

"Seen on the Streets of Montreal" can be considered part of an implicit series of posts on *Wooster Collective* which have a similar title structure, "Seen on the streets of" followed by a place name (usually a city, but sometimes also a country, building, or city district). Variants include "Seen on" followed by a street name, "Seen in", "Seen near", and "Seen under", plus several times the misspelling "Seen

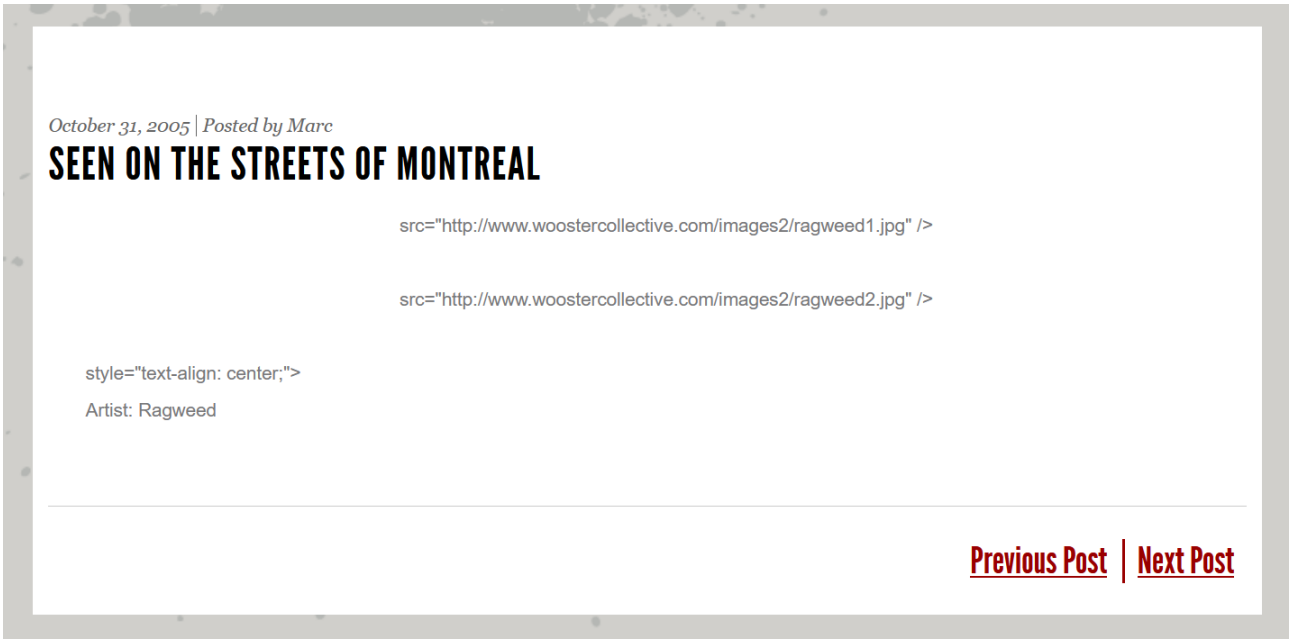


**Figure 2.** Detail of a screenshot of the blog post "Seen on the Streets of Montreal" by Marc Schiller from 2005 on *Wooster Collective* (http://www.woostercollective.com/post/seen-on-the-streets-of-montreal1).

on the steets of". If we look for these patterns in the HTML source of the *Wooster Collective* main page that lists all blog posts, we can extract the URLs of all relevant posts. However, www.woostercollective.com always only displays a certain number of posts, followed by a "Load more posts" button. One way to deal with this problem and still obtain the complete list of posts is to employ browser automation software such as Selenium (https://www.selenium.dev). Another, much simpler workaround is to fetch the monthly lists into which *Wooster Collective* compiles its posts (the "Filter By Date" button at the top of the main page); here, too, there are some months with too many blog posts to display at once without clicking "Load more posts", but one still gets a sizeable amount of posts via this method, which might be enough for the demonstration purposes in this context. Out of this sample of blog posts, a search (using regular expressions in a script written in the Perl programming language) for titles containing "seen on the streets of" and similar structures yields 85 posts, each one of which can now be turned into a metadata record relating to one artwork. (Other implicit series of suitable blog posts on *Wooster Collective* relating to individual artworks might be "X in Y", e.g. "Kaws in England", with 12 posts in the aforementioned sample; "X Shows Us How It's Done in Y", e.g. "Pøbel Shows Us How Its Done In Tokyo, Japan", 8 posts; "Fresh Stuff From X", e.g. "Fresh Stuff From Elbow Toe - 'Tastes Like Chicken'", 91 posts; or "Shit We're Diggin'", e.g. "Shit We're Diggin': NeSpoon", 44 posts.)

From the title, we can extract our first metadata field, the location of the artwork. In titles such as "Seen on the Streets of Montreal", we can simply assume that everything after "Streets of" will be a place name of some sort and put it in the location field of our metadata record. As for the three "Seen on..." posts, e.g. "Seen on Sullivan", we can safely assume that these are streets in New York (where Sara and Marc Schiller, the *Wooster Collective* creators, live), and thus automatically add "Street, New York" to each, so that the location field value becomes e.g. "Sullivan Street, New York". There are two titles out of which we cannot extract any location information: one early post from 2003 is simply titled "Seen on the Street… 'Buddy" [sic], and another


**Figure 3.** Pages 14–15 from Tapies, 2018.

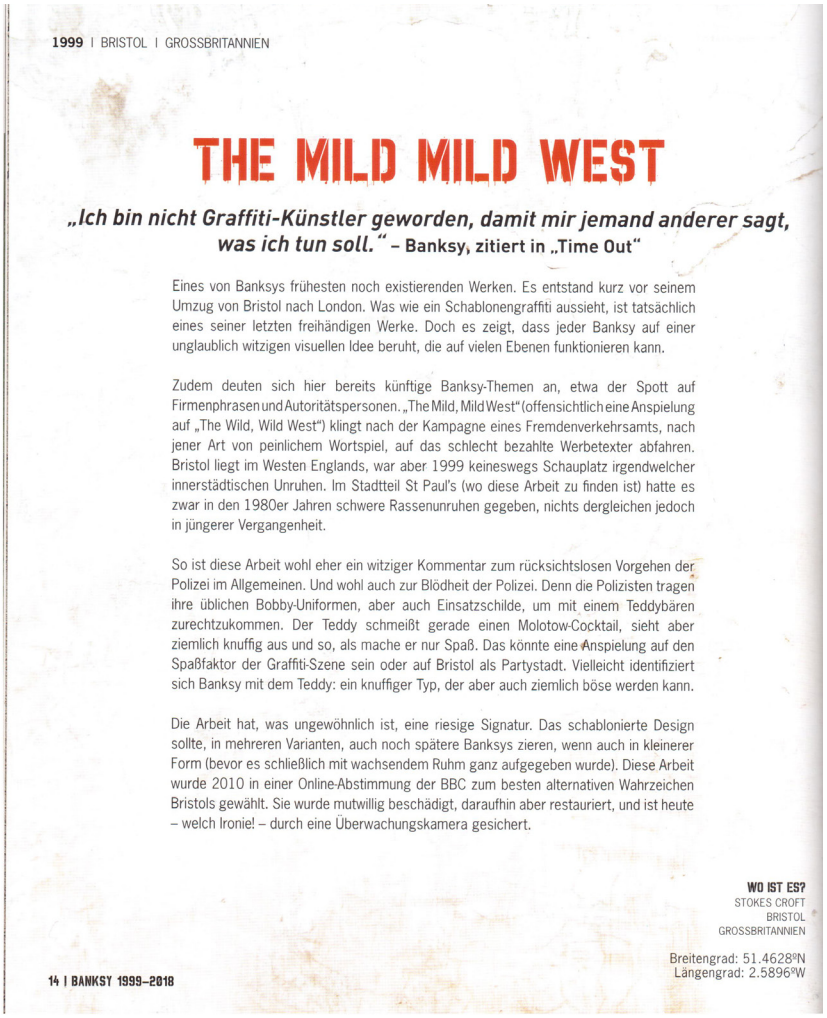


**Figure 4.** Page 14 from Tapies, 2018.

from 2016 is titled "Seen Near Lily's Juice Bar" (most likely Miss Lily's on West Houston Street, New York). The other 83 place designations are all correct, as can be seen when checking them against a geolocation web service. In this case, the place designations were queried in Google Maps, and they all returned a result and can thus be assumed to be correct. For instance, when one enters the search string "the Palazzo Reale in Milan" (from the title "Seen in the Palazzo Reale in Milan") in Google Maps, it returns the point with the coordinates N45.46319°, E9.19116°. As mentioned above, for two out of 85 location field values, no geolocation could be performed, so the location coverage in our sample is roughly 98%.

It is more difficult to extract artist names because the artist indication in the text body, if present at all, does not always follow an easily recognisable structure such as "Artist: ...". However, in 31 cases, that formula is used, and the artist name (or name of the artist collective) can be extracted. For instance, the artists in this sample include DS, N4T4, Rems182 and Zukclub. The coverage of the artist metadata field is thus approximately 36%.

It is far easier to extract dates and photographs. The date on which a post was published is always given in the format "October 31, 2005", which can be converted into a format more suitable for automatic sorting, such as "2005-10-31". Each blog post has at least one image, and for the reasons outlined above, we are only going to consider the preview image. We can automatically download all those preview images in case the website goes offline one day. In one case in the sample, however, the hyperlink to the image file is broken. Thus we have a date coverage of 100% and a photograph coverage of 99%.

Additionally, we could add each text body in its entirety as a kind of 'description' field. While it is difficult to automatically extract further information from those texts, they can still be useful to human readers. For instance, one such blog post text simply says, "More here.", the word "here" being the anchor of a hyperlink pointing to the URL http://www.coletivografico.com. Although this website is now defunct, one can still infer from the address that the artist group responsible for the artwork in question is Coletivo Gráfico, a street art collective from Rio de Janeiro.

At this point, we have assembled a data collection of 85 records, each relating to a piece of street art and consisting of four to five data fields (location, date, photograph, description text, and at least some artist names). The potential usefulness of this data collection is discussed below in Section 3.

2.2. Books and Magazines

As an example of printed matter, we are now going to consider a typical street art monograph: *Banksy 1999–2018* by Xavier Tapies, the German edition of *Where's B**ksy?* (Tapies, 2018). Conveniently, the book is rigidly structured, as each double page is dedicated to one artwork by Banksy, the right-hand page containing a photograph and the left-hand page giving textual information (Figure 3). Likewise, all of the text pages follow the same structure (Figure 4).

In the top left corner, the year of creation is given ("1999"), followed by location information on the city level ("Bristol | Großbritannien"). The centred heading of the text, in a red font that imitates stencilled letters, indicates the title of the object in question. This is followed by a quotation by either Banksy himself, as is the case here, or someone else. Then the main text body describes the artwork and gives some background information. In the bottom left corner, the page number is given, and the bottom right corner provides a more precise location ("Stokes Croft / Bristol / Großbritannien"), including geographic coordinates.

Once we have digitised the pages using a scanner or camera, we need to apply Optical Character Recognition (OCR) software to the digital images. There are different OCR applications available, and one can spend much time configuring and training them. If, however, one simply uploads the images to one of the many free OCR web services and takes the text output as it is, as will be shown here for demonstration purposes, some challenges will have to be faced when further processing the text. As Figure 5 shows, the biggest problem is not the character recognition per se—almost all of the individual characters were recognised correctly—but the layout; in the printed book, the text is distributed across several fields on the page, and the OCR software tries to re-

```
14 I BANKSY 1999-2818

1999 | BRISTOL | GROSSBRITANNIEN

THE MILD MILD WEST

was ich tun soll." - Banksy, zitiert in .Time Out"

Eines von Banksys frühesten noch existierenden Werken. Es entstand kurz vor seinem
Umzug von Bristol nach London. Was wie ein Schablonengraffiti aussieht, ist tatsächlich
eines seiner letzten freihändigen Werke. Doch es zeigt, dass jeder Banksy auf einer
unglaublich witzigen visuellen Idee beruht, die auf vielen Ebenen funktionieren kann.

Zudem deuten sich hier bereits künftige Banksy-Themen an, etwa der Spott auf
Firmenphrasen und Autoritatspersonen. , The Mild, Mild West" (offensichtlich eine Anspielung
auf „The Wild, Wild West") klingt nach der Kampagne eines Fremdenverkehrsamts, nach
jener Art von peinlichem Wortspiel, auf das schlecht bezahlte Werbetexter abfahren.
Bristol liegt im Westen Englands, war aber 1999 keineswegs Schauplatz irgendwelcher
innerstädtischen Unruhen. Im Stadtteil St Paul's (wo diese Arbeit zu finden ist) hatte es
zwar in den 1980er Jahren schwere Rassenunruhen gegeben, nichts dergleichen jedoch
in jüngerer Vergangenheit.

So ist diese Arbeit wohl eher ein witziger Kommentar zum rücksichtslosen Vorgehen der
Polizei im Allgemeinen. Und wohl auch zur Blödheit der Polizei. Denn die Polizisten tragen
ihre üblichen Bobby-Uniformen, aber auch Einsatzschilde, um mit einem Teddybären
zurechtzukommen. Der Teddy schmeißt gerade einen Molotow-Cocktail, sieht aber
ziemlich knuffig aus und so, als mache er nur Spaß. Das könnte eine Anspielung auf den
Spaßfaktor der Graffiti-Szene sein oder auf Bristol als Partystadt. Vielleicht identifiziert
sich Banksy mit dem Teddy: ein knuffiger Typ, der aber auch ziemlich böse werden kann.

Die Arbeit hat, was ungewöhnlich ist, eine riesige Signatur. Das schablonierte Design
sollte, in mehreren Varianten, auch noch spätere Banksys zieren, wenn auch in kleinerer
Form (bevor es schließlich mit wachsendem Ruhm ganz aufgegeben wurde). Diese Arbeit
wurde 2010 in einer Online-Abstimmung der BBC zum besten alternativen Wahrzeichen
Bristols gewählt. Sie wurde mutwillig beschädigt, daraufhin aber restauriert, und ist heute
- welch Ironie! - durch eine Überwachungskamera gesichert.



„Ich bin nicht Graffiti-Künstler geworden, damit mir jemand anderer sagt,

WO IST ES?
STOKES CROFT
BRISTOL
GROSSBRITANNIEN

Breitengrad: 51.4628°N

Längengrad: 2.5896°W
```

**Figure 5.** Text output of an OCR web service applied to page 14 from Tapies, 2018.

arrange this text into a linear order. The OCR text output starts with the text in the bottom left corner, i.e. the page number, followed by the text in the top left corner, i.e. the year of creation. After the title, the quotation by Banksy is cut in half, with the first part ("' Ich bin nicht Graffiti-Künstler geworden [...]") given after the main text block, and the second part ("was ich tun soll [...]") before.

Despite these problems with the layout, the text structure is still sufficiently preserved to extract several pieces of information, which was done again by means of regular expressions in a Perl script. Thus for each object in the book,

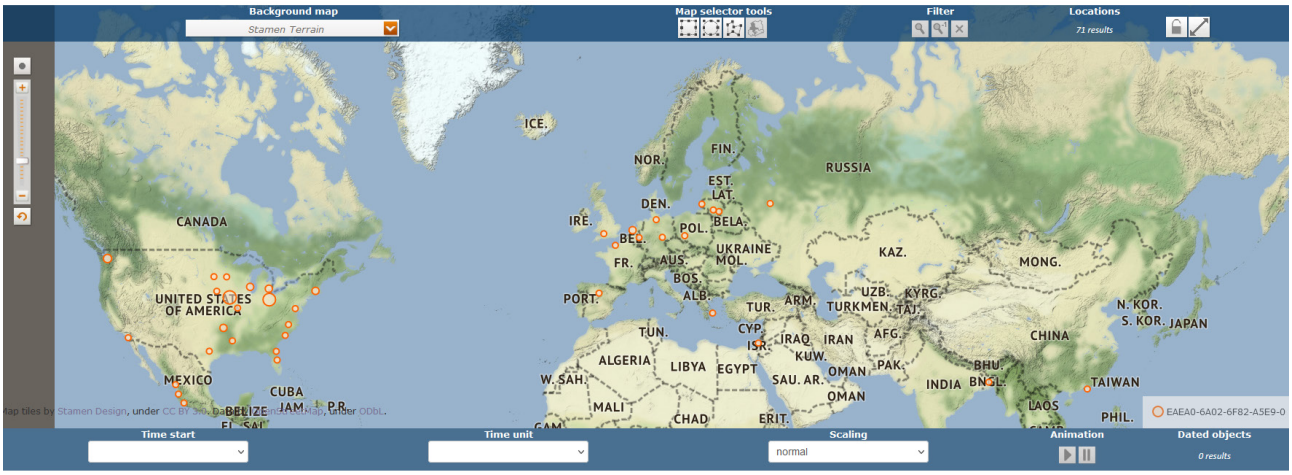


**Figure 6.** Map of records in combined dataset created with the DARIAH-DE Geo-Browser.

we are able to extract the year of creation, title, verbal location, coordinates, description text, and photograph, and of course, we already have the artist's name because it is always Banksy. However, the OCR result is not always as satisfactory as for page 14. In a sample consisting of the first 20 double pages of *Banksy 1999–2018*, all of the dates could be successfully extracted, and all of the verbal location information was correct and useful when checked against Google Maps, but only 60% of the titles and 85% of the coordinates could be recognised. Typical errors that prevented the successful recognition of these fields were the muddling up of the order of textual elements on the page, which led to the misplacement of the title, and the misidentification of the degree sign in the coordinates (printed here in the shape of °) as e.g. the percentage sign %.

### 3. The Combined Database

If we now combine the two sample datasets into one, we end up with 105 street art object metadata records (85 from *Wooster Collective* and 20 from *Banksy 1999–2018*). All of them contain a date (although that is only the year in the case of the Banksy works), all but one have a photograph, and all but two contain a location. However, only for 49% of the works, the name of the artist is known, and the title for only 11% (as only Tapies assigns titles to the works he covers and *Wooster Collective* does not). Geographic coordinates are present in 16% of the records, but as already mentioned above, more coordinates can be converted from the verbal location information. Thus the combined dataset is quite heterogeneous or 'messy'. Can we make use of it nonetheless?

Given the nearly complete coverage of location information, one of the most obvious ways to visualise the data would be to plot it on a map, which makes it easier for humans to see how the objects are distributed geographically. There are web services that carry out the conversion of addresses into coordinates and the plotting of coordinates on a map in a single step, such as the DARIAH-DE Geo-Browser (https://geobrowser.de.dariah.eu; registration required), but if we are not careful, the resulting map may look like the one in Figure 6. Orange dots represent street art objects in the dataset, and as we can see, most of them seem to be located in the United States. That is because of how the geolocation completion feature of the Geo-Browser works: many place names were seen as ambiguous and thus identified erroneously; for instance, when the location in *Wooster Collective* was simply given as "Athens", it was identified by the Geo-Browser as Athens, Michigan, whereas the location string "Athens, Greece" was located correctly. Other examples of problematic geolocation results include Paris, Texas and Rome, Illinois, when the correct locations would have been in France and Italy, respectively. Diligent selection and configuration of the geolocation application may

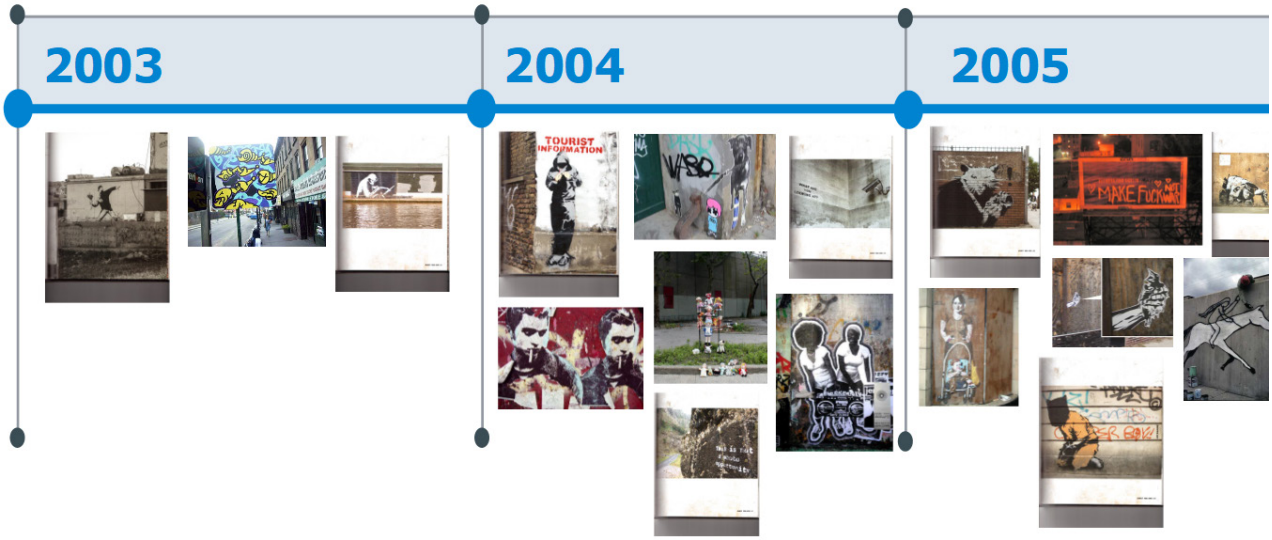


**Figure 7.** Mock-up, created manually by the author, of a timeline with photographs indicating objects from the corresponding year. Photographs from *Wooster Collective* and Tapies, 2018.

prevent such problems.

Apart from a geographic visualisation, we could also arrange our data chronologically, given the complete coverage of date (or at least year) information in our dataset. Furthermore, there happens to be a large overlap between the two subsets, as the years 2003 (when *Wooster Collective* started) through 2016 (when *Wooster Collective* stopped posting regularly) are also covered by *Banksy 1999–2018*, as its title already indicates. For instance, we could simply plot the objects on a timeline (Figure 7). Such a timeline is an efficient way to immediately convey the quantitative development (in this case increase) of objects over time, although it may not always reflect real-world developments in the field of street art but might be susceptible to possible biases in the data sources.

A different kind of utilisation of our data would be to query the database directly to obtain specific information. If, for instance, we wanted to find out what other street artists besides Banksy were active in the UK at the same time, we could simply search within the location field for places in the UK, excluding works by Banksy himself. (In other, larger datasets, we would also need to limit the date range to exclude works from before the beginning of Banksy's career.) This requires the location data to have been correctly recognised, normalised and expanded automatically by a geolocation service so that e.g. the location "Newcastle" in the source has become something like "Newcastle upon Tyne, Tyne and Wear, England, United Kingdom" in the database. If we then search for locations that end with "United Kingdom", we find four objects from *Wooster Collective*: one by N4T4 in Nuneaton from 2005, an anonymous work in Bristol from 2006, a piece by Mobster in Newcastle from 2008, and one by DS in London from 2011.

**4. Possible Issues**

4.1. Processing Complex Layouts

Extracting data from *Banksy 1999–2018* proved straightforward because of its rigid structure of one artwork per double page and one photograph on each right-hand page. However, many other street art books and magazines feature more complex layouts in which several photographs are arranged on the same page. Figure 8, for instance, shows several pages from a booklet on street art in a district of Braunschweig, Germany (Markwort, 2020). Some


**Figure 8.** Pages 23 (photographs by The Bridge e.V.), 58, 59, 30 and 31 (photographs by Dietlinde Schulze) from Markwort, 2020.

photographs, like the ones on the double page in the bottom right picture, are printed flush, i.e. directly next to each other, while others are separated by a narrow 'gutter' in the colour of the page background. Not only is it difficult for a computer to recognise where one image ends and another begins, but it is also hard to tell to which images the caption texts refer to. It would take advanced image segmentation or edge detection algorithms to successfully extract the photograph of each individual artwork and to assign the correct corresponding caption text.

4.2. Obtaining Transcriptions via OCR

As mentioned at the beginning of this paper, it would be highly useful if all metadata records came with a complete transcription of any textual content in the artwork in question, e.g. the words "THE MILD MILD WEST..." and "BANKSY!" in the Banksy mural shown in Figure 3. As we are using OCR anyway to extract textual data about the object, would it not be feasible to use OCR to also obtain the spray-painted letters within the artwork? In this example, OCR software has difficulties recognising all of the letters due to their irregular shapes, particularly in the idiosyncratic "BANKSY!" signature, but there are other, more severe problems with this approach in general. One such problem is the inability of OCR software to distinguish between background and foreground, and thus between irrelevant and relevant writing. For instance, Figure 9 shows a page from a book titled *Graffiti. From A to Z* (Campos & Valle Padilla, 2010). The number 12 is part of the artwork, but if we apply OCR to this image, we get not only "12." as the text output but also "GUAYABAS DE CALIDAD" from the print on the fruit box in the shop window, which is clearly not part


**Figure 9.** Page from *Graffiti. From A to Z*, 2010 (unpaginated). Photograph by Itzel Valle Padilla.


**Figure 10.** Page from *Graffiti. From A to Z*, 2010 (unpaginated). Photograph by Itzel Valle Padilla.

of the graffiti artwork. For a human, this fact is easy to recognise, but not for a computer.

Another problem—basically the opposite of the previous—is caused by pictures of poor quality that lead to wrong OCR results, or more precisely: the characters present may be recognised correctly by the OCR software, but the resulting text is still faulty because of truncated or partially covered words in the original photograph. Consider, for instance, Figure 10, also taken from the book *Graffiti. From A to Z*. The text on the largest of the depicted stickers is recognised by OCR software as "Que muera el celula", as the last letter, r, is obscured by another sticker. The original wording was "Que muera el celular" (roughly, "Death to the cell phone"). If the text "Que muera el celula" enters the database unchecked, the consequence is not only that a database search for the word "celular" yields no result, but also that there is the danger of mistaking that word for 'la célula' (cell). For these reasons, unsupervised mass OCR is not recommended, which raises the question if it would not be simpler to enter any transcriptions manually in the first place.

### 4.3. Qualifiers Add a Layer of Complexity

So far, the data fields in our metadata records had a simple key–value structure, i.e. each field, if present, was filled with a string of characters or a number, so that the entire dataset could be represented, e.g. as a table. Ideally, however, we would like to record more information *about* at least some of the data points, i.e. to add qualifying statements to them. For instance, for geographic coordinates, it would be desirable to record their accuracy, i.e. the number of decimals given in the source, so that we can tell whether e.g.

the points N34.0833, W118.3418 and N34.08325123, W118.34175987 refer to the same object. For photographs in our dataset, it is crucial to record any licence under which the photograph was originally published so that we know if and how we can re-publish it (more on that below). It would also be useful to record each source from which a piece of information was taken (or perhaps even the name of the researcher who performed the data extraction, and the name and settings of the software employed), especially if several sources refer to the same object so that there are e.g. several different titles given for the same artwork. Such additional 'meta-metadata' essentially turns a two-dimensional dataset into a three-dimensional one that can no longer be represented as a simple spreadsheet. The data gains scholarly soundness but becomes harder to handle and process.

4.4. Legal Issues

So where is this sample dataset of 105 records that has been described in this paper? Why is there no hyperlink to it so that other researchers can use it and add more data to it, instead of having to build their own dataset from scratch? The reason is that legal barriers make it difficult to put such a dataset on the Internet. (The following describes the situation in the author's home country, Germany, but the legal circumstances are similar all around the world.) For one thing, there is the 'description text' field which contains texts such as the 300-word piece by Xavier Tapies on Banksy's *The Mild Mild West*. Such texts are protected by copyright from being re-published without the author's consent (or that of his or her heirs for 70 years after the author's death). With photographs, the matter is more complicated. One might think that neither the original, publicly visible, two-dimensional artwork itself nor a photograph thereof is protected by copyright, but a recent much-discussed court ruling (Reiss-Engelhorn-Museen vs. Wikimedia, cf. Initiative Urheberrecht, 2018) suggests that such photographs do indeed have some sort of legal protection, lasting for 50 years after publication, from being shared without the photographer's permission. But even if we leave out description texts and photographs, take 'just the facts' such as location, artist, title etc. and put that information on the Internet, we might still run into trouble. While the individual factual statements (e.g. "Banksy's *The Mild Mild West* was created in 1999 and is located in Bristol") are not protected in any way and may be publicly re-stated, it can be argued that the sum of all those statements from e.g. *Banksy 1999–2018* constitutes a 'database' which took considerable effort to compile, and therefore we would be not be allowed to re-publish a substantial portion of that 'database' without the permission of the person who compiled it (Kreutzer & Lahmann, 2019). Of course, it is a matter of debate what exactly a 'substantial portion' is; this applies more likely to the 20 out of 90 artworks from *Banksy 1999–2018* than the 85 out of more than a thousand artworks on *Wooster Collective*.

In any case, if one wants to err on the side of caution, it is best not to publish any data gained in the ways described above on the Internet. It should be safe, however, to create such a dataset for one's own personal use, or even to share it among a limited number of other people, e.g. a research group. Another strategy would be to approach the rights holders and obtain permission to re-publish the data or to encourage them to apply a suitable licence to their data that facilitates re-use. This is something that e.g. the Japanese Visual Media Graph project has done with regard to fan-made databases of anime, Japanese video games, and other popular media from Japan, albeit with only a small number of data sources (Pfeffer & Kacsuk, 2021; see also the project website at https://jvmg.iuk.hdm-stuttgart.de). A third option to deal with the copyright restrictions would be to publish only small fragments of the source datasets online. For such a truncated database to be useful for research, one would have to make sure that the records selected for publication constitute a representative sample and do not contain any biases regarding, e.g. chronological or geographical coverage.

**5. Conclusions**

Is it worth it, then, despite all the difficulties described, to take the trouble and put together a database from websites and books in the manner outlined above, if the result is a collection of messy data that may not even be shared online? To answer this question, it is important to be aware of the capabilities and incapabilities of such a database. For instance, even if more data sources are added and the data pool grows to thousands of records, so-called known-item searches will rarely be successful, i.e. when you have

a particular artwork in mind about which you would like to conduct research, it is unlikely that your dataset will contain a record of it. Instead, the dataset could be useful for exploratory searches, e.g. if you want to see some examples of street art from a country or a period of time that you are not yet familiar with. Perhaps—although this would require a dataset of considerable size in order to be statistically valid—one could even use the data to devise hypotheses on the quantitative development of street art over space and/or time. The perhaps most convincing argument to simply try it for yourself and get your own dataset started is that there is so much information related to street art already out there, online or on our bookshelves, that it would be a pity not to make more use of it. Serendipitous connections between metadata records from diverse sources may be revealed that we would have never encountered by querying pre-existing online databases or flipping through books.

**Conflict of Interests**

The author declares no conflict of interests.

**References**

Campos, C., & Valle Padilla, I. (2010). *Graffiti. From A to Z = Graffiti. De A à Z = Graffiti. Von A bis Z = Graffiti. Van A tot Z*. BooQs.

de la Iglesia, M. (2007–2020). *Schablonengraffiti in Freiburg-Mittelwiehre*. http://graffiti.freiburg.bplaced.net

de la Iglesia, M. (2015). Towards the Scholarly Documentation of Street Art. *Street Art & Urban Creativity Journal*, *1*(1), 40-49.

Initiative Urheberrecht. (2018). *BGH: Fotografien gemeinfreier Gemälde geschützt*. https://urheber.info/diskurs/2018-12-20-bgh-fotografien-gemeinfreier-gemaelde-geschuetzt

Kreutzer, T., & Lahmann, H. (2019). *Rechte an Forschungsdaten und Datenbanken*. https://irights.info. https://irights.info/artikel/rechte-an-forschungsdaten-und-datenbanken/29587

Markwort, M. (2020). *Street Art & Graffiti im Westlichen Ringgebiet*. plankontor.

Pfeffer, M., & Kacsuk, Z. (2021, November 16). *Offene Forschungsdaten am Beispiel des Projekts „Japanese Visual Media Graph"*. Hochschule der Medien. Open Up! Webinar-Reihe, Stuttgart. https://openup.iuk.hdm-stuttgart.de/programm-der-webinar-reihe-winter-2021-22/

Schiller, M., & Schiller, S. (2003–2018). *Wooster Collective*. http://www.woostercollective.com

Tapies, X. (2018). *Banksy: 1999–2018* (D. Fehrmann, Trans.) (Genehmigte deutschsprachige Sonderausgabe). Frölich & Kaufmann.